The Racial Composition of Forensic DNA Databases

Erin Murphy* and Jun H. Tong**

Forensic DNA databases have received an inordinate amount of academic and judicial attention. From their inception, numerous scholars, advocates, and judges have wrestled with the proper reach of DNA collection, retention, and search policies. Central to these debates are concerns about racial equity in forensic genetic practices. Yet when such questions arise, critics typically just assert that forensic DNA databases are not demographically representative. Such assertions are expressed in vague or conclusory terms, without a citation to actual data or even to concrete estimates about the actual composition of DNA databases.

This Article endeavors to fill these gaps in the literature by providing demographic information about the composition of forensic DNA databases. We draw upon two sources. First, we obtained data from states in response to our requests under freedom of information laws. Second, we devised an original estimate based on public information about each state's DNA collection policies and the demographic data that matches those policies. In other words, we reverse-engineered the national DNA database. Both approaches revealed dramatic disparities in the racial composition of DNA databases, including that DNA profiles from Black persons are collected at two to three times the rate of White persons.

DOI: https://doi.org/10.15779/Z381G0HV8M.

* Professor, New York University School of Law. This Article benefitted from the feedback and comments provided by Montgomery Slatkin, Stephen Schulhofer, Doc Edge, Andrea Roth, the NYU School of Law Faculty Workshop, and the Public Law Workshop at the University of Minnesota School of Law. I am very grateful for extraordinary research assistance from Ben Morris and Carola Beeney.

** Associate, Desmarais LLP, J.D., New York University School of Law, 2018.

We then use our data on the actual and estimated racial composition of DNA databases to identify and illuminate four questions fundamental to forensic DNA policy. First, the data centers racial justice concerns as critical to debates about the proper scope of collection and search policies, as well as the impact of forensic DNA database practices more generally. Second, the data casts light on the significance, determinacy, and stability of race and ethnicity as meaningful biological and social categories. Third, the data provides insight into the advantages and disadvantages of choosing among architectural approaches when collecting, storing, and searching sensitive data such as genetic profiles. And finally, the data prompts questions about genetic privacy more generally, including how to weigh the significance of criminal justice practices in an increasingly genetically transparent society.

INTRODUCTION

For over two decades, U.S. law enforcement officials have been steadily amassing an enormous repository of genetic information in the form of the national forensic DNA database system. Nicknamed "CODIS," the Combined DNA Index System is actually several databases in one, all of which utilize a standard DNA profile that reports genetic information from either thirteen or twenty places on the genome.[1] CODIS indexes several profile categories, including unknown persons who left biological traces at crime scenes ("forensic" or "casework" samples), missing persons, the relatives of missing persons, and unidentified human remains.[2] The largest and most significant indexes, however, are the collections of DNA profiles from known persons with criminal justice histories, namely the convicted and arrested persons indexes.

Each state and the federal government require certain convicted persons—and in some states, arrested persons—to give a sample of their DNA to law enforcement for testing, storage, and inclusion in DNA databases. The Federal Bureau of Investigation (FBI) oversees the national DNA database (formally known as the National DNA Index System, or NDIS), which in September of 2018 contained profiles from 13.5 million offenders (primarily convicted persons) and over 3.2 million arrestees.[3] Each state also operates its own state-

1. *See generally Combined DNA Index System (CODIS)*, FBI, https://www.fbi.gov/services/laboratory/biometric-analysis/codis [https://perma.cc/HL8N-H9K6] (providing an overview of CODIS, including its development and future).

2. *See id.*

3. *CODIS - NDIS Statistics*, FBI (Sept. 2018), https://web.archive.org/web/20181016070747/https://www.fbi.gov/services/laboratory/biometric-analysis/codis/ndis-statistics [hereinafter *Sept. 2018 NDIS Statistics*]. NDIS data from September 2018 is used throughout this Article in computing relevant statistics, because the disclosures by the states were largely made in the summer of 2018. But the FBI periodically updates these statistics. As of July of 2020, the database contained 14.2 million offender profiles, 4.1 million arrestee profiles, and 1.0 million

level database, which may contain more profiles than are held at the national level, and some localities operate local databases as well.

At this time, the national database operates only as a pointer system. That is, the FBI does not seek or retain identifying or demographic information about the individuals whose profiles are held in the database as a result of state collection policies.[4] Instead, profiles are stored using a series of numbers that link them to the lab, specimen, and analyst from which they originated but nothing else; specific information about the person, case, or criminal history is kept only by the collection agency.[5]

The DNA database system is also a data*bank*. The physical specimen that is collected from the individual to generate the genetic profile is retained by the state.[6] Occasional retesting of previously tested samples is not uncommon. For instance, once states began conducting familial searches of DNA databases—that is, searches of the database not for the perpetrator of an offense, but for a close relative—it became common to retest samples of interest to identify the male-chromosome specific (or Y-STR) profile.[7] It also may be that, given the shift in 2017 from a 13-loci standard profile to a 20-loci standard profile, laboratories may undertake retesting of select samples to ensure that the database contains the most comprehensive profile and to diminish the risk of coincidental matches as the database grows.[8] However, there has not yet, to public awareness, been any campaign to systematically retest stored samples.

---

forensic profiles. *CODIS - NDIS Statistics*, FBI (July 2020), https://www.fbi.gov/services/laboratory/biometric-analysis/codis/ndis-statistics [https://perma.cc/YNK9-TZW7] [hereinafter *July 2018 NDIS Statistics*].

4. The FBI operates a "Next Generation Identification" system that integrates conventional biometrics such as fingerprints as well as cutting-edge technologies such as facial or iris recognition. It also maintains an index of individuals "of special concern" with biometric markers drawn from "the Immigration Violator File . . . , convicted sex offenders, and known or appropriately suspected terrorists." *Next Generation Identification (NGI)*, FBI, https://www.fbi.gov/services/cjis/fingerprints-and-other-biometrics/ngi [https://perma.cc/DK9Z-XQJV]. In the past, the FBI has suggested that it might seek to integrate DNA profiles into this NGI system, but thus far that has not occurred. However, as law enforcement agencies adopt technology that enables Rapid DNA analysis, the databases are more likely to merge, or at least a DNA index of persons of "special concern" may develop that serves as a bridge connecting biometric or other identifying information with certain DNA profiles. *See Rapid DNA*, FBI, https://www.fbi.gov/services/laboratory/biometric-analysis/codis/rapid-dna [https://perma.cc/A5B6-QJTQ]; *see also* ERIN E. MURPHY, INSIDE THE CELL 164–65 (2015) (explaining the anticipated integration of Rapid DNA into the FBI's CODIS and NGI systems through the "booking environment").

5. *See Frequently Asked Questions on CODIS and NDIS*, FBI, https://www.fbi.gov/services/laboratory/biometric-analysis/codis/codis-and-ndis-fact-sheet [https://perma.cc/8QAW-9XEH] (explaining that "[n]o names or other personal identifiers of the offenders, arrestees, or detainees are stored using the CODIS software" and listing the four pieces of specimen data that are stored).

6. *See* MURPHY, *supra* note 4, at 15.

7. *See id.* at 194–95 (describing Colorado's process for familial searching, which includes retesting all male candidate samples for the Y-STR loci).

8. The FBI's quality assurance standards require that the laboratory "have and follow a procedure for the verification and resolution of database matches," which may also involve routine retesting. *See* QUALITY ASSURANCE STANDARDS FOR FORENSIC DNA TESTING LABORATORIES stand.

Forensic DNA databases have received an extensive amount of academic and judicial attention. From the inception of forensic DNA, scholars, advocates, and judges have wrestled with the proper reach of DNA collection, retention, and search policies.[9] As Part I.B shows, central to these debates are concerns about racial equity in forensic genetic practices. Yet when the question of the racial and ethnic composition of DNA databases arises, commenters typically declare that forensic DNA databases are not demographically representative.[10] Such assertions are often stated in vague or conclusory terms without a citation to actual data or even a supposed estimate about the exact composition of DNA databases.[11]

This Article endeavors to fill this gap in the literature by providing demographic information about the composition of forensic DNA databases. We draw upon two sources. First, we submitted Freedom of Information Act (FOIA) requests for information about state DNA databases, which give a snapshot of a state's database composition at a moment in time. Second, we devised an original estimate of the racial and ethnic composition of DNA databases based on public information about each state's collection policies and the demographic data that matches those policies. In other words, we attempted to reverse-engineer the database.

Using these two methods, we generated the only quantitative picture in the literature of the racial, ethnic, and gender composition of DNA databases. Our efforts disclose dramatic racial disparities in the national DNA database. Most prominently, our FOIA requests, which gave us insight into the composition of the national database, paint a stark picture. For instance, in every jurisdiction, DNA profiles from Black persons are collected and stored in the state database at two to three times the rate of Black persons in the population. In contrast, only a tiny fraction of DNA profiles are collected and stored from persons of Asian descent.

Our estimation data paints a similar, though still more complicated picture. We show that although White people make up 62% of the total U.S. population, they make up only 49% of the disclosed DNA database. In comparison, although Black people make up only 13% of the U.S. population, they contribute 34% of samples to the disclosed DNA database. Put simply, DNA has been collected from Black persons at two and a half times the rate of White persons and from Native Americans at one and a half times the rate of White persons. And although people of color bear a disproportionate burden of DNA collection and storage, the burden is particularly concentrated on the Black population. We estimate that

12.5 (FBI 2020), https://www.fbi.gov/file-repository/quality-assurance-standards-for-forensic-dna-testing-laboratories.pdf/view [https://perma.cc/2426-JAFY].

9. *See generally* MURPHY, *supra* note 4, at 153–241 (describing evolution of law and policy across a range of DNA collection, testing, and searching practices).

10. *See infra* Part I.B.

11. *See infra* Part I.B.

2.26% of Black people have their DNA collected per year, whereas only 0.69% of the Hispanic population and 0.12% of the Asian population are subjected to DNA collection annually.

Of course, these efforts were inherently complicated by many factors, not least of which is that DNA databases are not static. The earliest databases often contained DNA profiles from only convicted felons, whereas today, DNA databases commonly include profiles from misdemeanants and arrestees.[12] In addition, compulsory collection has not always been enforced in practice. The logistical and financial challenge of processing an enormous number of genetic samples has led both to collection failures and to periods of backlog.[13] Other limitations are described in the sections that follow. Recognizing these limitations, our model helps fill the total vacuum of information that existed previously.

Part I reviews the existing literature and illustrates both that hard data on the demographic composition of databases is lacking and that there is a broad and longstanding interest in such information. Part II fills that gap in two ways. First, we report the results of our nationwide request to disclose such information. And second, we show estimates generated from a model that endeavors to reverse-engineer the composition of DNA databases based on publicly available information about state policies and practices. Part III explains the importance of this data to debates about forensic DNA policies.

I.
EXISTING REFERENCES TO DATABASE COMPOSITION

Interest in the demographic composition of forensic DNA databases spans back to the earliest days of their adoption. As time has passed, and genetic evidence has become a more central part of the criminal justice system, these concerns have only magnified. As this Section shows, references to the existence of possible racial or ethnic disparities in DNA databases regularly surface both in the scholarly literature and in judicial opinions across a wide array of topics.

None of those sources, however, include precise data on the database's actual demographic composition, or even a detailed estimate of that composition. Most simply assume that DNA databases reflect the existing disparities in the criminal justice system, relying on a common-sense intuition that the databases replicate documented disparities in conviction and arrest rates. This Section reviews the publicly available data about the DNA database system, as well as questions about the database's racial or ethnic composition.

12. *See* MURPHY, *supra* note 4, at 156–57 (summarizing current DNA collection laws).

13. *See, e.g.*, Kerry Abrams & Brandon L. Garrett, *DNA and Distrust*, 91 NOTRE DAME L. REV. 757, 778–79, 784 (2015) (describing problem of backlogs and failure to collect and test samples).

A. Published Data

Law enforcement in the United States releases shockingly little information about the national DNA database system. The only consistently released information is provided through a webpage managed by the FBI entitled "NDIS Statistics."[14]

The contents of the page have varied slightly over time, but generally they provide a precise number of: 1) convicted persons profiles; 2) arrestee profiles; 3) forensic profiles; 4) "investigations aided"; and 5) the number of NDIS participating labs.[15] This information is provided both as an aggregate national number and by each state.[16] The "investigations aided" figure encompasses both a raw number of "hits" (i.e., associations made within the database, either forensic-to-known or forensic-to-forensic) and the number of investigations such hits informed.[17] There is no data about the type of case in which the "hit" occurred, the person's characteristics (or even the offense that qualified the person for database inclusion), the outcome of the match (e.g., whether the match led to an investigation or identification of an actual probable perpetrator), or the criminal justice result of the match (including whether an arrest or conviction occurred). There is also no aggregate information about the demographic characteristics of either the forensic or known profiles—such as the fraction of forensic profiles by offense type or known profiles by qualifying offense.

Some of this lack of information may be ascribed to the early decision to exclude demographic and other data from the central database, a decision largely grounded in concerns related to the privacy of the known person and the security of the database overall.[18] But those concerns alone cannot explain the lack of information. States are permitted to make their own decisions about which data to track. This means, even following a decentralization model, the FBI could have nonetheless collected or tracked such information—or required states to track and

14. *Sept. 2018 NDIS Statistics*, *supra* note 3.

15. *See id.*

16. *See id.*

17. *See id.* A forensic-to-known match is a match between a crime scene sample and a known person's profile—what most people think of when they hear "DNA match." A forensic-to-forensic match is a match among samples from different crime scenes. Thus, for instance, a match might occur between profiles from DNA samples left by the perpetrator at two different bank robberies. Even if law enforcement is not able to match that profile to a known person, it is still useful to know that the same individual committed both robberies.

18. *See* FBI, NATIONAL DNA INDEX SYSTEM (NDIS) OPERATIONAL PROCEDURES MANUAL 5 (2020) https://www.fbi.gov/file-repository/ndis-operational-procedures-manual.pdf/view [https://perma.cc/TX27-GHPV] [hereinafter NDIS OPERATIONAL PROCEDURES MANUAL] ("In the early 1990s when the initial version of the CODIS software was being developed, the FBI Laboratory convened a group of privacy advocates to obtain feedback on its plans for this new law enforcement tool. Among the recommendations was the suggestion that, to protect the privacy of persons providing the DNA samples, that no personally identifying information be databased. This recommendation was incorporated into the CODIS software and the implementation of the National DNA Index and remains in effect today."); *see also* MURPHY, *supra* note 4, at 15–16 (explaining that national DNA database files are separated from identifying information, largely because of expedience and privacy concerns).

report it—as part of the Memorandum of Understanding (MOU) that each state signs when they join the national system.[19] The MOU already imposes a host of other informational requirements. For instance, in order to upload and access data in the national database, the laboratory agrees to maintain a case file with specific pieces of information in it.[20] While the FBI does not maintain that file, it does conduct regular audits to ensure compliance with the MOU and issues reports on the findings.[21] Moreover, the FBI has updated the software for the database system numerous times since its launch in 1998 and thus could have chosen to add fields if it wished to collect that data.[22] In short, the absence of information beyond the skeletal data provided is a choice, not an inevitable byproduct of operating a complex and decentralized system.

In fact, there is at least one exception to the general lack of public information about DNA database composition or efficacy. As of 2009, the Forensic Sciences Division of the Maryland State Police has published an annual report on the state database,[23] as required by law.[24] The reporting requirement was included as part of the legislature's enactment of a statute that permits DNA collection from arrested persons and is one of several mandates intended to gauge

19. *See, e.g.*, AUDIT DIVISION, OFF. OF THE INSPECTOR GEN., U.S. DEP'T OF JUSTICE, COMPLIANCE WITH STANDARDS GOVERNING COMBINED DNA INDEX SYSTEM ACTIVITIES AT THE HOUSTON POLICE DEPARTMENT CRIME LABORATORY 5 (2010), https://www.oversight.gov/sites/default/files/oig-reports/g6010009.pdf [https://perma.cc/RV9D-LLXV] (noting that a laboratory must sign an MOU before participating in NDIS, which "defines the responsibilities of each party").

20. *See generally* QUALITY ASSURANCE STANDARDS FOR DNA DATABASING LABORATORIES stand. 11 (FBI 2020), https://www.fbi.gov/file-repository/quality-assurance-standards-for-dna-databasing-laboratories.pdf/view [https://perma.cc/6BX4-9YXE]; *see also* NDIS OPERATIONAL PROCEDURES MANUAL, *supra* note 18, at 12–15 (detailing the requirements necessary for laboratories to upload to NDIS, including agreement to the NDIS MOU).

21. *See Combined DNA Index System Audits*, OFF. OF INSPECTOR GEN., U.S. DEP'T OF JUSTICE, https://oig.justice.gov/reports/codis-ext.htm [https://perma.cc/3G6R-7R5V]; *see also* MURPHY, *supra* note 4, at 139–41 (explaining how the audit process has revealed a wide range of compliance, including an average 6% error rate in uploading unauthorized DNA profiles across the labs audited since 2010).

22. *See, e.g.*, JOHN M. BUTLER, ADVANCED TOPICS IN FORENSIC DNA TYPING: METHODOLOGY 223 (2011) ("Software versions are updated periodically and provided to all CODIS laboratories by the FBI."); NDIS OPERATIONAL PROCEDURES MANUAL, *supra* note 18, at 51–52 (noting that newest version of software will include the capacity to do real-time "[r]apid" DNA searches).

23. *Forensic Sciences Reports*, MD. ST. POLICE, https://mdsp.maryland.gov/Organization/Pages/CriminalInvestigationBureau/ForensicSciencesDivision/ForensicSciencesReports.aspx [https://perma.cc/L7D4-9TMB].

24. *See* MD. CODE REGS. 29.05.01.16 (2020). That section prescribes the specific content of the report and includes:

> (3) Individual Data and Analysis. The Department of State Police shall include in the annual report, for the preceding calendar year, the racial demographics of all individuals who have been charged with qualifying crimes upon arrest in the following categories:
>
> (a) Asian;
>
> (b) African-American;
>
> (c) White;
>
> (d) Hispanic; or
>
> (e) Other.

the efficacy and equality of expanding DNA collection to arrestees. Thus, the required demographic data are limited to that category. But the Maryland report also includes fairly detailed information about the type and outcome of matches, such as offense types, whether criminal charges were filed, and whether conviction resulted.[25] Other jurisdictions also issue annual reports, but none contain demographic data or this kind of granular outcome data; most focus on expenditures or generalized match figures.[26]

Other models for greater transparency can be found outside of the United States. The Home Office in the United Kingdom produces a regular report (originally annually and now biennially) about the National DNA Database (NDNAD), in addition to a number of other informational documents.[27] As in the United States, that report includes the total number and type of profiles held, but it also provides a wealth of additional information. The report includes the number of forensic profiles by crime type;[28] detailed information about match rates and crime types;[29] and error rates by category.[30] Most importantly for purposes of this Article, it includes the sex, age, and ethnicity of databased

25. *See Forensic Sciences Reports*, *supra* note 23.

26. California, for example, publishes expenditure information as well as some efficacy information in an annual report. *See* DNA DATABASE AND DATA BANK PROGRAM, CAL. DEP'T OF JUSTICE, ANNUAL DNA IDENTIFICATION FUND REPORT FOR CALENDAR YEAR 2016 (2017), https://oag.ca.gov/sites/all/files/agweb/pdfs/bfs/2016-dna-fund-report.pdf [https://perma.cc/VU5U-6F99]; DIV. OF LAW ENF'T, CAL. DEP'T OF JUSTICE, ANNUAL REPORT: FISCAL YEAR 2013-2014 (2014), https://oag.ca.gov/sites/all/files/agweb/pdfs/publications/dle-annual-report-2013-14.pdf [https://perma.cc/YS7N-MPH3]. New York publishes collection rates by county. *See* N.Y. STATE DIV. OF CRIMINAL JUSTICE SERVS., DNA COLLECTION RATES BY COUNTY (2020) https://www.criminaljustice.ny.gov/crimnet/ojsa/DNA_Rates_Statewide.pdf [https://perma.cc/C32U-3TMM].

27. *See National DNA Database Documents*, HOME OFFICE (May 20, 2013), https://www.gov.uk/government/collections/dna-database-documents [https://perma.cc/5FP6-AQWN]. The national DNA database in the United Kingdom was previously maintained by the National DNA Database Strategy Board, but that entity assumed responsibility for the national fingerprint database as well and is now called the Forensic Information Database Strategy Board. Database statistics are available at *National DNA Database Statistics*, HOME OFFICE (Apr. 23, 2013), https://www.gov.uk/government/statistics/national-dna-database-statistics [https://perma.cc/ZVD9-445P]. As of the first quarter of 2020-21, the U.K.'s database contained roughly 5.6 million known person profiles and roughly 650,000 forensic profiles. HOME OFFICE, NDNAD STATISTICS, AS OF 30TH JUNE 2020 (2020), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/900981/NDNAD_Website_statistics_Q1_20-21.ods [https://perma.cc/59YP-TXY9].

28. HOME OFFICE, NATIONAL DNA DATABASE STRATEGY BOARD BIENNIAL REPORT 2018 - 2020, at 13 tbl.1 (2020), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/913011/NDNAD_Strategy_Board_AR_2018-2020_Web_Accessible.pdf [https://perma.cc/763X-LMYT].

29. *Id.* at 20–26. The data reflect match rates from both routine and emergency searches and show that routine matches occur most in burglary and vehicle cases. They report how often a search is successful for each crime type, how many searches were done, and matches made in absolute numbers.

30. *Id.* at 35–36 tbls.5, 6. The four reported categories of error are 1) a sample or record handling error by police personnel; 2) a sample or record handling error by lab personnel; 3) interpretation errors; and 4) transcription errors.

persons.[31] Thus, for instance, as of the 2018-2020 report, the U.K. database held between five and six million known persons,[32] with characteristics as follows (the parenthetical after each group provides the corresponding percentage from the 2011 general census in England and Wales[33]):

80.4% Male 19.1% Female	75.5% White/North-European (*86%)*[34] 2.3% White/South-European (") 7.5% Black (*3.3%)* 5.3% Asian (*7.5%)* 0.8% Middle-Eastern *(0.4%)* 0.6% Chinese, Japanese, or SE Asian (*7.5*%) 8.0% Unknown[35]

B. Scholarly and Judicial Assumptions About Database Composition

1. Compulsory Collection Laws

The earliest references to the demographic composition of DNA databases appeared with the first series of debates about them. As scholars, advocates, and courts grappled with the proper scope of compulsory DNA collection laws, they raised concerns that DNA databases would amplify the disparities already evident in the criminal justice system.[36] There was also concern that linking race and genetics ran the risk of reviving ugly debates about the biological origins of crime.[37] Fears about racial disparity also surfaced in connection with pragmatic

31. *Id.* at 16–19.

32. The database actually contains 6,568,035 known person records, but it is estimated that 14.7% are duplicates, and thus it actually contains 5,491,832 known persons. *Id.* at 10.

33. *See* OFFICE FOR NAT'L STATISTICS, ETHNICITY AND NATIONAL IDENTITY IN ENGLAND AND WALES 2011, at 3–4 (2012), https://webarchive.nationalarchives.gov.uk/20160105213319/http://www.ons.gov.uk/ons/dcp171776_290558.pdf [https://perma.cc/EL42-RP2N] (80.5% of the population was of White British descent, 7.5% was of "Asian" descent (Pakistani, Indian, Bangladeshi, Chinese, or other); 3.3% was Black; 2.2% was multiple race or ethnicities; and 1.0% was "Other," which includes 0.4% Arab). Critics have charged that the collected DNA data categories ought to mirror the census data categories. *See, e.g.*, David Skinner, *'The NDNAD Has No Ability in Itself To Be Discriminatory': Ethnicity and the Governance of the UK National DNA Database*, 47 SOC. 976, 981, 985 (2013) ("Anyone attempting to generate such estimates must reconcile two different datasets – one detailing the ethnic composition of the NDNAD and the other drawn from census-based projections of the current ethnic minority population of England and Wales.").

34. The census does not distinguish between Northern and Southern Europeans. Both groups are subsumed within the 86% statistic.

35. HOME OFFICE, *supra* note 28, at 16–17.

36. *See, e.g.*, Jill C. Schaefer, Comment, *Profiling at the Cellular Level: The Future of the New York State DNA Databanks*, 14 ALB. L.J. SCI. & TECH. 559, 578–79 (2004).

37. *See generally* Christian B. Sundquist, *The Technologies of Race: Big Data, Privacy and the New Racial Bioethics*, 27 ANNALS HEALTH L. 205, 205 (2018) (canvassing ways in which DNA technologies "threaten[] a disturbing return of nineteenth century 'race science'").

arguments for expungement and other provisions to limit the scope of DNA databases.[38]

These debates resurfaced a decade or so later as states began to expand their compulsory collection laws to include arrested persons, which the Supreme Court in *Maryland v. King* ultimately ruled constitutional.[39] In the words of one scholar discussing the *King* case, "The most compelling reason not to draw the line at arrestees who are never convicted, given the meager incremental benefits in doing so, is the profound racial inequity it creates in the makeup of the databases."[40] Other critics assailed arrestee DNA sampling on racial and equitable grounds as well.[41]

Indeed, scholars of forensic DNA cited racial and ethnic disparities as a major argument in support of a *universal* DNA collection policy, and the *New York Times* even published an op-ed stating as much.[42] The earliest advocates of universal collection laws were D.H. Kaye and Michael Smith, who published an article "question[ing] the rationales for drawing the line [of compulsory collection] at all convicted [persons]—which is fast becoming standard practice—or at all arrestees—which may be where we are headed."[43] Asserting that "[t]here can be no doubt that any database of DNA profiles will be dramatically skewed by race if the sampling and typing of DNA becomes a routine consequence of criminal conviction," they concluded that "we are fast producing a racially distorted system in which, however lawfully the DNA samples are taken, they are taken disproportionately from members of racial

38. *See, e.g.*, Peter A. Chow-White & Troy Duster, *Do Health and Forensic DNA Databases Increase Racial Disparities?*, PLOS MED. (Oct. 4, 2011), https://journals.plos.org/plosmedicine/article/file?id=10.1371/journal.pmed.1001100&type=printable [https://perma.cc/FV6X-ZC9S]; Valerie Werse, Note, *A "Lengthy, Uncertain, and Expensive Process": A Comparison of Types of Expungement from DNA Databases of Arrestees*, 39 RUTGERS COMPUTER & TECH. L.J. 282, 310–11 (2013).

39. *See* 569 U.S. 435, 465–66 (2013).

40. Andrea Roth, Maryland v. King *and the Wonderful, Horrible DNA Revolution in Law Enforcement*, 11 OHIO ST. J. CRIM. L. 295, 308 (2013).

41. *See, e.g.*, SHELDON KRIMSKY & TANIA SIMONCELLI, GENETIC JUSTICE: DNA DATA BANKS, CRIMINAL INVESTIGATIONS, AND CIVIL LIBERTIES 252–74 (2011); Erin Murphy, *License, Registration, Cheek Swab: DNA Testing and the Divided Court*, 127 HARV. L. REV. 161, 181–83, 188–91 (2013); Michael T. Risher, *Racial Disparities in Databanking of DNA Profiles*, *in* RACE AND THE GENETIC REVOLUTION: SCIENCE, MYTH, AND CULTURE 47 (Sheldon Krimsky & Kathleen Sloan, eds., 2011); John D. Biancamano, Note, *Arresting DNA: The Evolving Nature of DNA Collection Statutes and Their Fourth Amendment Justifications*, 70 OHIO ST. L.J. 619, 650–51 (2009).

42. Michael Seringhaus, Opinion, *To Stop Crime, Share Your Genes*, N.Y. TIMES (Mar. 14, 2010), https://www.nytimes.com/2010/03/15/opinion/15seringhaus.html [https://perma.cc/TUP5-7GB9] ("[T]he national DNA database is racially skewed, as [B]lacks and Hispanics are far more likely than [W]hites to be convicted of crimes.").

43. D.H. Kaye & Michael E. Smith, *DNA Identification Databases: Legality, Legitimacy, and the Case for Population-wide Coverage*, 2003 WIS. L. REV. 413, 414–15; *see also* Paul M. Monteleoni, Note, *DNA Databases, Universality, and the Fourth Amendment*, 82 N.Y.U. L. REV. 247, 278 (2007) ("[A] universal database would represent all racial groups proportionately, in contrast to a database created by arrest or conviction.").

minorities."[44] They worried that "such coverage . . . exacerbates racial tensions and undermines the preventative and investigative value of the databases," and instead proposed a universal DNA database.[45] Scholars opposed to universal databases did not contest that existing databases were racially skewed, only that a universal database would in fact be "race neutral."[46]

Another strand of commentary on compulsory collection laws argued that the racial-disparity claim should be evaluated in light of gender and other intersectional traits. Assuming the disparity of the databases, one author observed that "databank expansion can be used as a tool to draw attention to the problem of violence against Black women by Black men," since "Black men perpetrate 85.8% of rapes and sexual assaults of Black women, [and thus] more perpetrators of sexual violence against Black women will be identified."[47] In this respect, a racially skewed database might be laudatory, in that it would solve more of the crimes committed against women of that same ethnicity or race.

By and large, the debates over compulsory collection simply accepted that the racial composition of the database was skewed. For instance, the *New York Times* op-ed speculated that "the database could *approach universal population coverage* for certain races or groups and not others."[48] Other scholars provided a web of citations, most of which led back to the same source. Typical is one article in the *Nation* magazine, which referred to a statement by Jeremy Gruber, the executive director of the Council for Responsible Genetics. According to the article, Gruber stated, "By 2011, African-Americans made up 40 percent of the Combined DNA Index System (CODIS)"—but no information was given about how that precise figure was calculated.[49]

The "40 percent" figure surfaces repeatedly in the literature, at times without direct attribution, at times citing derivative sources.[50] It seems most

---

44. Kaye & Smith, *supra* note 43, at 415, 452.

45. *Id.* at 415.

46. *See, e.g.*, Tania Simoncelli, *Dangerous Excursions: The Case Against Expanding Forensic DNA Databases to Innocent Persons*, 34 J.L. MED. & ETHICS 390, 395 (2006) ("To the question of whether a universal database will help correct racial distortions, it is important to recognize that racism is not simply a symptom of DNA databases, but is systemic to our criminal justice system. . . . These patterns of racial disparity mean that our DNA databases are also racially skewed. But placing everyone in the database will not result in a more 'race neutral' system, because the makeup of the database has no bearing on who is targeted for suspicion and arrest. Even if everyone is in the database, the majority of hits will continue to identify minorities, as long as the types of crime, neighborhoods and populations monitored and investigated are racially driven.").

47. Marie-Amélie George, Note, *Gendered Crime, Raced Justice: A Critical Race Feminist Approach to Forensic DNA Databank Expansion*, 19 NAT'L BLACK L.J. 78, 102, 103 (2005).

48. Seringhaus, *supra* note 42 (emphasis added).

49. *See* Jason Silverstein, *The Dark Side of DNA Evidence,* THE NATION (Mar. 27, 2013), https://www.thenation.com/article/dark-side-dna-evidence/ [https://perma.cc/5KYM-YVA8].

50. *See, e.g.*, Brett Mares, *A Chip off the Old Block: Familial DNA Searches and the African American Community*, 29 L. & INEQ. 395, 408 (2011) ("Unsurprisingly, African American DNA profiles constitute an incongruent proportion—roughly forty percent—of DNA databases at the state and national levels." (citing Grimm, *infra* note 60, at 1176)).

likely that this commonly cited number traces back to a 2006 article. That article arrived at the "40 percent" figure as follows:

> African-Americans constitute about thirteen percent of the U.S. population, or about thirty-eight million people. In an average year, over forty percent of people convicted of felonies in the United States are African-American. As a result, the set of individuals in the Offender Index is not racially neutral with regard to the American population. Although we have not been able to find confirmation of this, we assume, based on the felony conviction statistics, that African-Americans make up at least forty percent of the CODIS Offender Index, or roughly 1.1 million people out of 2.75 million.[51]

That estimate is subject to a range of critiques, most prominently that individual state collection policies vary—there is no single "all felony" collection mandate. But it has stood for decades as the most viable proxy for more nuanced estimates.

2. Familial Searches

Questions about the racial composition of DNA databases have arisen not just in regard to DNA collection policies, but also with regard to DNA search policies. As originally conceived, DNA databases helped solve crime by matching unsolved forensic samples to known persons or by matching unsolved crimes to one another in a pattern that might point to particular suspects.[52] To justify the compulsory collection of DNA from convicted persons, law enforcement analogized DNA to a high-tech fingerprint and made the case that the biometric identifiers of persons who had violated the law were fairly preserved and used by police to solve not just present but future crimes.[53]

But once DNA databases gained in size, law enforcement realized they had even more crime-solving potential. If a search for an exact DNA match failed, then police could use the DNA database as a suspect-generating system.[54] Because two people who are biologically related are likely to share genetic material, police could search DNA databases not just for profiles that exactly matched samples left at crime scenes, but also for ones with a close resemblance. In this way, law enforcement uses a DNA database not to find an *exact* match, but rather to find some *near misses*—leads of known persons in the database who might be biologically related to the actual perpetrator.

Familial searches are controversial for two reasons. First, they represent a dramatic expansion of DNA databases because they make databases a resource

---

51. Henry T. Greely et al., *Family Ties: The Use of DNA Offender Databases to Catch Offenders' Kin*, 34 J.L. MED. & ETHICS 248, 258 (2006).

52. *See Frequently Asked Questions on CODIS and NDIS*, *supra* note 5 (answering "How do these DNA databases using CODIS work?" by referencing known-to-forensic or forensic-to-forensic matches).

53. MURPHY, *supra* note 4, at 157–59.

54. *Id.* at 191–93 (describing familial search process).

for locating and putting under suspicion people not already in the database.[55] Effectively, they turn all relatives of persons in DNA databases into suspects. Second, and most pertinently for this Article, scholars argue that the "genetic surveillance" enabled by familial searches would be concentrated on particular demographic populations.[56] Specifically, if DNA databases reflect demographic disparities in the criminal justice system, and thus are racially skewed, then those populations will shoulder an unfair burden of suspicion.

Accordingly, much of the debate over the propriety and constitutionality of familial searches centered on their likely disparate effect. In the earliest years of familial searching, one group of scholars led by Hank Greely attempted to assess the reach of such searches into particular communities. Their calculations, which they admit are "simplified" out of necessity:

> Assume first that family structures are the same for African-Americans and for non-Hispanic U.S. Caucasians in the CODIS Offender Index. Assume further that the average person in the database has five living first degree relatives. . . . Under these assumptions, the 1.1 million African Americans in the Offender Index will have 5.5 million first degree relatives, leading to a total of 6.6 million African-Americans "findable" through the database – the offenders and their relatives. That constitutes about seventeen percent of all African-Americans. U.S. Caucasians (including non-African-American Hispanics) make up about sixty percent of the Offender Index or currently about 1.65 million people. They would have 8.25 million first degree relatives, for total coverage of 9.9 million people "findable" through the database. U.S. Caucasians, including non-African-American Hispanics, constitute about eighty-three percent of the American population or about 247 million people. The 9.9 million U.S. Caucasians who would be either in the Offender Index, or a first degree relative of someone in the Index would make up just four percent of the white population. Thus, more than four times as much of the African-American population as the U.S. Caucasian population would be "under surveillance" as a result of family forensic DNA and the vast majority of those people would be relatives of offenders, not offenders themselves. (If non-African-American Hispanics were analyzed separately from non-Hispanic U.S. Caucasians, the disproportion between African-Americans and U.S. Caucasians would be even greater.)[57]

55. *See id.* at 206–07.
56. *See id.* at 207.
57. Greely et al., *supra* note 51, at 259.

The Greely et al. estimate—and in particular the 17% figure—has been cited innumerable times in the scholarly and forensic science literature,[58] as well as in the popular press.[59]

A student Note took a slightly different approach in its attempt to quantify the effects of familial DNA searches for the Hispanic population.[60] Building from Greely's estimate, the Note included incarceration rates, population growth, and criminal justice trends for the Hispanic community as well as in-depth reports on family structure and reproductive patterns. Crafting a formula in which:

> X_n represents generations of family members; S is the number of original relatives; C is the number of persons eventually convicted of the crime at issue for each group; A is the average number of children under the age of eighteen for the given demographic group; [and] H_n represents the number of additional potential partial allele hits attributed to a new generation,[61]

the author concluded that "more members of the Hispanic community than the African American and white communities will be subjected to investigation following a given CODIS search."[62] Specifically, "Hispanics were exposed to a risk of surveillance approximately 3% higher than whites and 2% higher than African Americans after two generations. After three generations, the difference increased to 5% more Hispanics than African Americans placed at risk, and 8% more Hispanics than whites."[63]

Other arguments over familial searches simply stated the concern that familial searches would have a particularly adverse effect on people of color, without attempting to quantify it concretely. Typical is one scholar, who observed that "[t]hese disparities in conviction and arrest rates may also be represented in the racial composition of DNA data banks because the vast majority of profiles come from convicted offenders and arrestees."[64] Similar expressions of concern were also found in popular media. One such opinion piece argued that:

---

58. *See, e.g.*, Roth, *supra* note 40, at 308; Joyce Kim et al., *Policy Implications for Familial Searching*, at 6, INVESTIGATIVE GENETICS (Nov. 1, 2011), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3253037/pdf/2041-2223-2-22.pdf [https://perma.cc/T5UM-RFXT].

59. *See, e.g.*, Jeffrey Rosen, *Genetic Surveillance for All*, SLATE (Mar. 17, 2009) https://slate.com/news-and-politics/2009/03/genetic-surveillance-for-all.html [https://perma.cc/R3F8-88R2]; Ellen Nakashima, *From DNA of Family, a Tool to Make Arrests*, WASH. POST, Apr. 21, 2008, at A.1.

60. *See* Daniel J. Grimm, Note, *The Demographics of Genetic Surveillance: Familial DNA Testing and the Hispanic Community*, 107 COLUM. L. REV. 1164, 1180 (2007).

61. *Id.* at 1180–81.

62. *Id.* at 1182.

63. *Id.* at 1182–83. The author's methodology relied upon data about the average number of children in existing Hispanic households, and then confirmed the findings using another approach that substituted that data with fertility rates.

64. Sonia M. Suter, *All in the Family: Privacy and DNA Familial Searching*, 23 HARV. J. L. & TECH. 309, 369 (2010). *See also* Mares, *supra* note 50, at 397 (contemplating the "constitutionality of familial DNA testing" in light of the "disproportionate effect it will likely have on African Americans").

> [P]artial matching and familial searching will greatly aggravate the racial inequality already embedded in offender-based DNA databases. Certain racial and ethnic populations are already overrepresented in CODIS, owing simply to the reality of crime statistics. But implicitly expanding database coverage to include relatives will grossly and unfairly amplify this bias. By applying partial matching and familial searching to a database that includes anyone ever arrested—which seems to be where the system is headed—CODIS could one day approach universal coverage for some races and not for others.[65]

Familial searches highlighted another issue: the breakdown of DNA databases by sex. Familial search techniques rely on a second stage of testing that examines the Y, or male, chromosome in order to winnow the list of leads to manageable levels.[66] As a result, it cannot be used to find female perpetrators or perpetrators whose only relative in the database is a female. While this issue has occasioned some academic and popular interest,[67] it has in no way generated the degree of concern as its racial and ethnic counterpart.

3. Recreational and Genealogical Database Searches

Law enforcement's use of recreational genetic databases—such as the popular online platform FamilyTreeDNA—to conduct familial searches has offered an intriguing counterpoint to debates over the racially disparate impact of forensic DNA practices. Like familial searches, recreational or genealogical database searches (also called "long-range familial searches") rely upon principles of inheritance to pinpoint persons in the database who are not the actual perpetrator, but are a potential relative of the perpetrator. Because these searches use hundreds of thousands of single nucleotide polymorphisms (SNPs), rather than the thirteen to twenty forensic short tandem repeats (STRs) of a classic forensic DNA profile, they have far greater power than familial searches to

---

65. Natalie Ram & Michael Seringhaus, *O Brother, Where Art Thou?*, SLATE (June 14, 2010), http://www.slate.com/articles/news_and_politics/jurisprudence/2010/06/o_brother_where_art_thou.html [https://perma.cc/F5XC-F5JK]; *see also* Eli Rosenberg, *Family DNA Searches Seen as Crime-solving Tool, and Intrusion on Rights*, N.Y. TIMES (Jan. 27, 2017), https://www.nytimes.com/2017/01/27/nyregion/familial-dna-searching-karina-vetrano.html [https://perma.cc/5K3R-ZTAR] (equating familial searches to the creation of a "database of suspects largely defined by their race and class").

66. *See, e.g.*, EMILY NIEDZWIECKI ET AL., ICF INT'L, UNDERSTANDING FAMILIAL DNA SEARCHING: COMING TO A CONSENSUS ON TERMINOLOGY 4 (2016), https://www.ncjrs.gov/pdffiles1/nij/grants/251080.pdf [https://perma.cc/Q7KJ-R8BG] ("Because the majority of samples profiled are from males, the most common form of lineage testing is Y-STR analysis . . . .").

67. *See, e.g.*, Mary McCarthy, Note, *Am I My Brother's Keeper?: Familial DNA Searches in the Twenty-first Century*, 86 NOTRE DAME L. REV. 381, 402 (2011) ("[I]f Y-chromosome testing is increasingly used, males may be disproportionately represented in DNA databases and partial match searches. But since the majority of criminal perpetrators are male, increased use of Y-chromosome testing might not have a great effect." (footnotes omitted)); Richard Williams, *DNA: All in the Family*, STATE LEGISLATORS, June 2019, at 14 (noting that a familial search "works only for men").

uncover leads.[68] However, like familial searches they are ultimately still a tool for generating leads and crafting suspect pools, not for finding an exact match to the crime scene sample.

In contrast to law enforcement DNA databases, the composition of recreational and commercial databases skews heavily White.[69] In one recent paper, researchers estimated the success rate of long-range familial searches using a dataset of roughly 1.28 million profiles voluntarily submitted to the MyHeritage database.[70] Of those persons, roughly 75% were of Northern European descent.[71] As a result, they estimated that "[i]ndividuals of primarily North European background were 30% more likely to have a [highly probative] match than individuals whose genetic background was primarily from sub-Saharan Africa."[72] In fact, the authors estimated:

> [W]ith a database size of ~3 million US individuals of European descent (2% of the adults of this population), more than 99% of the people of this ethnicity would have at least a single third-cousin match and more than 65% are expected to have at least one second-cousin match. With the exponential growth of consumer genomics, we posit that such a database scale is foreseeable . . . in the near future.[73]

In short, "[l]ong-range familial searches create racial disparity that is the opposite of disparities documented in traditional forensic databases."[74]

Considering these data alongside assumptions about the racial disparities of forensic databases have led scholars to argue that police access to recreational genetics may "help to remedy the racial and ethnic disparities that plague traditional forensic searches."[75]

---

68. Ellen M. Greytak et al., *Genetic Genealogy for Cold Case and Active Investigations*, 299 FORENSIC SCI. INT'L 103, 103–04 (2019).

69. Yaniv Erlich et al., *Identity Inference of Genomic Data Using Long-range Familial Searches*, 362 SCIENCE 690, 690 (2018); *see also* Michael D. Edge & Graham Coop, *Attacks on Genetic Privacy Via Uploads to Genealogical Databases*, at 1, 14, ELIFE (Jan. 7, 2020), https://doi.org/10.7554/eLife.51810 [https://perma.cc/NT6X-49XV] ("DTC genetics companies generally do not release this kind of information on their users, but their research papers suggest that they have access to especially large samples with European ancestries—for example, a 23andMe paper on demography in the United States included almost 150,000 self-described European Americans and less than 10,000 each of self-described African Americans and Latino Americans. For a qualitatively similar sample composition in a study from Ancestry, see *Han et al.* (*2017*)." (citation omitted)).

70. Erlich et al., *supra* note 69, at 1.

71. *Id.*

72. *Id.*

73. *Id.* (citation omitted).

74. *Id.*

75. Natalie Ram et al., *Genealogy Databases and the Future of Criminal Investigation,* 360 SCIENCE 1078, 1078–79 (2018). *But see* Erin Murphy, *Law and Policy Oversight of Familial Searches in Recreational Genealogy Databases*, FORENSIC SCI. INT'L, Nov. 2018, at e5, e7 (pondering whether "the prevalence of genealogical DNA database searches will instead begin to infect the debate about the use of government databases, and prompt the loosening of existing regulations rather than the enhancement of the regulatory architecture for genealogical searches").

4. Rogue Databases

The racial composition of DNA databases is also a salient issue in assessing the policy and practice of "rogue" databasing. This phrase captures a number of different kinds of informal, non-CODIS database systems. In each instance, a law enforcement or district attorney's office creates a database of profiles that do not qualify for inclusion in the state or national database. This is typically because the quality of the test results was too low, the sample was insufficiently linked to the perpetrator of the offense, or the state law does not explicitly authorize collection from that individual.[76] These databases may have varying levels of sophistication: some localities rely on commercial platforms that mimic the CODIS database whereas others may be as basic as an Excel spreadsheet.[77] Rogue databases also have varying degrees of legal status. In some jurisdictions, a municipality may pass a local law authorizing the creation of the database whereas in others they are informal and unregulated.[78]

A series of issues arise regarding rogue database practices, including issues of quality control, security, and abuse. Another concern is that police may more readily coerce samples from people of color or engage in surreptitious collection of DNA from communities of color and place those samples in unregulated databases lacking any meaningful oversight. As one author remarks:

> While local databases have the potential to mitigate some of the racial inequities in the criminal justice system by replacing police reliance on intuition and hunches with more reliable investigative leads based on DNA evidence, local databases increase distributional inequities because local police have total discretion about who to target for inclusion in these databases. This has resulted in police seeking out the "usual suspects"—poor people of color—to secure DNA samples for these databases.
>
> . . . .
>
> . . . [O]n balance, local databases will contribute to the disproportionate burdens people of color face in the criminal justice system.[79]

76. *See, e.g.*, Jan Ransom & Ashley Southall, *N.Y.P.D. Detective Gave a Boy, 12, a Soda. He Landed in a DNA Database*, N.Y. TIMES (Aug. 15, 2019), https://www.nytimes.com/2019/08/15/nyregion/nypd-dna-database.html [https://perma.cc/68YA-XT29] (documenting collection of DNA by New York law enforcement after stop and release).

77. MURPHY, supra note 4, at 181–88 (describing different degrees of sophistication of rogue databasing).

78. *See, e.g.*, Andrea Roth, *"Spit and Acquit": Prosecutors as Surveillance Entrepreneurs*, 107 CALIF. L. REV. 405, 417-32 (2019) (describing legal regime surrounding Orange County database).

79. Jason Kreag, *Going Local: The Fragmentation of Genetic Surveillance*, 95 B.U. L. REV. 1491, 1497–1524 (2015) (footnote omitted); *see also* Elina Treyger, *Collateral Incentives to Arrest*, 63 KAN. L. REV. 557, 558 (2015) (identifying DNA collection authority as "collateral incentives to arrest" that may be abused by police).

5. Statistical Analysis of a DNA Match

Finally, questions about the racial composition of DNA databases have arisen around the statistical probability of a DNA match. A DNA match statistic reports the probability that an erroneous match between a defendant's DNA profile and crime scene evidence occurred by chance. The preferred means for reporting matches have greatly preoccupied scholars and legal actors alike and even occasioned a series of conflicting reports from blue-ribbon committees of the National Research Council.[80] The full account of this debate distracts from the central points of this Article, but a brief sketch merits recitation.

First, in the early days of DNA typing, population geneticists set forth a series of tables that aimed to quantify the frequency of the genetic traits (i.e., alleles) used for forensic DNA typing.[81] Fittingly, these were known as the "allele-frequency tables," and they were used as the baseline data for calculating the significance of a match between crime scene evidence and a known person's profile. In other words, analysts would determine the twenty-six alleles in the known profile, see that they matched the twenty-six alleles in the crime scene evidence profile, and then issue a report indicating the probability that such a match occurred by chance rather than because the defendant left the genetic material.

In the 1990s, as forensic DNA typing was starting to take root, these frequencies were reported for broad racial or ethnic categories like "White" or "Black" or "Hispanic."[82] The theory behind reporting frequencies based on these broad racial and ethnic categories was that reproduction occurred within ethnic or racial groups as a result of historical patterns of migration as well as social and legal prohibitions against mating across racial and ethnic lines.[83] As a result, there is a higher probability of finding particular alleles within a specific racial or ethnic population.[84]

Two prominent scientists challenged those groupings publicly, noting that sufficient genetic substructure—i.e., variations in allele frequencies—within

80. *See* JAY D. ARONSON, GENETIC WITNESS: SCIENCE, LAW, AND CONTROVERSY IN THE MAKING OF DNA PROFILING 120–46 (2007) (detailing the DNA "wars" over various aspects of population genetics).

81. MURPHY, *supra* note 4, at 85–89 (explaining basic statistical approach).

82. *See, e.g.*, COMM. ON DNA FORENSIC SCI., NAT'L RESEARCH COUNCIL, THE EVALUATION OF FORENSIC DNA EVIDENCE 116 tbl.4.9 (1996) [hereinafter NRC II] (reporting data for "Black," "White," and "Hispanic").

83. *See id.* at 28 ("The population of the United States is made up of subpopulations descended from different parts of the globe and not fully homogenized."); *see also id.* at 57–58, 98–99, 111–12 (explaining data). *But see id.* at 94 (noting "a point often made by population geneticists—namely, that differences among individuals within a race are much larger than the differences between races").

84. *See id.* at 28 ("Extensive studies from a wide variety of databases show that there are indeed substantial frequency differences among the major racial and linguistic groups (black, Hispanic, American Indian, east Asian, and white).").

these general categories undermined the broader categorization effort.[85] In layman's terms, they argued that a category like "White" could include "Southern European" or "Northern European" or "Middle Eastern" or "South American," and yet the frequencies of a particular allele within each of those groups would vary as a result of migration patterns just as much as would the allele frequency between a "Black" person descended from Africa and a "White" person descended from southern Europe.[86] Thus, to report "White" data without accounting for the variation within a category, while suggesting that there was enough meaningful variation to distinguish "White" from "Black," was inconsistent. Moreover, frequencies of alleles might be more similar across "White" and "Black" categories when the person came from a part of the world with less rigid historical segregation, or even for regions in which reproduction more readily occurred across categories (whether voluntarily or as a result of systematic rape).[87] Along the same lines, a category like "Hispanic," which ultimately is a linguistic grouping with social resonance that is not always tied to ancestral migration patterns, artificially lumped together Caribbean with Mexican, Latin American, South American, and European Spanish-speakers.[88] Yet those populations (as a result of those underlying migration patterns) might have significantly different patterns of allele inheritance.[89]

In response, two other prominent scientists agreed that these claims were fundamentally true, but argued that their impact was "trivial."[90] They noted that mating was closer to random than assumed, because it took only two or three generations to smooth observed differences.[91] Moreover, they argued that the variation that did exist was so minor as to be inconsequential to the ultimate task at hand, which was computing a match probability that—taking such differences into account—would vary so minimally as to be insignificant.[92] Indeed, there were many racial and ethnic populations that had *no* specific statistical data, and yet DNA match statistics were routinely presented in general terms for those

85. *See* R. C. Lewontin & Daniel L. Hartl, *Population Genetics in Forensic DNA Typing*, 254 SCIENCE 1745, 1745–46 (1991).

86. *Id.* at 1747 ("That is, for these genes, there is, on average, one-third more genetic variation among Irish, Spanish, Italians, Slavs, Swedes, and other subpopulations, than there is, on the average, between Europeans, Asians, Africans, Amerinds, and Oceanians.").

87. *Id.* at 1748 ("Even today, the typical adult 'Caucasian' in the United States is the grandchild of immigrants. For 'Hispanics,' the situation is at least one generation delayed (counting Puerto Ricans as immigrants). The key point for DNA typing is that there has been very little time for mixing of genes from diverse populations of origin."); *id.* at 1749 ("American blacks in different localities have various amounts of European and American Indian ancestry acquired since their introduction into North America.").

88. *Id.* at 1749 ("'*Hispanics*.' This heterogeneous assemblage is perhaps the worst case for calculating reliable probabilities. The census designation 'Hispanic' is a biological hodgepodge.").

89. *Id.*

90. Ranajit Chakraborty & Kenneth K. Kidd, *The Utility of DNA Typing in Forensic Work*, 254 SCIENCE 1735, 1735–38 (1991).

91. *Id.* at 1737.

92. *Id.* at 1738.

groups.[93] The authors contended that the use of data pooled across different groups should be considered well within accepted ranges of uncertainty.[94]

A third set of critics, exemplified by Jonathan Kahn's influential article, observed that the care and deliberation evident in the development of DNA testing made an interesting contrast to the sloppy, intuitive way race and ethnicity were introduced into the debate.[95] For instance, Kahn argued that the scientists who minimized the intra-group differences as significant nonetheless supported the use of inter-group differences even though, in many cases, the intra-group difference was manifold while the inter-group difference was much smaller.[96] By insisting on recognizing inter-group difference, while papering over intra-group difference, scientists appeared to endorse racialized ideas about genetics.[97] A judge or jury, after hearing match statistics for the "Black" or "White" population, would assume that meaningful genetic difference separated Black people from White people.[98] They would not hear that, although such differences were broadly observable, even more meaningful genetic differences were evident within the categories of "White" or "Black."[99] Even the samples used to construct the frequency tables were viewed as haphazardly assembled. They were small in size, geographically bounded (e.g., one hundred people from New Mexico), and relied chiefly on self-reported race and ethnicity.[100] In contrast, the techniques underlying genetic testing more generally had been refined through more exact scientific examination.[101]

To be clear, the debates about the significance of historical migratory and mating patterns as regards match statistics did not cut along clean political or ideological lines. Although enthusiastic embrace of dubious racial groupings is readily associated with other forms of explicit and implicit bias found throughout

93. *Id.*

94. *See, e.g.*, People v. Cua, 119 Cal. Rptr. 3d 391, 408–10 (2011). In *Cua*, an Asian American defendant argued that the provided random match probability impermissibly excluded his racial group, since the expert provided figures for the African American, Hispanic, and Caucasian populations. *Id.* at 408. However, the court dismissed this claim, first noting that there was no evidence defendant was, in fact, the ethnicity he claimed. *Id.* The court also cited authorities who underscored "the limited role that the defendant's ethnic or racial status plays in evaluating the evidence of a match," and concluded that the testimony "gave the jury relevant information as to the relative rarity in the general population of the genotype found in the crime scene sample." *Id.* at 408–10.

95. Jonathan Kahn, *Race, Genes, and Justice: A Call to Reform the Presentation of Forensic DNA Evidence in Criminal Trials*, 74 BROOK. L. REV. 325 (2009).

96. *See id.* at 341–42.

97. *See id.* at 350.

98. *See id.* at 356.

99. *See* NRC II, *supra* note 82, at 94 (underscoring that inter-group differences are larger than intra-group differences).

100. *See, e.g.*, State v. Champ, No. A-00-617, 2001 WL 273071, at *12 (Neb. Ct. App. Mar. 20, 2001) (rejecting the argument that a database of 100 to 200 African Americans was not representative enough to supply frequency statistics, finding no error in admission of DNA analysis "done in accordance with generally accepted scientific principles").

101. *See generally* DAVID H. KAYE, THE DOUBLE HELIX AND THE LAW OF EVIDENCE (2010) (chronicling development of forensic DNA testing).

the criminal justice system, it also might be marshaled in support of a defendant's interests. Failing to account for meaningful substructure—particularly in populations with a long tradition of intra-group mating (such as Native Americans)—might unfairly prejudice the defendant. This view received a bump in attention when an Arizona analyst reported a large number of pairwise 9- and 10-loci matches within the relatively small Arizona database.[102] One population geneticist endeavored to predict the likelihood of such matches using the statistical probabilities used in criminal cases and concluded that they were highly unlikely—suggesting the match statistics might be wrong.[103]

These battles were not without consequence. At least one court excluded DNA evidence as a result of discomfort over the unsettled nature of the statistical component.[104] It also set the stage for a series of court decisions that wrangled over how to account for racial and ethnic patterns of mating in formulating forensic statistical approaches. For instance, if there is no racial or ethnic information about the perpetrator of the offense, but the defendant belongs to a particular group, should the court admit match statistics for all groups? Just the group to which the defendant belongs? Or a general statistic that is the most conservative of those available?[105] If the race of the perpetrator is known, then

---

102. *See* KATHRYN TROYER ET AL., A NINE STR LOCUS MATCH BETWEEN TWO APPARENTLY UNRELATED INDIVIDUALS USING AMPFLSTR PROFILER PLUS AND COFILER (2001) (presented at International Symposium on Human Identification); David H. Kaye, *Trawling DNA Databases for Partial Matches; What is the FBI Afraid of?*, 19 CORNELL J. L. & PUB. POL'Y 145, 153–54 (2009).

103. Laurence D. Mueller, *Can Simple Population Genetic Models Reconcile Partial Match Frequencies Observed in Large Forensic Databases?*, 87 J. GENETICS 101, 107 (2008). In 2015, the FBI acknowledged that there were errors in the original datasets used to generate allele frequencies, which were attributed both to clerical mistakes (e.g., transcription errors) as well as technological failures (e.g., treating stutter as a true allele); corrections were required for 255 of 1239 (around 20%) of the recorded allele frequencies. Tamyra Moretti et al., FBI Laboratory, Presentation at the International Symposium on Forensic Science Error Management: Genotyping Errors in the FBI STR Allele Frequency Database Used for Estimating Match Probabilities in Forensic Investigations 14, 28 (2017) (citing Tamyra R. Moretti et al., *Erratum*, 60 J. FORENSIC SCI. 1114–16 (2015)), https://www.nist.gov/system/files/documents/2017/08/23/anthonyonoratotuesdayafternoonsession.pdf [https://perma.cc/6BNN-B65K]; *see also* Spencer S. Hsu, *FBI Notifies Crime Labs of Errors Used in DNA Match Calculations Since 1999*, WASH. POST (May 29, 2015), https://www.washingtonpost.com/local/crime/fbi-notifies-crime-labs-of-errors-used-in-dna-match-calculations-since-1999/2015/05/29/f04234fc-0591-11e5-8bda-c7b4e9a8f7ac_story.html [https://perma.cc/Y5QS-GGZW].

104. People v. Barney, 10 Cal. Rptr. 2d 731, 743 (1992).

105. *See, e.g.*, People v. Wilson, 136 P.3d 864, 872 (Cal. 2006) (surveying options of which statistic to present when the perpetrator's race is unknown). *Wilson* put to rest a series of cases in the appellate courts that raised the issue in the context of a defendant who was "half Hispanic and half Caucasian," especially when the testimony concerning the perpetrator's racial or ethnic identity was ambiguous. *Id.* (citing People v. Pizarro (*Pizarro I*), 12 Cal. Rptr. 2d 436 (Cal. Ct. App. 1992), and People v. Pizarro (*Pizarro II*), 3 Cal. Rptr. 3d 21 (Cal. Ct. App. 2003)). The appellate court ultimately found that a defendant's race was not relevant to prove a random match probability unless the defendant's race was "sufficiently established." *See Pizarro I*, 12 Cal. Rptr. 2d at 443; *Pizarro II*, 3 Cal. Rptr. 3d at 31; *see also* People v. Prince, 36 Cal. Rptr. 3d 300, 317 (Cal. Ct. App. 2005) ("We do not know whether Caucasian, Hispanic, and African–American databases are 'generally representative' of the population as a whole. For such evidence to be admissible, as an initial matter an expert witness would have to be

should only that statistic be introduced because all others are irrelevant? Should it matter how the perpetrator's race was identified and whether that perception was reliable or accurate?[106] Who determines the defendant's race—the defendant, "common sense," or the defendant's DNA profile? It was also controversial when ethnic or racial allele frequencies empowered investigators to use DNA—and at times, tests for additional genomic markers directly associated with biogeographical ancestry—to predict the "race" or "ethnicity" of a perpetrator.[107]

Concrete evidence that forensic databases are heavily skewed demographically might also undermine the accuracy of statistical match probabilities in a case built upon the identification of the perpetrator as a result of a "cold hit" or match in a DNA database. In such a case, the overrepresentation of a person's subpopulation may increase the probability that the match occurred by chance. In other words, if the racial composition of the DNA database is not randomly selected, but instead heavily skews toward inclusion of particular subpopulations, then a match predicated upon a random match probability is less likely to be accurate.

---

able to testify that extrapolation from specific ethnic populations to the population as a whole is scientifically appropriate and that, for example, a DNA profile which is shown to be rare in three major ethnic populations will be equally (or comparatively) rare in the general population."), *superseded*, 132 P.3d 210 (Cal. 2006).

106. *See, e.g.*, State v. Daye, No. CR110234742, 2013 WL 1189441 (Conn. Super. Ct. Feb. 28, 2013). In *Daye*, the court wrote:

> In fact, using the victim's ethnicity to determine which ethnic group should be used for probability comparison would be improper, as it could insinuate that the perpetrator was Indian. On the other hand, if the perpetrator were known to be Indian, comparing the matched profile with non-Indians would be irrelevant; however that is not the case here.
>
> The contention that the database which is used for calculating the probability that the matched sample obtained from the crime scene originated from an individual other than the defendant should have contained Jamaicans specifically, rather than African–Americans in general, is likewise unfounded. A comparison of the matched profile to profiles of African–Americans, Caucasians and Hispanics in general, based on the testimony of Ms. Roy, is clearly relevant to the jury's determination of the identity of the victim's killer.
>
> Finally, the record is bereft of any evidence regarding the existence of a methodology, principle, or protocol by which an expert, consistent with accepted scientific principles, is required to include other population subgroups in calculating the frequency of an allelic profile in question, or whether there exists an accepted scientific population threshold in a subgroup that must be achieved in order for such a subgroup to be included in calculations, and if achieved the mechanism by which the subgroup is included. In addition, there was no evidence that the Connecticut State Forensic Laboratory failed to appropriately apply any such protocols, if they exist.

Id. at *6–7 (citation omitted).

107. MURPHY, *supra* note 4, at 215; Bahrad A. Sokhansanj, Note, *Beyond Protecting Genetic Privacy: Understanding Genetic Discrimination Through Its Disparate Impact on Racial Minorities*, 2 COLUM. J. RACE & L. 279, 296 (2012) ("Where no match was found in the existing database, DNA samples have been used to attempt to predict a suspect's racial or ethnic origin. This explicitly racial use of forensic DNA analysis echoes the race-based medical research questions discussed above and raises important issues regarding racial profiling and the potential for reinforcing stereotypes associated with criminal behavior—or perhaps the use of more insidious categories, such as those associated with genetic traits thought to be explicitly predictive of behavior.").

C. Conclusion

Over the twenty-plus years in which DNA databases have operated and forensic DNA testing has occurred with regularity, there have been questions about the demographic composition of such databases. In the absence of any clear data, scholars have resorted to reciting general platitudes or relying on simplified formulas to estimate database composition. Those estimates reverberated throughout the literature and even into public debates about database practices with real consequence. This enduring need for more concrete and granulated data propelled us to seek a more satisfying and robust answer to the question of whether, and to what extent, DNA databases are racially disparate.

II.
ACTUAL AND ESTIMATED DATABASE COMPOSITION

To learn more about the demographic composition of DNA databases, we undertook two complementary investigations. First, we submitted requests to every jurisdiction seeking information pursuant to the local FOIA or sunshine law.[108] Second, we endeavored to reverse-engineer the database on our own. The results of our disclosure requests are discussed in Part II.A, and the results of our estimates are discussed in Part II.B. Part II.C compares the disclosed data to our estimated data.

Before turning to those results, however, there are a few caveats. First, our effort to ascertain the racial and ethnic composition of DNA databases admittedly relies on a socio-cultural practice (e.g., ascertaining "race") that genetics in many ways itself belies. Recognizing this limitation, which we discuss further in Part III.B, we nonetheless think it valuable to explore database composition in light of the enduring socio-cultural significance of these categories.

Second, there is wide variation in how categories like race and ethnicity are experienced, perceived, and recorded. Such variation presents several problems. Most importantly, the lack of care in how these categories are used or assigned may undermine the ultimate integrity of the data and the labels used may not always map perfectly either within or across groups. These areas of contention can lead to confusing or conflicting data that obscure the questions at hand. Accordingly, we have elected to streamline and simplify at the expense of important nuance.

Relatedly, we received data that used a variety of terms to describe demographic groups. But, in nearly all cases, we do not have information on how judgments about those categories were made. For instance, some jurisdictions report on the "African American" percentage, but we doubt that this category includes only Black persons of African descent. We expect it also includes native-born Africans and self-identified Black persons of Caribbean, European,

108. For convenience, we refer to these inquiries as "FOIA" inquiries regardless of the actual title of the enabling law.

or other descent. Similarly, the "White" category covers an enormous geographic range, possibly including persons from or descended from persons in Europe as well as North, South, or Latin America; Northern Africa; or the Middle East. And the "Asian" category may be used for persons from or descended from places as wide ranging as India, Japan, or even the Middle East. Or it may be confined to only countries labeled as the "Southeast" or "Far East."

The Hispanic category is particularly vexing. In some places, "Hispanic" may be treated as a racial identity and thus presented as an option to choose in place of identifiers like "White," "Black," or "Asian." In other places, Hispanic is treated as an ethnic identity rather than a racial category, meaning that it is selected in addition to racial identity. These problems also plague our estimation project which likewise relied on reported data from censuses and official reports not all of which use the same terminology and have varying means of assigning categories.

Given the complexity of the data, and the impossibility of reconciling those differences, we chose to report all received data according to six categories: White, Black, Asian, Hispanic, Native American, and Other. These categories mapped most closely onto the data we received, even if occasionally different labels (such as "African American") were used. Despite its problematic connotations, we chose to use "Other" as a category to mirror the practice common in many states. Finally, because we often wish to make comparisons between the "White" category and all other groups, we use the term "BIPOC" to refer to persons identified as a race or ethnicity other than only White. Colloquially, that abbreviation stands for "Black, Indigenous, and People of Color,"[109] and in this Article it is used to encompass the data provided about persons identified as Black, Asian, Hispanic, Native American, and "Other."

A. Freedom of Information Requests

1. Methodology

Our first source of data was state-level public disclosure of DNA databases. Between January and August of 2018, we sent public disclosure requests to all fifty states seeking data on the composition of their state-level DNA databases based on race, ethnicity, and sex.[110] Because DNA databases are dynamic—new profiles are continuously added—we asked for a snapshot as of a date and time within a year of our request. Twenty-eight jurisdictions replied.

---

109. *See* Sandra E. Garcia, *Where Did BIPOC Come From?*, N.Y. TIMES (June 17, 2020), https://www.nytimes.com/article/what-is-bipoc.html [https://perma.cc/P2SG-ZMJ5] (tracing history of the term BIPOC to 2013, and noting its rising popularity but also critiques).

110. We sought information about the State DNA Index System (SDIS) because we thought it most likely that jurisdictions would hold demographic data at the SDIS level even if they did not retain that data at the NDIS level. Our results suggest that although there may be variation in the composition of SDIS as compared with what is uploaded to NDIS, it is nominal. Accordingly, we think it fair to assume that the disclosed SDIS results provide an accurate depiction of the NDIS data.

Of the jurisdictions that responded but did not supply data,[111] states' most common explanation was that such data are not maintained by the state or that it was not regularly compiled and thus not required to be created under existing standards of disclosure. Three states[112] claimed that the data was altogether exempt from their disclosure laws.

Two states released partial data. New Jersey disclosed sex data but does not maintain racial or ethnic data. Maryland has public data concerning its arrestee population, issued annually by legislative command, but does not have data on convicted persons. Seven states disclosed some, or all, of the requested demographic information: California, Florida, Indiana, Maine, Nevada, South Dakota, and Texas. These results are presented in Part II.A.2 below.

Given that disclosures came from some of the largest state contributors to the national database, the cumulative reported data provides a direct glimpse into the national data. Taken together, the disclosed responses cover just over 5.6 million known persons' profiles, roughly 33% of the national database.

2. Results

Table 1 presents the disclosed data in a single comprehensive chart. The first row shows the total number of known profiles in NDIS from each contributing state as of September 2018.[113] The second row shows the number of profiles in SDIS from each state, as disclosed in the summer of 2018 by each listed state in response to our request.[114] Within that row, in italics beneath the raw number of profiles, is the percentage of the NDIS database that those profiles represent.

The next rows show the percentage of the state's database profiles that comes from each racial or ethnic group. Below each of those, in italics, is that group's share of the general population in the state. All general population demographic data are taken from www.census.gov, which draws from an array

111. Colorado, Delaware, Hawai'i, Kansas, Kentucky, Massachusetts, Missouri, New Hampshire, New Mexico, New York, North Dakota, Oregon, Utah, and Wisconsin.

112. Montana, Washington, and Wyoming.

113. This information is useful for two reasons. Because we requested state-level data, it was possible that some states might have kept data at that level (SDIS) and not transmitted it to the national database (NDIS). By comparing the disclosed data figures with the national figures from each state, we are able to see that the data disclosed does, in fact, represent a close approximation of what is held at the national level. In addition, being able to compare the total number of profiles held in NDIS from those states with the total number disclosed allows us to assert with confidence that we have garnered an accurate snapshot of one-third of the national database.

114. *See* Letter from Shannon Patterson, Staff Services Manager II, Office of the Chief, Div. of Law Enf't, Cal. Dep't of Justice (July 10, 2018) [hereinafter California Letter]; E-mail from Office of the Gen. Counsel, Fla. Dep't of Law Enf't (Dec. 19, 2017); Letter from Kristine Crouch, CODIS Adm'r, Ind. State Police (July 24, 2018) [hereinafter Indiana Letter]; E-mail from Lt. Scott A. Gosselin, Me. State Police Crime Lab. (Aug. 13, 2018); E-mail from Mindy McKay, Dir.'s Office, Nev. Dep't of Pub. Safety (July 12, 2018); Letter from Patricia Archer, Assistant Attorney Gen., Office of the Attorney Gen., State of S.D. (July 24, 2018); Letter from Jennifer Howard, Lab. Records Program Specialist, Tex. Dep't of Pub. Safety Crime Lab. (June 20, 2018). All correspondence on file with authors.

of sources.[115] States also disclosed demographic information about sex, which is provided in the lowest rows.

The far-right column represents total data. The top numbers represent the totals as a percentage of the disclosed data, whereas the bottom numbers represent the totals as a reflection of the national population. Thus, for instance, the box where the "Total" column meets the "Black" row shows that 23.6% of the profiles from the disclosed data came from persons identified as Black, whereas only 13.4% of the national U.S. population identifies as Black.

115. *See QuickFacts: United States*, U.S. CENSUS BUREAU (July 1, 2019), www.census.gov/quickfacts/ [https://perma.cc/NQS7-VX2N] (in the search bar at the top left, type in the name of each state). The figures for the "White" population are drawn from the census figure for "White alone, not Hispanic or Latino." There are no figures for other racial groups that exclude "Hispanic or Latino," so the population percentages reported for those groups may include persons who identify both with the race listed and as Hispanic.

Table 1: Disclosed data by state

	CA[116]	FL[117]	IN[118]	ME[119]	NV	SD	TX[120]	TOTAL
Total in NDIS by state as of Sept 2018[121]	2,768,269	1,350,667	307,714	32,847	167,726	67,600	960,985	5,655,808
Total in SDIS as disclosed by state	2,771,721 *16.5%*	1,175,391 *7.0%*	300,741 *1.8%*	33,711 *.2%*	344,097 *2.1%*	67,753.*4* *%*	918,953 *5.5%*	5,612,367 *33.4%*
White (not Hispanic)	**29.6%** *37%*	**61.4%** *53.5%*	**70%** *78.9%*	**92.8%** *93.1%*	**69.4%** *48.7%*	**66.8%** *81.4%*	**37.3%** *41.5%*	**42.9%** *60.4%*
Black	**17.1%** *6.5%*	**35.2%** *16.9%*	**26%** *9.8%*	**3.9%** *1.6%*	**25.6%** *10.1%*	**6.0%** *2.4%*	**29.1%** *12.8%*	**23.6%** *13.4%*
Hispanic	**34.7%** *39.3%*	**2.4%** *26.1%*	**4%** *7.1%*	**.5%** *1.7%*	- 29%	**4.4%** *4.1%*	**32.7%** *39.6%*	**23.2%** *18.3%*
Asian	**.99%** *15.3%*	**.23%** *3.0%*	- *2.5%*	**.4%** *1.2%*	**2.3%** *8.7%*	**.08%** *1.7%*	**.42%** *5.2%*	**.75%** *5.9%*
Native American	- *2.1%*	**.06%** *.6%*	- *.5%*	**1%** *.7%*	**1.7%** *2.5%*	**21.5%** 9.1%	**.02%** *1.1%*	**.38%** *1.5%*
Other/multi racial if separate from unknown	- *3.9%*	**.1%** *2.2%*	 *2.1%*	**-** *1.8%*	**-** *4.5%*	**1.3%** *2.4%*	**.28%** *2%*	**.08%** *2.7%*
Unknown		**.61%**		**1.4%**	**1.0%**			**.2%**
Male	76.8%	76.7%	80%	82.7%	81.6%	75.6%	84.7%	
Female	18.7%	22.2%	20%	17%	18.4%	24.4%	15.3%	
Other	-	1.1%	-	-	.02%			

116. California provided separate data for arrestees and convicted persons; the data presented combine those percentages.

117. Florida did not disclose how it determines race or ethnicity.

118. Indiana did not provide numerical breakdowns other than the total arrested and convicted persons by category; instead Indiana reported the information as percentages.

119. Nevada did not report data for the Hispanic category.

120. Texas reported only female convicted persons figures.

121. *Sept. 2018 NDIS Statistics*, *supra* note 3. As of September 2018, the National Database contained 13,528,363 offender profiles (which are primarily from convicted persons, but also include certain detained persons and legally mandated samples) and 3,280,752 arrestee profiles, for a total of 16,809,115 profiles of known persons. *Id.*

3. Reflections

The disclosed data offer rare insight into the actual demographic impact of DNA compulsory collection policies. The disclosed data represent roughly a third of the national database. Of course, it is possible that data from other states might paint a different picture than the one depicted by the disclosed data. Thus, the conclusions drawn from these data are necessarily limited by the possibility that this snapshot is not, in fact, representative. Even considering the data on their face, several conclusions leap to the fore.

First, the disclosed data confirm that DNA databases are racially disproportionate and, in some instances, starkly so. Figure 1 below shows the demographic composition of the U.S. population, whereas Figure 2 shows the demographic composition of the seven states who disclosed, which constitute a third of the national database. Most prominently, although White people constitute almost two-thirds (60%) of the combined population of the disclosing states, they make up less than half (43%) of the DNA database. Conversely, people of color make up almost a half of the DNA database (48%) disclosed by the states, even though they constitute only a little over a third (39%) of the population.

Figure 1 Racial and ethnic composition of U.S. population

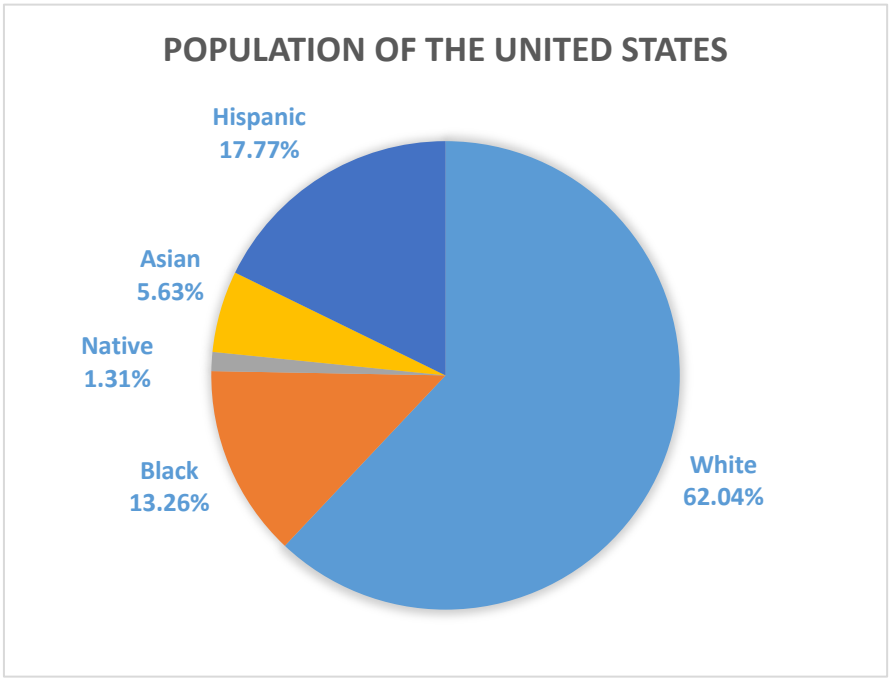
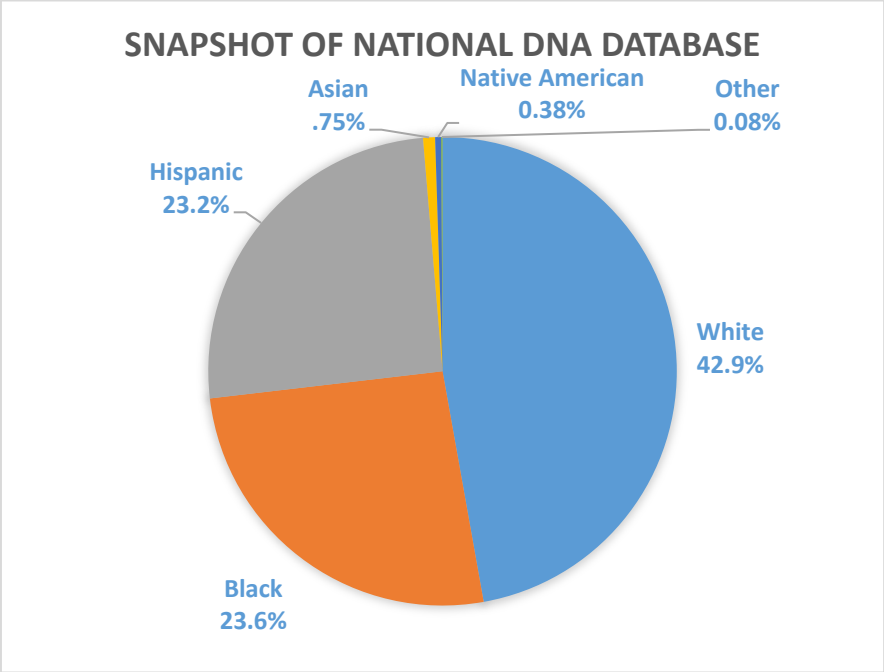


Figure 2 Racial and ethnic composition of 33.4% of national DNA database



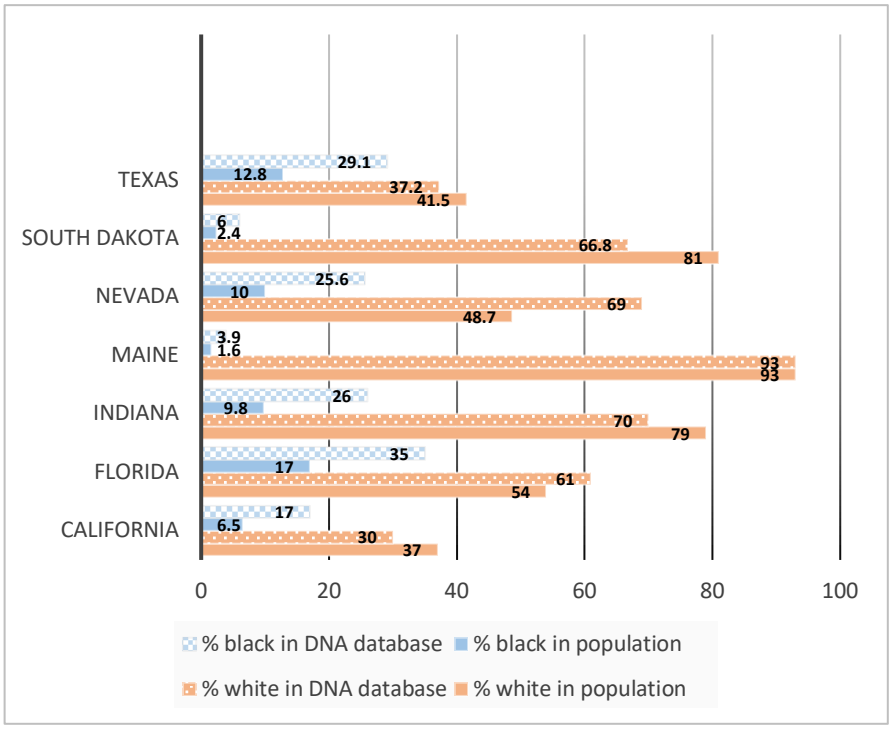
These data also provide insight into a range of demographically different states.[122] California, Texas, and Florida are all large states with aggressive collection policies and diverse general populations. Maine, Indiana, and South Dakota are smaller states with much more homogenous populations. And Nevada rests somewhere in between, as a mid-size state with a somewhat diverse population. All states except Maine allow for arrestee DNA collection, although not every state broke out its data in that way. Whether taken as emblematic of the larger database, or simply considered individually, these data reveal several important things.

Most importantly, the White, non-Hispanic population was never overrepresented in the databases of states that disclosed data,[123] whereas the Black population was overrepresented in every state. This disparity was dramatic.

122. *See generally* Rich Williams, *Forensic Science Database: Search by State*, NAT'L CONF. ST. LEGISLATURES (Nov. 17, 2014), http://www.ncsl.org/research/civil-and-criminal-justice/dna-database-search-by-state.aspx [https://perma.cc/LEE7-MC9M] [hereinafter NCSIL Database] (cataloging DNA policies by state). The website for the NCSL database formerly displayed an interactive map that is no longer accessible. Each state's page is still active, however. For California's page, for example, see NAT'L CONF. ST. LEGISLATURES, CALIFORNIA, https://www.ncsl.org/portals/1/documents/cj/dna/california.pdf [https://perma.cc/7KQK-RT3N] [hereinafter NCSL, CALIFORNIA]. Other states' pages may be accessed by substituting "[state]" with the name of the state here: https://www.ncsl.org/portals/1/documents/cj/dna/[state].pdf.

123. Florida and Nevada were the only two states in which the percentage of DNA collected from White persons exceed the percentage of that group in the general population. But there is reason to exclude those states in determining whether the White, non-Hispanic population is overrepresented in

Figure 3. Comparison of percentage of state's DNA database by category versus general population



In every state, the share of DNA profiles from the Black population was double or triple that of the state's general Black population. This remained true whether the state had a large Black population (e.g., Florida, where 16.9% of the population is Black) or small (e.g., Maine, where only 1.6% of the population is Black).[124] In contrast, the share of DNA taken from White persons more closely approximated their share of the population at large, although, in every state, the percentage of White profiles was smaller than that group's share of the general population—in some places by a difference of over 15%. Also notable is that Asians, across all states, show much smaller proportional rates of contribution than suggested by their population size. In California, for instance, Asians constitute nearly 15% of the general population but constitute less than 1% of the DNA database.

any of the disclosed data. Namely, Nevada did not report any figures for Hispanic persons despite a state population that is 29% Hispanic. Similarly, Florida reported on 2.4% of persons as Hispanic, despite a population that is 26.1% Hispanic. Given that many Hispanic persons also identify as White and given data on the prevalence of arrest and conviction of Hispanic persons in those states, it is certain that these group figures include significant numbers of White Hispanic persons. *See infra* Parts II.C.2, III.B.

124. *See QuickFacts: United States*, *supra* note 115.

The data also painted a confusing picture of collection from persons of Hispanic ethnicity. The states showed erratic patterns for recording that data, unrelated to the size of the state or the size of its Hispanic population at large. For instance, California and Texas both have large Hispanic populations and tracked Hispanic ethnicity. Nevada, on the other hand, also has a large Hispanic population but did not appear to track Hispanic ethnicity. Indiana and South Dakota, with relatively small Hispanic populations, also tracked Hispanic ethnicity, while Maine did not. The decision to track or not to track that Hispanic ethnicity did not seem to turn on any readily ascertainable factor.

Similarly, there are reasons to question the reported data on Hispanic—and, by association, White—populations even more so than with regard to other categories. Because Hispanic is typically considered an ethnic category that is additive to a racial category, it is unclear how states chose to prioritize particular aspects of identity.[125] This is especially true since the data show that persons were assigned to only one category. For instance, Florida has a large Hispanic population (26.1% of the population) but reported Hispanics as very low database contributors (2.4%). These figures raise questions about accuracy, especially given general criminal justice data from the state that suggest higher rates of arrest and conviction among Hispanic people.[126] That the ethnic data are tracked, but seemingly inconsistently and poorly, is a telling comment on their apparent salience to law enforcement as well as the reliability of the remaining data.

Lastly, it is notable that in one state—South Dakota—Native Americans constitute an outsized proportion of the DNA database. In South Dakota, Native Americans make up only 9% of the population, but constitute 22% of the DNA database. To be fair, that state has by far the highest share of Native Americans in the general population among the seven—9% versus the next largest neighbor, Nevada, at 1.7%. But the contribution rate is more than double: Maine is second in terms of the contribution rate. There, Native Americans constitute 0.7% of the population but contribute 1% of the database. In the other states, Native American collection rates are much more proportionate to the general population. That discrepancy raises questions about the concentration of policing or other policy choices with regard to that particular population or the dynamics of that particular state.

B. Estimates

Though we were unable to get official DNA database data from the majority of states, we nonetheless wished to estimate the disparities more broadly. We thus attempted to deduce the racial and ethnic composition of DNA databases based on public information. In short, we endeavored to reverse-engineer the

125. *See* Lewontin & Hartl, *supra* note 85.

126. *See, e.g.*, *Florida Profile*, PRISON POLICY INITIATIVE, https://www.prisonpolicy.org/profiles/FL.html [https://perma.cc/L36M-J27Q] (showing 14% of the prison population as "Latino").

composition of the national DNA database. Part II.B.1 generally explains the methodology used to estimate the level of disparity. Part II.B.2 presents our results. Part II.C compares our estimates with the information provided in response to our FOIA requests from the seven states that responded with information. This comparison allowed us to check the accuracy of our estimation methodology.

1. Methodology

For each state, our goals were to 1) identify the state's compulsory DNA collection laws; 2) find demographic data on the population governed by those laws, either directly or through a proxy; 3) aggregate each state's demographic data; and 4) normalize the data across states by finding the proportion of each state's population to which the state's collection policy applies. But precisely reverse-engineering DNA databases would require access to data at a level of granularity and specificity that is simply publicly unavailable.[127] We thus had to use proxies and rough estimates, even while acknowledging that this approach suffers from several shortcomings of varying degrees of scope and severity.

Most fundamentally, any estimate that we produce is by necessity a snapshot of the database at a moment in time, because at best it reflects the collection policies and demographic trends of that moment. Because DNA collection laws have changed over time and demographic populations have shifted, our method fails to capture variations that may have occurred. Such variation would affect the overall composition of the DNA database, even though it would accurately depict the composition under present policies.

More particular problems arise with respect to states or categories for which there was no published demographic data directly pertinent to our estimates. For instance, a state may report general demographic data about misdemeanants, but not break that category down by offense type. Yet compulsory DNA collection laws in the state might apply only to particular crimes. We explain in greater detail below the specific challenges we faced in each piece of our estimation process, and an appendix with detailed data is on file that explains the specific sources we relied upon for each state.

Notwithstanding these challenges, we began by identifying the compulsory collection policy for the state—that is, the type of criminal justice contacts that require an individual to submit a DNA sample. Then we aggregated the total number of such events in the most recent year such aggregation was disclosed and determined the racial composition of this aggregation. For example, suppose

127. Such data is inaccessible in part because it is not collected, and in part because even if collected, it is not shared or reported in a publicly accessible forum. For instance, a state might track convictions for misdemeanors, but fail to track them in detail that maps onto DNA collection statutes—such as one that requires DNA only from certain categories of misdemeanants (e.g., sexual offenders with prior convictions). Or a state may collect such information, but fail to publicly report demographic characteristics of persons within a specific group.

we determined that a state requires all persons convicted of a felony and all sex offense misdemeanants to submit a DNA sample. Based on those requirements, we tallied the number of felony and relevant misdemeanor convictions, identified the demographic characteristics of each of those groups in the state, and then mapped each piece together to determine the comprehensive demographic picture of probable DNA contributors for that jurisdiction. Ultimately, we pieced each state together in proportion to its overall participation in the national database system to generate a national-level estimate.

a. Incidents That Trigger DNA Submissions

Our first step required us to determine which incidents trigger compulsory DNA submission in each state. The National Conference of State Legislatures maintains a Forensic Science Database that catalogues state law.[128] For each state, the database aggregates the laws that would require the submission of a DNA sample.[129]

In general, almost all states require the submission of DNA samples for sex crime misdemeanor and felony convictions. Several states also mandate DNA collection for other misdemeanor convictions. Lastly, some states require DNA samples at the arrest stage for all or specified felonies, misdemeanors, or a combination of the two.

The main shortcoming in using statutory collection requirements as a proxy for DNA submission is that compulsory collection laws have changed over time. A snapshot of the composition of the database today may not in fact reflect the actual composition of the database if the law recently changed to expand to new or different classes of persons.

The static picture that we develop also likely overcounts on occasion because we cannot account for repeat offenders. For instance, if a person commits two qualifying offenses in one year, and the demographic data we use reports the racial composition of the arrestee population by arrest rather than by person, then that individual will be double-counted in our estimate.

b. Conviction and Arrest Data

The next step is to find demographic data concerning all individuals who qualify for collection according to existing law and to determine the total number of samples collected and the racial demographics of those from whom samples were taken. Unfortunately, states typically do not provide total compiled data of convictions and arrests for a specified period of time—much less for the specific time period in which DNA collection laws have been in place. Instead, states provide annual data. Thus, we used this annual data as the basis for estimating the total data, recognizing that actual offense rates differ from year to year.

128. NCSIL Database, *supra* note 122.
129. *See, e.g.*, NCSL, CALIFORNIA, *supra* note 122.

With these limitations in mind, we searched for information about persons with convictions and arrests that, based on the states' laws, would require the individual to submit DNA samples. Many states provide some form of yearly felony conviction data. Some states also provide information on yearly arrests for certain crimes. For states that do not provide this information, we instead turned to their prison admission data as a functional, if imperfect, proxy for conviction rates.

We note that these data are imperfect because they leave out many arrests or convictions that would nonetheless result in DNA collection. For instance, in states in which prison admission data is the only information disclosed, the information does not include convictions that do not lead to imprisonment or convictions that lead to incarceration in a facility other than a prison. The data also obviously exclude arrests that do not lead to conviction. Similarly, for those states that provide only conviction data, data of DNA submission upon arrest are missing. Finally, although a state may have a collection policy, it may not implement that policy perfectly in practice: the state may fail to actually collect from eligible persons notwithstanding statutory authorization.[130] These problems likely lead to an underestimation of the number of DNA submissions. This is particularly problematic regarding misdemeanor offenses—both arrests and convictions. Those offenses often had little or no recorded data, much less data broken down into categories (such as sex offenses) that key to DNA submission.

On the other end of the spectrum, there could be discrepancies where both convictions and arrests are disclosed. Some states require DNA submission only for certain types of arrests. But the conviction and arrest data do not usually differentiate based on the type of crime leading to the conviction or arrest. Thus, it is possible that there could be individuals that the state counts twice, as the state would record the individual as both an arrested and as a convicted data point. We should also note that we find such duplication in the actual DNA databases.[131] Since it is possible that some of those individuals arrested are not convicted, and some of those convicted may have been arrested for a crime that does not require DNA submission, it is not safe to rely on just the arrest data or just the conviction data. The double counting from this issue would contribute to some overestimation.

---

130. *See, e.g.*, Rachel Dissell, *DNA From Thousands of Cuyahoga County Felony Arrests Never Taken, Not in CODIS Crime-solving Database*, CLEVELAND PLAIN DEALER (June 16, 2017), https://www.cleveland.com/metro/2017/06/dna_from_thousands_of_cuyahoga_county_felony_arrests_never_taken_not_in_codis_crime-solving_database.html [https://perma.cc/DS7N-8C6P] (estimating up to 10,000 samples eligible for inclusion in database were never collected); Rachel Dissell, *Cleveland Police Not Following State DNA Collection Laws; Other Cities Get Court Orders When Arrestees Refuse Swabs*, CLEVELAND PLAIN DEALER (Sept. 29, 2014), https://www.cleveland.com/court-justice/2014/09/cleveland_police_not_following.html [https://perma.cc/9H53-M3ET].

131. *See, e.g.*, Julie Samuels et al., *Collecting DNA From Arrestees: Implementation Lessons*, 270 NIJ J. 18, 22–23 (2012).

c. Contributor Demographics

The most difficult data to collect and aggregate are the data related to the racial or ethnic composition of convicted or arrested persons. Many states do provide some racial demographic data concerning annual convictions or newly admitted prisoners. However, multiple problems arise. First, the data often are limited to select groupings. For instance, states typically show the percentages of "White", "Black," and "Other." The use of "Other" leads to insufficient data on the number of non-Black minorities whose DNA was collected (e.g., Hispanics, Asians, Native Americans). Furthermore, many states do not clearly explain how they identify Hispanic persons. Instead, those persons may be subsumed under "White" or "Black," or conversely may be counted only in one column ("Hispanic") without regard to their race. Depending on how the state counted, this could lead to over- or under-estimations in different categories.

Lastly, some states do not provide any racial demographic data about arrests or convictions. For those states, we used the composition of the state's prison population. The Prison Policy Initiative provides data for all states on the percentage of Whites, Blacks, and any other races prevalent in the state in the state's prison population.[132] This information is in turn gathered from the 2010 census. For obvious reasons, these data are imperfect proxies, not least because we presume that the racial composition of recent convictions and imprisonments is consistent with that of the prison population in 2010. Moreover, our annualized approach also effectively presumes stability in the racial composition of convictions and arrests year to year (whether in terms of the fraction of the total that any particular offense might represent, or in terms of the demographic characteristics of persons arrested for that offense). It thus presumes that the racial percentages of the most current year's data are representative of the racial percentages of the total conviction and arrest data. Nevertheless, it serves at least some anchoring function in estimating the total population.

d. Racial Disparity Analysis

With our estimates, we can approximate the percentage of persons within a particular racial group who contributed DNA to the DNA database each year. To make this comparison, we use the 2010 U.S. census data, which includes the total population of each state.[133] We then find the percentage of individuals whose DNA was collected for each race in each state—e.g.,

$$\frac{\textit{Annual number of Black persons who submitted their DNA}}{\textit{Total number of Black persons in the state}} \times 100\%$$

132. *Discover Your State*, PRISON POLICY INITIATIVE, https://www.prisonpolicy.org/profiles/ [https://perma.cc/L8NN-GB8M].

133. *QuickFacts: United States*, *supra* note 115.

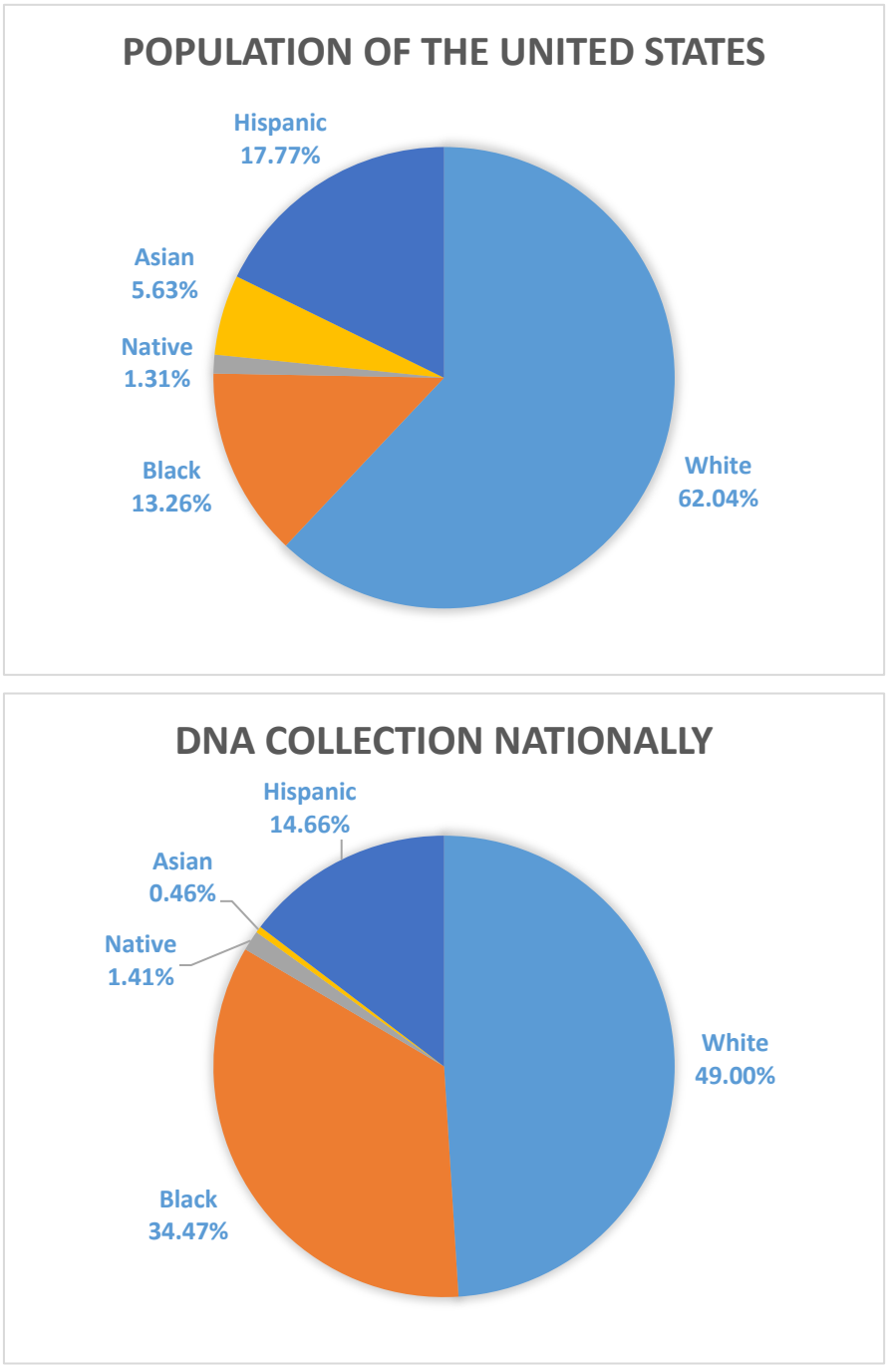
This allows us to compare the percentage for each race within each state, as well as the percentage of each race in the United States generally. An appendix with detailed data is on file that reproduces our state-level data.

2. Results

Based on the available data collected from the state databases and from the demographic data disclosed in the 2010 U.S. census, we found a clear disparity between the percentage of BIPOC persons whose DNA has been collected and the percentage of White persons whose DNA has been collected. The data show that the DNA of BIPOC persons is collected 1.37 times more often than the DNA of White persons. Importantly, this is likely an underestimate, as many states did not provide data on Asian or Hispanic persons. Within the BIPOC group, it is clear that Black persons bear the brunt of the disparity: we estimate that 2.26% of the Black population have their DNA collected in a year, whereas only 1.21% of all BIPOC persons have their DNA collected within that same year.

As a result, according to our calculations, the national DNA database contains DNA profiles from a disproportionate number of Black persons. Figure 4 below compares U.S. population demographics with the DNA database demographics. It shows that although White people make up 62% of the U.S. population, they make up only 49% of the DNA database. In contrast, although Black people make up only 13.26% of the U.S. population, they make up 34.47% of the DNA database.

Figure 4. Comparing the estimated racial breakdown of the DNA collected nationally with the racial composition of the U.S. population.



The racial breakdown of the DNA database in Figure 4 assumes that the racial breakdown of the DNA collected annually for each state is the same as that of the total DNA profiles collected by each state, as disclosed by the National DNA Index.[134] We aggregated the data for the states to create the racial breakdown for the total DNA profiles nationally.

The following subsections contain graphs that highlight aspects of this disparity. The first subsection presents nationwide data, while the second subsection compares data among states.

a. Nationwide

Figure 5 shows the percentage of DNA collected from the population annually for each racial group. The graph aggregates the total number of persons of each racial group whose DNA had been collected in our annual data and compares that to the total number of persons of each race based on the 2010 U.S. census.

Figure 5. For each race, the percentage of the population whose DNA had been collected annually. The percentage for BIPOC persons is the aggregate of the data for each category other than White.

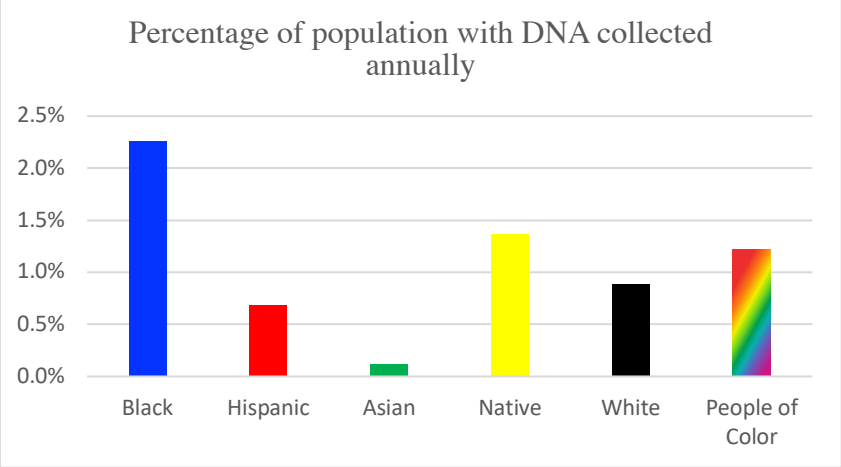
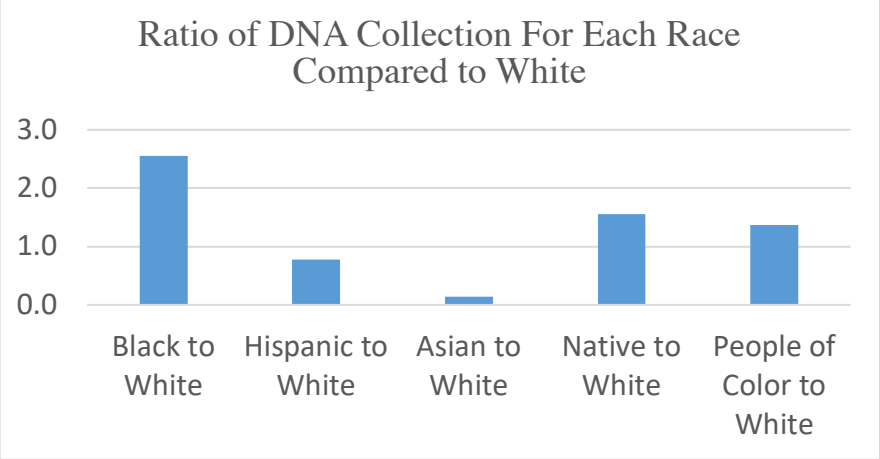


Figure 5 shows that a greater percentage of BIPOC persons have their DNA collected annually than White persons. Within the BIPOC category, Black persons experience the greatest disparity as more than twice as many Black persons as White persons are required to contribute DNA. Both the percentages of Hispanic persons and of Asian persons are less than that of White persons. These percentages, however, may be underestimates as data for those groups were not disclosed in many states.

134. *Sept. 2018 NDIS Statistics*, *supra* note 3.

Figure 6 provides another look at this disparity by presenting the ratio of the percentage for each race to the percentage of White persons. This graph is generated by dividing the data for each race shown in Figure 5 by the data for White persons. Thus, if the ratio given is above one, the percentage of that population whose DNA is collected is greater than for White persons.

Figure 6. The ratio of the percentage of each BIPOC population whose DNA is collected to the percentage of the White population whose DNA is collected.



As Figure 6 shows, DNA has been collected from Black persons at two and a half times the rate of collection from White persons. Figure 6 also shows that Native Americans have had their DNA collected at one and a half times the rate of collection from White persons.

b. State-by-State Comparison

The following graphs compare the data for all states. The data show that in all but two states, the percentage of BIPOC persons required to contribute DNA annually is greater than that of White persons. Within the BIPOC population, the collection of DNA for Black persons is much greater than the collection for other groups. Lastly, in every state, a greater percentage of Black persons gets their DNA collected annually than that of White persons.

Figure 7 shows, for each state, the ratio of the percentage of BIPOC persons whose DNA is collected annually to that of White persons. The larger the ratio, the greater the discrepancy in DNA collection between the two groups. The data show that the ratio is less than one only in Florida and Utah. However, for both states, there was a lack of data concerning the arrests and convictions of Hispanic persons, perhaps due to "Hispanic" being treated as an ethnic, and not racial, group. It is likely that including the missing data would raise the ratio above one. In fact, Florida's disclosed results paint a complex picture, as discussed in Parts II.A and II.C.2. On the other end of the spectrum, the data shows the starkest ratios in Delaware, West Virginia, and New York. In these states, the percentage

of the BIPOC population whose DNA was collected greatly exceeds that of the White population.

For each state where the data was publicly available, Figure 8 shows the percentages of Black, Hispanic, and Asian American populations whose DNA was collected annually. As noted in Part II.B, many states were missing data for groups other than White persons or Black persons. The data clearly show that the percentage of the Black population for each state whose DNA was collected greatly exceeds that of the other BIPOC groups. In many states—notably South Dakota, North Dakota, and Utah —the disparity is sizeable. The graphs show that the negative impact of DNA collection clearly affects Black persons more than persons of any other race or ethnicity.

Figure 9 narrows the dataset of Figure 7 and shows the ratio of DNA collection specifically between Black and White persons. These ratios show that, in every state, the percentage of Black persons whose DNA is collected annually is greater than that for White persons. Furthermore, in the states where the disparity is greatest—Delaware, New York, and West Virginia—the disparity is even larger than the ratio between BIPOC persons and White persons.

Figure 7. The ratio of the percentage of BIPOC persons whose DNA is collected to the percentage of White persons whose DNA is collected for each state.

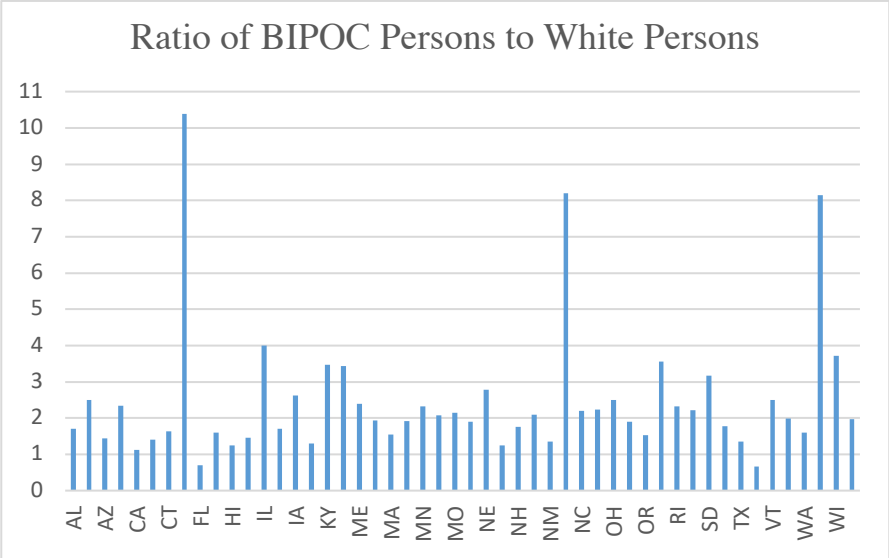


Figure 8. For each group, in each state, the percentages of the Black, Hispanic, and Asian populations whose DNA was collected annually.

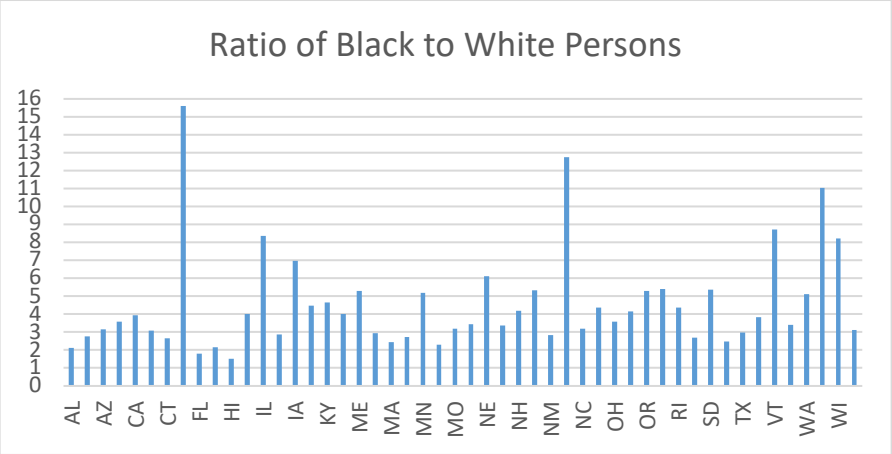
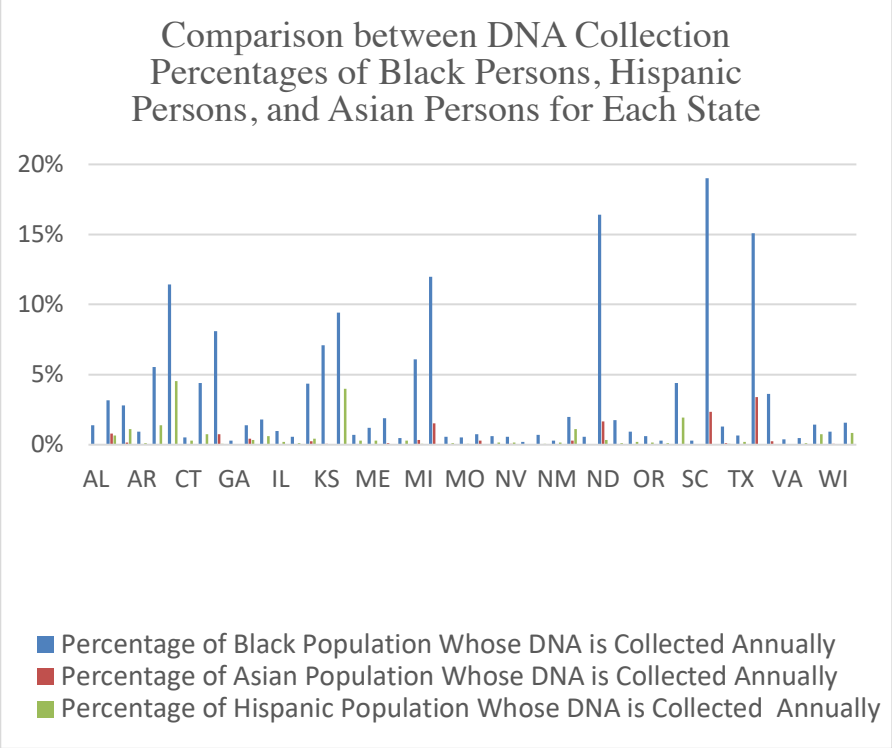


Figure 9. The ratio of the percentage of the Black population whose DNA is collected to the percentage of the White population whose DNA is collected for each state.



C. Comparison Between Disclosed Data and Estimated Data

As described in Part II.A, we received actual data about the demographic composition of the DNA databases of California, Florida, Indiana, Maine, Nevada, South Dakota, and Texas. These data provide some baseline against which to gauge the accuracy of our estimation methodology. However, this test is not perfect.

First, the publicly disclosed data consists of overall data whereas our estimates are based on annual data. Thus, comparing the percentage of each group whose DNA is collected will lead to a comparison between annual data and aggregate data. The better comparison between the disclosed data and our estimated data uses ratios of the percentage of DNA collection among the racial and ethnic groups for both sets of data. To determine each state's ratio, we used each state's disclosed figures to identify the number of persons in each group who submitted DNA to the state database (e.g., 20,000 Black persons; 50,000 White persons). Using that figure and the 2010 U.S. census data, we then figured out what percentage of that group had their DNA collected (e.g., 20,000 of 100,000 Black people in the state is 20% of the state's Black population; 50,000 of 500,000 White people is 10% of the White population). We then used these numbers to compute the ratio of contribution between racial and ethnic groups (e.g., a 2:1 Black-to-White ratio).

This subpart examines each of the states that disclosed data and compares that data to our estimates for the same states using this ratio approach.

1. California

Table 2. Comparing the percentages of different racial groups whose DNA was collected from the estimated California data and the actual California data.

	CA (estimated)	**CA (disclosed)**
BIPOC to White	1.29	1.10
Black to White	3.94	3.38
Hispanic to White	1.24	1.15
Asian to White	0.06	0.09
Black to All BIPOC	3.05	3.07
Hispanic to All BIPOC	0.96	1.04
Asian to All BIPOC	0.04	0.08

California's disclosed data show that there are 1,629,012 convictions and 655,695 arrests in the system. Table 2 compares this data with our estimated data.

The table shows that for many ratios, the estimated and the disclosed data are similar. However, there are larger discrepancies in the estimated data for the

Black to White ratios and the BIPOC to White ratio. The comparison between the sets of data show that our estimate presents a rough, although imperfect, picture of the discrepancies.

2. Florida

Table 3. Comparing the percentages of different racial groups whose DNA was collected from the estimated Florida data and the actual Florida data.

	FL (estimated)	**FL (disclosed)**
BIPOC to White	0.70	0.78
Black to White	1.79	1.89
Hispanic to White	0.02	0.09
Asian to White	0.17	0.07
Black to All BIPOC	2.57	2.46
Hispanic to All BIPOC	0.03	0.11
Asian to All BIPOC	0.24	0.09

The disclosure from Florida gives the racial composition of the group of 1,175,391 people whose DNA was collected. Table 3 compares this data to our estimated data.

The table shows more discrepancies than the comparison of the California data. The ratio between BIPOC and Hispanic persons differs by almost four times, with our figure grossly underestimating the ratio. The other ratios in Table 3 are also off by matters of degrees.

3. Indiana

Table 4. Comparing the percentages of different racial groups whose DNA was collected from the estimated Indiana data and the actual Indiana data.

	IN (estimated)	**IN (disclosed)**
BIPOC to White	1.70	1.88
Black to White	2.80	3.09
Hispanic to White	0.45	0.68
Black to All BIPOC	1.65	1.64
Hispanic to All BIPOC	0.26	0.36

The disclosed data from Indiana give the racial composition of the group of 300,738 people who had their DNA collected. Table 4 compares this data with the estimated data.

The table shows a clear underestimate of the ratios between BIPOC and White persons. The ratios of other races to White persons are all lower for the estimated data compared to the disclosed data. Thus, the estimated data underestimate the discrepancy in DNA collection between BIPOC and White persons. In other words, there is an even greater disparity than predicted. However, most of the estimated ratios are similar to the ratios based on the disclosed data.

4. Maine

Table 5. Comparing the percentages of different racial groups whose DNA was collected from the estimated Maine data and the actual Maine data.

	ME (estimated)	**ME (disclosed)**
BIPOC to White	2.40	1.47
Black to White	5.27	2.78
Hispanic to White	1.33	0.32
Asian to White	0.00	0.32
Black to All BIPOC	2.19	1.88
Hispanic to All BIPOC	0.56	0.22
Asian to All BIPOC	0.00	0.22

The FOIA response from Maine gives the racial composition of the group of 33,711 people who had their DNA collected. Table 5 compares this data to the estimated data.

The table shows a series of stark overestimates, including the ratios between BIPOC and White persons and between individual groups. The closest estimate was the Black to all BIPOC estimate, although that figure is still a significant overestimate. Ultimately, the actual data show much less disparity than predicted.

5. Nevada

Table 6. Comparing the percentages of different racial groups whose DNA was collected from the estimated Nevada data and the actual Nevada data.

	NV (estimated)	**NV (disclosed)**
BIPOC to White	1.29	0.47
Black to White	3.47	2.01
Native American to White	1.24	0.76
Asian to White	0.29	0.20
Black to All BIPOC	2.68	4.27
Native American to All BIPOC	0.95	1.61
Asian to All BIPOC	0.23	0.42

The disclosed data from Nevada give the racial composition of the group of 344,097 people whose DNA was collected. Table 6 compares this data to the estimated data.

The table shows an overestimate of the ratios between BIPOC and White persons. The ratios between other races to White persons are all higher in the estimated data compared to the disclosed data. Thus, the estimated data overestimate the disparity in DNA collection between BIPOC and White persons. However, these ratios likely do not capture the actual ratios in the state given the lack of public data on Hispanic persons. For instance, if 50% of the population grouped under "White" would also identify as "Hispanic," then reducing the size of the "White" population and separately comparing the "Hispanic" category would cause the ratio between the White population and populations of persons of color to shrink considerably.

6. South Dakota

Table 7. Comparing the percentages of different racial groups whose DNA was collected from the estimated South Dakota data and the actual South Dakota data.

	SD (estimated)	**SD (disclosed)**
BIPOC to White	3.17	2.63
Black to White	5.32	4.12
Hispanic to White	0.00	1.50
Native American to White	4.40	3.01
Black to All BIPOC	1.68	1.57
Hispanic to All BIPOC	0.00	0.57
Native American to All BIPOC	1.39	1.14

The disclosed data from South Dakota give the racial composition of the group of 67,753 people whose DNA was collected. Table 7 compares this data to the estimated data.

The table shows an overestimate of the ratios between BIPOC and White persons. The ratios between other races to White persons are all higher for the estimated data compared to the FOIA data with varying degrees of departure. The estimated data overestimate the disparity in DNA collection between BIPOC and White persons. Contrary to Nevada, South Dakota disclosed figures for the Hispanic population, but public data sources used in our estimation did not provide that data. This may account for some of the discrepancies.

7. Texas

Table 8. Comparing the percentages of different racial groups whose DNA was collected from the estimated Texas data and the actual Texas data.

	TX (estimated)	**TX (disclosed)**
BIPOC to White	1.38	1.27
Black to White	3.05	2.68
Hispanic to White	1.05	0.97
Asian to White	0.00	0.10
Black to All BIPOC	2.21	2.12
Hispanic to All BIPOC	0.76	0.77
Asian to All BIPOC	0.00	0.08

The disclosed data from Texas give the racial composition of the group of 918,953 people whose DNA was collected. Table 8 compares this data to the estimated data.

The estimated data and the disclosed data from Texas lead to very similar ratios concerning the DNA collection of BIPOC and White persons. The estimated data show slight overestimates, but the numbers allow one to summarize the DNA collection disparity in Texas. The disclosed data also capture the population of persons of Asian descent, which was absent from our estimate.

D. Conclusion

Our estimated data were within a respectable range of the disclosed data. Nevertheless, it appears that our estimates sometimes faltered, especially with respect to smaller states where ratios from the estimated data departed more drastically from the actual data. It is possible that in those states, recent increases in the BIPOC population could cause recent DNA collection data to include a larger percentage of BIPOC than the total DNA collected in the states. It is also possible that the data in these smaller states are less accurate. Regardless, there is enough of a disparity between the estimated data and the actual data to affirm that certainty about the demographic composition of DNA databases will only come from deliberate decisions by those in control of the database to record and publish such information.

Nevertheless, both our estimated data and the disclosed data reveal a clear disparity between the DNA collected from BIPOC, especially from Black people, and the DNA collected from White people. The picture is even clearer when these data are aggregated to show the disparity nationally—or at least in a third of the nation—as in Figure 4.

Despite acknowledged shortcomings in our methodology, our estimates serve an essential purpose. In the absence of actual data about the racial composition of DNA databases—whether due to a failure to collect that information or a refusal to disclose it—scholars and policy-makers are left with speculation and conjecture. The more such speculation can be confirmed or discredited, the better tailored next steps can be. Moreover, frustration with the shortcomings of such estimates may inform debates about what data should be collected and how. The final Section of this Article further explores these implications.

III.
INSIGHTS FROM THE DATA AND ESTIMATES

Both the released data and the estimated data affirm the intuition that DNA databases contain disproportionate, sometimes dramatically disproportionate, profiles from particular communities. Across states, databases house DNA from

Black populations at rates twice or more their share of the state's general population. The only other group to approximate such disparity was Native Americans in the state of South Dakota. Interestingly, the disclosed data do not seem to support dramatically higher rates of DNA collection from the Hispanic community. However, the perceived parity may be due in all or in part to discrepancies in how or whether ethnicity is recorded.

The overrepresentation of Black populations in DNA databases likely comes as no surprise. After all, if DNA databases are tied to criminal justice policy and practice, and criminal justice policy and practices generate racially disproportionate rates of arrest and conviction, then the source of the disparity is obvious, and its solutions equally so. We might focus only on why and how arrests and convictions reflect as much racial disparity as they do and assume debates about the equities of the DNA database will be tethered to the answers.

While conversations around racist criminal justice policies and practices are certainly worthwhile, they overlook the specific need to reckon with racial bias and disparity in the particular context of DNA collection. In this instance, the disclosed data and our estimates provoke inquiry into fundamental questions specific to DNA policy and our understanding of racial equity in the *genetic* context. This Section addresses these questions in turn and considers the implications of these disparities as regards: (A) racial justice in DNA collection, retention, and search policies; (B) the biologization of race; (C) debates about data collection and centralization more generally; and (D) overarching questions of genetic privacy.

A. Implications for Racial Justice

1. Collection Policies

The database composition data highlight several policy questions that DNA collection practices raise. For instance, some scholars have speculated that expanding collection policies to include arrestees might diminish racial disparities in the database, particularly if collection occurs prior to a judicial finding of probable cause and the jurisdiction lacks automatic expungement provisions.[135] This theory is predicated on data that indicates more Whites are arrested than ultimately charged or convicted of qualifying offenses.[136] Other

135. *See, e.g.*, Kaye & Smith, *supra* note 43, at 454 ("Racial imbalance in the databases would be further reduced if, as leading law enforcement leaders have urged, arrest rather than conviction becomes the occasion for sampling DNA and including profiles in the database." (footnote omitted)).

136. *See id.* at 454–55 ("[E]xpanding DNA databases to include arrestees would diminish the racial disparity by bringing many more whites into the databases—about half of *all* males experience at least one misdemeanor or felony arrest in their lifetimes."); *see also* LINDSEY DEVERS, U.S. DEP'T OF JUSTICE, PLEA AND CHARGE BARGAINING 3 (2011), https://www.bja.gov/Publications/PleaBargainingResearchSummary.pdf [https://perma.cc/R7AF-CPBD] ("Studies that assess the effects of race find that blacks are less likely to receive a reduced charge compared with whites. Additionally, one study found that blacks are also less likely to receive the

scholars have argued the opposite. They claim that greater sampling from arrested persons, especially prior to judicial findings of probable cause, are likely to entangle more people of color because people of color are more likely to be baselessly entangled by overbroad arrest policies or excessive policing.[137]

The disclosed data, although limited, paint a complicated picture. Of the six states that have arrestee collection laws and disclosed data, only two states segregated their data by convicted persons versus arrestee (California and Texas). Both states' arrestee policies yielded slightly greater proportions of samples from Hispanic and White persons and lesser proportions from Black persons. In Texas, the general population is 41.5% White (not Hispanic), 13% Black, and 40% Hispanic. By comparison, the arrestee database is 45% White, 18% Black, and 33% Hispanic, and the convicted persons database is 37% White, 30% Black, and 33% Hispanic. A less dramatic but parallel dynamic is evident in the California data where the general population is 37% White (not Hispanic), 6.5% Black, and 39% Hispanic. In contrast, the arrestee database is 31% White, 14% Black, and 41% Hispanic, and the convicted persons database is 29% White, 18% Black, and 32% Hispanic. Thus, it appears that DNA policies that include arrestees tend to collect slightly higher rates of DNA from the White population than those that collect from convicted persons alone. However, the effects as regards the Hispanic population are less evident.

Interestingly, each state's arrestee laws differ in significant ways. Texas requires DNA samples from a complex array of persons.[138] At a basic level, Texas requires all convicted felons to provide a sample along with persons

benefits of shorter or reduced sentences as a result of the exercise of prosecutorial discretion during plea bargaining. Studies have generally found a relationship between race and whether or not a defendant receives a reduced charge." (citations omitted)); Carlos Berdejó, *Criminalizing Race: Racial Disparities in Plea-bargaining*, 59 B.C. L. REV. 1187, 1191 (2018) ("White defendants are twenty-five percent more likely than black defendants to have their most serious initial charge dropped or reduced to a less severe charge (i.e., black defendants are more likely than white defendants to be convicted of their highest initial charge). As a result, white defendants who face initial felony charges are approximately fifteen percent more likely than black defendants to end up being convicted of a misdemeanor instead. In addition, white defendants initially charged with misdemeanors are approximately seventy-five percent more likely than black defendants to be convicted for crimes carrying no possible incarceration, or not to be convicted at all." (footnotes omitted)). *But cf.* BESIKI LUKA KUTATELADZE & NANCY R. ANDILORO, VERA INST. OF JUSTICE, PROSECUTION AND RACIAL JUSTICE IN NEW YORK COUNTY ii (2014), https://storage.googleapis.com/vera-web-assets/downloads/Publications/race-and-prosecution-in-manhattan/legacy_downloads/race-and-prosecution-manhattan-technical.pdf [https://perma.cc/TYS4-SU5R] (finding that the New York City District Attorney "prosecutes nearly all cases brought by the police with no marked racial or ethnic differences at case screening," and that "[f]or all offenses combined, compared to similarly-situated white defendants, black and Latino defendants were more likely to be detained, to receive a custodial plea offer, and to be incarcerated; but they were also more likely to benefit from case dismissals").

137. *See, e.g.*, Risher, *supra* note 41, at 47–67 (detailing with precision the way that racialized enforcement and prosecution practices are likely to entrench racial disparities in DNA databases); Roth, *supra* note 40, at 308 (noting that arrestee DNA sampling will exacerbate the "implicit bias and explicit racism that create[s] inequity in every stage of the criminal justice process").

138. *See* TEX. GOV'T CODE ANN. § 411.1471 (West 2020). In addition, some arrestee sampling laws are only triggered if a person has a prior conviction.

convicted of certain misdemeanors such as public lewdness, indecent exposure, terroristic threats, or promoting prostitution. In addition, persons indicted for various sex offenses and specified burglary or kidnapping and persons arrested for—having been previously convicted of—certain sex offenses must also provide DNA samples. Thus, in Texas, an indictment is required before DNA profiling arrestees who have not been previously convicted of eligible offenses. In contrast, California permits collection for all felony arrestees at booking prior to any formal finding of probable cause by either a grand jury or judicial officer. Finally, Texas provides automatic expungement whereas California allows expungement upon request.

It is difficult to know how much the differences in racial disparities among arrestee collection are due to the state's collection policies. In both Texas and California, there is a fairly significant difference between the arrestee and convicted persons demographics. Texas's disparity for arrest versus conviction in the Black population is starkest. The Black fraction of the arrestee database much more closely resembles the Black population's fractional share in general (18%), and is nearly half the convicted persons percentage (30%). It may be that this reinforces a funneling effect wherein Black persons are more likely to be processed through the system and convicted than persons of other races or ethnicities. It may also relate to the specific profile of persons covered by the arrestee policy; for instance, sex offenders as a class might be more likely to skew White. Or it could reflect the higher standard—indictment—for collection from Texas arrestees.

However, it is difficult to conclude from these data alone the effect of any one policy, much less isolate which feature of the policy (e.g., qualifying offenses, probable cause finding requirement, automatic expungement) is responsible for such effect. Moreover, from the limited data provided, it was too difficult to conclude whether different states' policies (e.g., conviction only versus arrest or conviction) are more or less likely to result in disparate database composition. But this information does suggest that more comprehensive data might yield answers to these important policy questions.

Notably, the data also reveal that a state's size and diversity had little effect on the size or scope of the observed disparity in the DNA database. The DNA databases generally included two to three times as many Black persons as demographically proportionate regardless of the size of the state or whether Black persons constituted a large or small fraction of the general population. Interestingly, however, the data regarding the Hispanic population showed much less disparity. In California, Texas, and South Dakota, the fraction of Hispanics in the database roughly matched their share in the general population while in Maryland and Nevada it was significantly less. Of course, there are reasons to think that in some states these data were too unreliable (due to difficulties in the way in which that category is reported) to draw clear conclusions.

At a most fundamental level, these data confirm that states have stockpiled genetic material from a large and demographically disproportionate share of the Black population, which should inform debates about collection, retention, and testing policies. If Black persons constitute 13% of the population but 24% of the database while White persons are 60% of the population and only 43% of the database, then decisions about the retention and storage of samples will have an outsized effect on the Black population. Ensuring democratic accountability over search and retention policies, which unlike collection policies tend to be set by the executive rather than legislative branch, may thus require added care to ensure diverse viewpoints are heard. The need for broader sources of legitimacy and accountability may also dictate that search and retention rules be fixed legislatively rather than by executive fiat.

Finally, with improved data collection practices—for instance, linking individual contributions with qualifying offenses, or recording whether particular profiles resulted in additional solved crimes or even charges—we might be able to draw conclusions about how best to optimize DNA databases while minimizing their racially disparate impact. This might also make it easier to track trends such as which populations commit DNA-eligible offenses or whether developing or improving testing methods might improve detection rates. As noted in Part I, current systems fail to track even the most basic measurements for an efficacy analysis. At the same time, DNA databases exhaust greater and greater resources as they move unilaterally in the direction of expansion without any meaningful analysis of costs and benefits.

2. Search Policies

More information about DNA database composition not only increases our understanding of optimal collection policies, but also allows us to better assess the wisdom of *search* policies, specifically those pertaining to familial searches. As Part I recounts, critics have called for the release of profiles held in the national DNA database. This would allow researchers to interrogate the assumptions underpinning match statistics, including "the extent to which DNA profiles cluster due to identity by descent."[139] Most pointedly, there has been marked debate over the wisdom of allowing familial searches in forensic DNA databases. At present, eleven states explicitly authorize such searches and two jurisdictions (Maryland and the District of Columbia) expressly forbid them.[140]

---

139. *See, e.g.*, D. E. Krane et al., *Time for DNA Disclosure*, 326 (5960) SCIENCE 1631, 1631 (2009).

140. MD. CODE ANN., PUB. SAFETY § 2–506(d) (West 2020); D.C. CODE § 22–4151(b) (2020). *See also* Graham Rayman, *Legal Aid Lawyers Challenge New York's Use of Familial DNA Testing*, DAILY NEWS (Feb. 16, 2018), https://www.nydailynews.com/new-york/legal-aid-lawyers-challenge-n-y-s-familial-dna-testing-article-1.3823989 [https://web.archive.org/web/20200802191341/https://www.nydailynews.com/new-york/legal-aid-lawyers-challenge-n-y-s-familial-dna-testing-article-1.3823989] (highlighting the debate over familial

Familial searches effectively turn national DNA databases into genetic informants.[141] By searching for near (rather than exact) profile matches to forensic samples, analysts can come up with a list of possible persons who may be a relative of the perpetrator. One of the greatest concerns about such searches is their disparate impact. If DNA databases are racially disproportionate, then certain communities will regularly fall under "genetic suspicion" while others go unbothered.

The disclosed data confirm the intuition that, at least for the Black population, the dramatic overrepresentation in DNA databases opens greater shares of that community to suspicion using familial searches. Furthermore, because DNA profiles are retained indefinitely, the net of coverage from a familial search will continue to expand as later generations come of age. Thus, to the extent that there are privacy concerns in amassing a database of genetic profiles that disproportionately includes entries from particular populations, such concerns are exacerbated given that principles of genetic inheritance mean such disparity will be reproduced and magnified so long as familial searches are permitted.

Research models show that "individuals from certain marginalized groups may be disproportionately more often subject to false familial identification."[142] This is because errors in assumptions about the allele frequency distributions (the predictions about the commonness or rarity of certain traits in a population) cloud the assessment of relatedness.[143] Moreover:

> Because some of these groups (Native Americans and some immigrant groups) are correlated with social groups already over-represented in the criminal justice system, group members would be more likely to have a relative in the database, and that relative would be more likely to have a coincidental partial match with a crime scene sample.[144]

To the extent that the disclosed data affirm that Native Americans in certain states (such as South Dakota) are dramatically overrepresented in DNA databases, policy-makers and advocates might want to caution against search policies (whether familial searches or searches for incomplete profiles) that have a higher probability of falsely implicating members of that group.

DNA searches in New York and noting that Maryland and the District of Columbia had prohibited such searches).

141. *See generally* Erin Murphy, *Relative Doubt: Familial Searches of DNA Databases*, 109 MICH. L. REV. 291, 320 (2010) (explaining that familial searching may turn databased persons into involuntary "genetic informants").

142. Rori V. Rohlfs et al., *Familial Identification: Population Structure and Relationship Distinguishability*, at 2, PLOS GENETICS (Feb. 9, 2012), https://doi.org/10.1371/journal.pgen.1002469 [https://perma.cc/9PYA-M9Q4].

143. *Id.*

144. *Id.* at 9.

B. The "Biology" of Race

The disclosed and estimated data also expose some more fundamental questions about race and ethnicity in relation to criminal justice. They shed light on how and why we talk about racial or ethnic categories generally in criminal justice and, specifically, forensic genetics. As summarized by scholar David Skinner:

> A recurring theme of work on racialization and the new genetics is the slipperiness of race as an object of expert and public discussion. Scientists use racial and ethnic categories while acknowledging these to be flawed and contentious. They accommodate to local, common sense understandings of difference and often willingly acknowledge that race is a "social construct."[145]

These tensions between scientific and folk knowledge, or between bio-geographical truth and social understanding, surface directly in the composition of DNA databases. The difficulty in drawing sustainable inferences from these tensions further underscores how impoverished our vocabulary is with regard to race and policing.

For instance, there were subtle variations in the ways in which the disclosed data by race or ethnicity were reported. All states reported figures for categories we might loosely label "White," "Black," and "Asian." However, the way these categories were labeled varied. For instance, states varied in using general descriptors like "White" and "Black" as opposed to pseudo-geographical words like "Caucasian" and "African American." The disclosures also did not indicate how groups were constituted—either as a categorical matter (i.e., what constitutes "White") or as applied (i.e., what determines whether a particular contributor is "White"). History teaches that these categories have changed over time: "White" today was not necessarily "White" yesterday.[146] Nonetheless, the categories were perceived as sufficiently immutable that both categorization and classification were deemed possible. To be sure, some states problematized the data by noting that it was self-reported.[147] California, moreover, added that "racial classification is not considered a required field on the collection card" and "the Department of Justice does not verify the accuracy of reported racial classifications."[148] Rather, such data are "either self-reporting by the offender or

145. David Skinner, *Race, Racism and Identification in the Era of Technosecurity*, 29 SCI. AS CULTURE 77, 82 (2020).

146. *See generally* NELL IRVIN PAINTER, THE HISTORY OF WHITE PEOPLE 72–90 (2010) (discussing coinage of the "Caucasian" ideal, which then included Europeans and many Scandinavians, excluding descendants of the Lapps, or modern-day Finnish; North Africans; Indians; and those from certain parts of Russia); *id.* at 139 (quoting Ralph Waldo Emerson's list of "races" who can never "occupy any very high place in the human family," which included that "[t]he Irish cannot; the American Indian cannot; the Chinese cannot"; rather, "[b]efore the energy of the Caucasian race all the other races have quailed").

147. *See, e.g.*, Indiana Letter, *supra* note 114 ("Race and Ethnicity are self-reported.").

148. California Letter, *supra* note 114.

speculation on the part of the law enforcement officer supervising the collection."[149]

But generally, the lack of rigor in either devising or implementing such categories reveals how race as a category is both critically important and largely meaningless. At present and throughout history, race reflects important social and political differences in the lived experiences of individuals—particularly as regards policing and criminal justice. And yet, as a scientific category, race has weak traction at best. It is significant that, as Professor Kahn pointed out, there is considerable sloppiness in racial and ethnic categories as used and ascribed in forensic genetics.[150] When a DNA match report lists a series of alleles that make up the genetic profile and then reports match statistics in terms of racial or ethnic categories like "Black" or "African-American," it can imply that the scientific foundation that undergirds the determination of the profile similarly buttresses the match statistic. In reality, the act of devising and applying these categories is characterized by casual intuition, not scientific expertise.

Consider, for instance, that no state reported "multiracial" as a category, even as current data suggest roughly 6.9% of the population is multiracial (a percentage higher than many of the other categories reported).[151] Multiracial contributors were either sorted into one of the component identities or placed as "Other"; interestingly, only Florida separated "Other" from "Unknown." This incapacity to deal, at the most fundamental level, with the sizeable portion of the population that claims multiracial or multiethnic heritage speaks volumes about the states' attachment to rigid divisions despite biological or social reality.[152] Similarly, states had divergent approaches to the category of "Hispanic." Some states appeared to treat it as though it were a separate racial category, independent of White, Black, Asian, or Native American. Others may have double-counted by allowing an individual to elect both a racial and ethnic identity. One state, despite its sizeable Hispanic population, seems to have generally ignored it as a demographic identifier. The variety in these responses suggests that most "racial" categories in fact code as ancestral geographical categories—and even the great

149. *See id.*

150. *See* Kahn, *supra* note 95, at 346–47 ("The casual and perfunctory assignment of social categories of race to biological samples in professional discussions of forensic DNA stands in marked contrast to the meticulous care taken concerning the more technical aspects of DNA extraction, amplification, and analysis.").

151. Current data estimate roughly 6.9% of the population is multiracial, and birth data from 2013 show 10% of babies born that year were born to parents who self-identify as from different demographic groups. Kim Parker et al., *Multiracial in America: Proud, Diverse & Growing in Numbers*, PEW RESEARCH CTR. (June 11, 2015), https://www.pewsocialtrends.org/2015/06/11/multiracial-in-america/ [https://perma.cc/53TU-RW83].

152. Compare, for instance, the United Kingdom's practice of reporting census figures not just for "Mixed/multiple ethnic groups" but for specific subgroups like "White and Black Caribbean," "White and Asian," and "White and Black African," and also for breaking broad categories like "Asian/Asian British" and "Black/African/Caribbean/Black British" into ancestral subcategories. OFFICE FOR NAT'L STATISTICS, *supra* note 33, at 3.

American melting pot surrenders even the pretense of categorical certainty when the geography gets complicated.

In this regard, it makes sense that states varied—albeit slightly—in the categories they selected for data gathering. Although all states reported figures for White, Black, and Asian populations, some states also reported figures for Native Americans or Hispanics. Significantly, these fluctuations did not seem to reflect the percentage share of these groups in a state's population. Florida, with its relatively small Native American population, reported figures for that group while California did not. On the other hand, Nevada, with its large Hispanic population, did not report figures for Hispanics while Indiana, with its small population share of Hispanics, did. It is difficult to discern the significance of these choices; what emerges is only that the salience of particular racial or ethnic categories is variable, even if what drives that variability remains opaque.

Relatedly, the breadth of the categories reported also seems to reflect socially contingent ideas. The "Asian" category was universally used; yet descriptively, that category could sweep in an enormous number of individuals.[153] States either perceived the salience of "Asian" for criminal justice purposes to be low—with no utility in breaking that group into constituent parts—or considered the share of the population *in toto* so small that disaggregating it would be to risk unhelpful fragmentation. Nevertheless, subsuming a numerically small subgroup under the broad tent of "Asian" could mask unique disparities experienced by that subgroup. In other words, simply the act of categorizing may serve to reveal or mask existing biases.

By comparison, the United Kingdom's figures show somewhat greater precision. In defining the category "Asian," the United Kingdom distinguishes the categories of Middle Eastern from "Chinese, Japanese, and SE Asian."[154] Similarly, the United Kingdom breaks "White" into Northern European and Southern European,[155] a categorical distinction that would likely quickly collapse if attempted in the U.S. context.[156] Interestingly, the United Kingdom sorts by

153. *See generally* Anna Purna Kambhampaty, *At Census Time, Asian Americans Again Confront the Question of Who 'Counts' as Asian. Here's How the Answer Got So Complicated*, TIME (Mar. 12, 2020), https://time.com/5800209/asian-american-census/ [https://perma.cc/ME56-5LN5] (noting that the census defines the estimated 20 million Asians in the United States to include persons "having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam," noting disagreement among Whites polled over whether persons of Indian or Pakistani descent should be included as "Asian").

154. HOME OFFICE, *supra* note 28, at 17 figs.3b, 4b.

155. *See id.* The salience of particular groups within a population is most visible in one scholar's account of the "16+1" categories used in the 2001 U.K. census: "Indian, Pakistani; Bangladeshi; Other Asian; Black Caribbean; Black African; Other Black; Chinese; Other ethnic group; Mixed White and Black Caribbean; Mixed White and Black African; Mixed White and Asian; Other Mixed; White British; White Irish; and Other White." Skinner, *supra* note 33, at 985.

156. It is true that ancestral genetic patterns are evident in surveys of regional DNA data that reflect the history of immigration in that area. *See* Katarzyna Bryc et al., *The Genetic Ancestry of African Americans, Latinos, and European Americans Across the United States*, 96 AM. J. HUM. GENETICS 37,

"ethnic appearance," suggesting that such groups are readily identified through a superficial appraisal.[157] The self-conception of the "races" within the society—their relative stability and ascertainability—is evident in the manner in which the DNA database records and reports this information.

In sum, efforts to glean the racial composition of DNA databases ultimately unmask the shallowness of racial categorization, yet they also cannot be dismissed given the depth of the real problem of racial discrimination in criminal justice. It is true that reported racial categories are weakly defined and even more weakly populated. At the same time, these poorly constructed categories have much to teach us about the salience of particular characteristics in our cultural and political context and are imperative to assessing the actual impact of DNA policy within those culturally and politically identifiable communities. As David Skinner observed in connection with the categories reported from the U.K. National DNA Database, or NDNAD:

> The NDNAD example tells us much about the novelty of contemporary biopolitics and the 'reinscription of race' associated with new genetics. The ethnic categories used in its operation and debate, like those in other areas of contemporary genomics, cannot be thought of as either purely social or biological. These categories are hybrid, mutable boundary objects that move back and forth between scientific, governmental and political domains. DNA is implicated in the politics of ethnicity, racism and criminal justice without a presumption that criminal behavior, or indeed 'race', has a biological basis.[158]

Thus, it cannot be said that race and ethnicity are irrelevant or even necessarily that collecting such data with greater rigor is the solution. A "race-neutral" database is as disingenuous as a "racially scientific" one, because race matters in criminal justice. Rather, the key is to walk the precarious line between social and biological ideas of race in order to reach a better understanding of the function of criminal justice systems. At the same time, one must acknowledge the danger of inadvertently allowing the robust science of genetic testing to lend credibility to the feeble socio-cultural practice of racial and ethnic sorting. Without attentiveness to how and why these categories exist, DNA science may inadvertently legitimate racialized ideas of biological determinism. In simpler terms: one should pay attention to the racial composition of DNA databases while

---

49 (2015) ("The distributions of the European subpopulation ancestries in European Americans illustrate that the distribution of within-European ancestry is not homogenous among individuals from different states, and instead, reflects differences in population migrations and settlement patterns across the US."). However, in the United States there are high rates of intermarriage within ancestral groups of European descent. *See, e.g.*, Dribe Martin et al., *Becoming American: Intermarriage During the Great Migration to the United States*, 49 J. INTERDISCIPLINARY HIST. 189, 193 (2018) ("Most groups of European origin showed high rates of intermarriage with the native-born population and a clear trend over time to more intermarriage and less endogamy.").

157. Skinner, *supra* note 33, at 981.

158. *Id.* at 987 (citations omitted).

taking care not to reinforce reflexive and unfounded ideas about the relationship between biology and race.

C. Data Centralization, Data Diffusion, Data Ignorance

The disclosed data also illuminate debates about genetic privacy, one of the most enduring and difficult questions in forensic DNA testing. Since its inception, the national DNA system has served only as a pointer system with decentralized data. As a result, the "national database" is nothing more than a collection of profiles attached to lab, analyst, and case identifiers; the FBI cannot disclose the racial composition of the national database because it does not keep that information. On the other hand, decentralization makes it difficult to carry out widespread privacy attacks such as a hack of the national database, simply because the task is logistically complicated. Indeed, even compromising each state-level database is difficult. The FBI requires that access to the database be limited to specialized computers licensed to run the software and that those computers be physically behind a locked door, accessible only by specific authorized personnel. Moreover, once personnel gain access, the actual information they can retrieve (especially beyond their own state borders) is of limited value—they can only get a series of numbers, which the FBI then links to a specific lab and specific case.

The deliberate decision not to standardize the data beyond a bare minimum of a lab identifier also means that there is no uniformity in how DNA samples are tracked. As our disclosure requests demonstrate, one state might elect to collect demographic data and associate that with its internal DNA profiles, while another determines not to do so. Because the FBI requires only basic pointer information, there is no threshold of added data that a state must have to associate with any particular profile, much less any consistency in how such data is recorded or transcribed. As a result, it is impossible to get a national or comparative picture of the DNA database drawn with any meaningful clarity, and the state or local picture is often obscured as well.

However, the advent of Rapid DNA testing is likely to change some of these practices. Once regulations are in place to permit DNA testing by rapid machines outside of the laboratory context—for instance, in police precincts and perhaps even squad cars or mobile crime labs—the demand for speed in the uploading and matching process will increase. For most of the DNA database's history, submissions were sent manually. Twice weekly, the national database searched within itself and reported any existing matches. But with Rapid DNA, the urge to allow instant digital uploads and real-time searching is enhanced. In fact, the FBI already has plans in place to craft a parallel national DNA database for serious crimes, called the DNA Index of Special Concern. [159] In short, Rapid

159. Tom Jackman, *FBI Plans 'Rapid DNA' Network for Quick Database Checks on Arrestees*, WASH. POST (Dec. 13, 2018), https://www.washingtonpost.com/crime-law/2018/12/13/fbi-plans-rapid-

DNA may lead to a database system that is centralized and cross-linked to other bio-identifiers or personal information.

In this respect, the disclosed data raise numerous questions surrounding the data collection, retention, and search policies for DNA samples. There are obvious tradeoffs: greater transparency may enhance the risk of abuse while ignorance may serve as protection. This Article's glimpse into one set of questions—namely, how privacy and transparency may impact the racial composition of DNA databases—underscores how these arguments tilt in both directions.

On the one hand, the determination to strip almost all useful information from profiles submitted to a centralized repository helps guard against unauthorized access, use, or release of information. The FBI has defended its decentralized approach to the national DNA database as chiefly about protecting privacy.[160] That is, the danger posed by unauthorized access to over 15 million DNA profiles in the database is greatly diminished if there is little to glean from a break-in. Decentralization also minimizes the risk of net-widening and function creep, or pressure to expand the use of the data beyond the reasons for which it was initially collected, because the narrow scope of the collected information precludes ready expansion. In addition, given that the physical biological specimen is stored when DNA is collected, the decentralized approach circumvents pressure to re-examine such samples—whether for efficiency as new tests develop or for nefarious reasons. Specifically, broad testing campaigns are difficult when biological samples are stored not only away from a centralized repository but also in a web of laboratories even within a single jurisdiction.

Decentralization may also mitigate against bias. A person searching the database cannot execute searches or initiate investigations on the basis of race if such information is not readily available. Tellingly, California's disclosure took pains to point out that the provided information resides in a separate database that "is not searched for criminal identification purposes," and thus "race is not, and cannot, be used as a search criterion when operating CODIS databases, and does not appear in any search result."[161] In divorcing demographic data from biologic data, law enforcement helped blunt against the abuse of that data.

But it is important to acknowledge that, notwithstanding these benefits, the decentralization of data related to DNA profiles, and the associated fluctuations in what data is collected, poses serious limits on the capacity to assess the costs and benefits of our DNA policies. Scholars, researchers, and policy-makers cannot answer basic questions regarding the demographic composition of the

---

dna-network-quick-database-checks-arrestees/?utm_term=.515e0e005ac8 [https://perma.cc/MCE3-36K9].

160. *See, e.g.*, United States v. Mitchell, 652 F.3d 387, 400 (3d Cir. 2011) (upholding DNA database in part because "[t]he FBI's restrictions on the type of information stored in CODIS reflect Congress's concern about creating 'strict privacy protections.'" (citation omitted)).

161. California Letter, *supra* note 114.

national DNA database. They have even more difficulty assessing what kinds of offenses are most commonly qualifying, whether certain offense types or offender characteristics are more or less likely to lead to matches or cold hits, or whether certain collection or search policies are more or less efficacious. The lack of standardization also makes comparison across jurisdictions difficult. Indeed, the refusal to collect certain pieces of information may render those categories effectively immune from scrutiny, even regarding the fundamental issue of whether DNA databases are racially disproportionate. At most, observers can speculate about conditions, but that speculation is always abstract and uncertain. And finally, stripping race from the database, although ostensibly a protection against bias, may in fact help insulate existing bias from review. Concrete data often have a more powerful rhetorical impact than abstract ideas do. Decentralization forecloses uncomfortable conversations about collection policies and practices, whether along the lines of racial disparity or other demographic dimensions.

In line with the debate over genetic privacy, DNA databases reflect three different models through which to understand government collection and use of big data, which we might call a centralization model, a diffusion model, and an ignorance model. The centralization model has the benefit of allowing information optimization because the data can be sliced a million different ways to enhance understanding. However, this model poses the greatest risk of abuse by both governmental and non-governmental actors. The ignorance model cuts dramatically in the other direction. The deliberate decision *not* to know more about volatile data ostensibly neutralizes its most lethal form. Yet it is deeply unsatisfying to know that large quantities of genetic information are amassed from members of our society in a manner that defies much objective assessment of its merits. Between these two poles, a diffusion model may seem like a happy compromise: amass the big data, but in a way that makes it more (although not wholly) impenetrable to misuse.[162] If each state is required to collect and hold the data, then the data can only be collected and examined with painstaking effort. Conversely, any incidents of abuse or error are more likely to be contained. In this way, the national DNA database architecture could satisfy Paul Ohm's and Jonathan Frankle's category of a "desirably inefficient" system—a system in which efficiency is sacrificed in service of other goals.[163]

---

162. The Department of Health and Human Services' efforts to implement the national Sentinel System—a nationwide database of health and insurance records that allows the FDA to "monitor the safety of FDA-regulated medical products"—confronted debates about centralization and decentralization early in its inception, and ultimately implemented a "distributed database" approach. *FDA's Sentinel Initiative*, FDA, https://www.fda.gov/safety/fdas-sentinel-initiative [https://perma.cc/3XC9-9USD]; *see generally* Barbara J. Evans, *Congress' New Infrastructural Model of Medical Privacy*, 84 NOTRE DAME L. REV. 585, 606 (2009) (describing "archipelago" of data).

163. Paul Ohm & Jonathan Frankle, *Desirable Inefficiency*, 70 FLA. L. REV. 777, 821–22 (2018). Ohm & Frankle cite "decentralization" as one of the four distinctive attributes of desirably inefficient systems. *Id.* at 815.

Yet at least in the genetic context, diffusion as a solution may be an illusion. First, like many compromises, it is susceptible to criticism and thus pressure from both sides. Privacy advocates may worry that the DNA databases are still too invasive, while law enforcement advocates will question why they are not more useful. Second, while technological and commercial forces helped propel Rapid DNA testing, this change was also policy-driven.[164] Once matching in any form began, enthusiasm grew for immediate testing to find matches.[165] Thus, the diffusion model for big data may always be moving toward making as much information as possible fully accessible. Diffusion may be less a bulwark than a speed bump.

Moreover, genetic data leaks in a way that other forms of data may not, filling in blanks otherwise deliberately left open. It is increasingly evident that ostensibly de-identified or purposely limited genetic samples can, rather readily, be re-identified or extrapolated. For instance, certain genetic markers have been shown to correspond to surname prediction quite reliably.[166] Another set of researchers determined that they could take an ostensibly anonymized dataset of CODIS loci and a similarly de-identified dataset of genealogical markers, and match records in 90% to 98% of cases.[167] Still other researchers have shown that, given genealogical markers and publicly available data, it is quite easy to identify specific individuals by name.[168] If it is possible to take two data sets and quickly reunite them, then the diffusion model becomes an even greater illusion. It may complicate the task of weaponizing information, but it does not foreclose it. Anonymous big data may be an illusory idea to begin with, and thus a conversation about tradeoffs between the ignorance and transparency model—rather than outright data protection—simply misses the mark.

D. The Illusion of Genetic Privacy

The disclosed and estimated data also shed light on deeper questions about the conceptualization of genetic privacy. Our focus on the racial composition of forensic databases and their potential inequities derives from a series of underlying assumptions: (1) the government's ability to compel people to give DNA samples matters because the government is an actor worthy of special attention, because it possesses unique power to compel otherwise unavailable and thus valuable information; and (2) the government's exercise of this power in a potentially discriminatory fashion is both intrinsically and consequentially bad. But are such assumptions accurate, and if not, are debates about the racial

164. MURPHY, *supra* note 4, at 164–66 (describing utility of and lobbying interests for rapid systems).

165. *Id.* at 187 (describing pilot programs aimed at on-site testing).

166. *See, e.g.*, Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, 339 SCIENCE 321, 322 (2013).

167. Jaehee Kim et al., *Statistical Detection of Relatives Typed with Disjoint Forensic and Biomedical Loci*, 175 CELL 848, 852 (2018).

168. *See* Gymrek, *supra* note 166, at 324; Erlich, *supra* note 69, at 692.

composition of law enforcement DNA databases by those concerned about racial equity misguided?

In other words, statements about the racial composition of forensic DNA databases or the potential inequities of law enforcement amassing a treasure trove of genetic data rest heavily on the assumption that forensic DNA databases *matter*—that they give law enforcement access to information it would not have otherwise, or that they represent a special incursion into a person's privacy. But if other sources of genetic information eclipse forensic databases in their utility, then should concerns about racial or ethnic inequities (at least as regards law enforcement databases) disappear? Or does there remain something special about the government's power to compel or store genetic information even when such information is already available in the public sphere?

The issue presents itself chiefly as a result of the rise of recreational and commercial genetics in the form of testing and databasing entities such as 23andMe, Ancestry.com, MyHeritage, and GEDMatch. According to one survey, more people took consumer genetic tests in 2018 than in all the previous years combined,[169] and some experts predict that the direct-to-consumer genetic testing market will be worth more than $2.5 billion in 2024.[170] More importantly, the type of genetic testing that these entities engage in—looking at hundreds of thousands of single nucleotide polymorphisms (SNPs)—far outstrips the identification and informational capacity of the forensic standard, which is 20-loci short tandem repeat (STR) testing. And of course, both methods are fully eclipsed by whole-genome sequencing, which may one day become the standard of care for clinical medicine.

Although there are logistical obstacles that prevent law enforcement from using commercial testing databases in high volume, those obstacles may not always stand. By way of a basic illustration, suppose the Supreme Court upholds as lawful a police officer's power to subpoena a commercial database for identifying information of a person who matches (or perhaps is a likely close relative of) a forensic sample's genetic profile. Then, it would not be difficult to imagine that police would do "John Doe" searches in commercial databases with regularity. Such searches might even become the first, rather than last, line of investigation. Indeed, law enforcement have already solved an impressive

169. Antonio Regalado, *More Than 26 Million People Have Taken an At-home Ancestry Test*, MIT TECH. REV. (Feb. 11, 2019), https://www.technologyreview.com/s/612880/more-than-26-million-people-have-taken-an-at-home-ancestry-test/ [https://perma.cc/L82W-8PGV].

170. SUMANT UGALMUGLE & RUPALI SWAIN, GLOBAL MARKET INSIGHTS, DIRECT-TO-CONSUMER (DTC) GENETIC TESTING MARKET SIZE BY TEST TYPE (CARRIER TESTING, PREDICTIVE TESTING, ANCESTRY & RELATIONSHIP TESTING, NUTRIGENOMICS TESTING), BY DISTRIBUTION CHANNEL (ONLINE PLATFORMS, OVER-THE-COUNTER), BY TECHNOLOGY (TARGETED ANALYSIS, SINGLE NUCLEOTIDE POLYMORPHISM (SNP) CHIPS, WHOLE GENOME SEQUENCING (WGS)) (2020), https://www.gminsights.com/industry-analysis/direct-to-consumer-dtc-genetic-testing-market [https://perma.cc/T7ZK-NLST].

number of cold cases using forensic genealogical methods, notwithstanding that those methods are resource-intensive in their current iteration.[171]

More pertinently, the Supreme Court's reasoning in *Maryland v. King*—which upheld the constitutionality of requiring arrested persons to contribute DNA to a law enforcement database[172]—seems to leave open the possibility that the government could compel an individual to give a DNA sample under a much wider array of circumstances.[173] If genetic information is viewed as a neutralized form of "identity" akin to a fingerprint, as the majority in *King* argued,[174] then there is no reason why genetic databases would not mirror the scope of fingerprint databases and cover everything from driver's license holders to student loan recipients to state employees or licensees.

If genetic databases continue to proliferate in this way, then focusing on the particular composition of law enforcement DNA databases misses the mark. For those interested either in privacy or racial justice, the core concern is law enforcement's use of or access to *any* genetic information (or particular kinds of information), rather than the specific act of compelling and stockpiling DNA profiles from particular people.

In this respect, the debate over the wisdom of universal databases and their likely impact as regards racial equity is illustrative. As recounted in Part I, advocates for a universal DNA database (i.e., a law enforcement database containing profiles from everyone in the population) often primarily defend that position by citing concerns about racial equity. If everyone is required to be in the database, the reasoning goes, then the adverse effects of acknowledged policing and enforcement biases will be blunted. Police will be chastened in their use of the database because they know it contains the powerful and not just the weak. And as an expressive matter, there will no longer be a need to reckon with the discomfort of amassing genetic profiles of an underclass or of racial and ethnic minorities because everyone will be in the database together.

Critics, however, argue that universal databases do little to rectify the real problem with racially disparate policing, which is not that more people of color qualify for inclusion in DNA databases but that police use their policing discretion in racially disparate ways.[175] For example, a universal database may help police solve "all" marijuana possession, underage drinking, assault, or domestic violence cases, but it will do nothing to change whether police enforce drug laws in suburban White enclaves or respond to domestic violence calls in majority-minority communities.

---

171. *See generally* Murphy, *supra* note 75.

172. *See* 569 U.S. 435, 465–66 (2013).

173. Murphy, *supra* note 75, at e7–e8.

174. *See* 569 U.S. at 465–66.

175. Dorothy Roberts, *Collateral Consequences, Genetic Surveillance, and the New Biopolitics of Race*, 54 HOW. L.J. 567, 586 (2011) ("Although DNA testing can correct injustices when used narrowly to confirm a suspect's guilt or innocence, the massive genetic surveillance we are witnessing threatens to reinforce the racial roots of the very injustices that need to be corrected.").

Therefore, it may in fact be salient that it is the government, and in particular the police, that has collected, tested, and stored genetic data—even if similar or better data is publicly available for law enforcement purposes elsewhere. The special power of police, and their particular history of wielding criminal law as a tool of oppression,[176] matters. Even if the physical incursion is trivial (a cheek swab) and the informational intrusion nonexistent (because the government could amass the same information another way), the simple fact that the government has targeted a particular class of persons in a racially disparate manner is enough to raise alarm.

Analogy might be drawn to the reasoning in cases such as *Brown v. Texas*[177] and *Hiibel v. Sixth Judicial District Court*.[178] Those cases underscore that "stop and identify" statutes, which require a person to give their name or address upon request to a law enforcement officer, are only permissible if the initial stop is lawfully predicated on reasonable suspicion or probable cause.[179] As with DNA testing, the physical intrusion is minimal and the informational intrusion—the name or address—may often be readily ascertained in other ways (and, indeed, in many cases may already be in the hands of another government agency, such as the Department of Motor Vehicles). But even this simple act is restricted, ostensibly due to the concern over law enforcement having unfettered discretion to choose whom to stop and penalize for noncompliance.

Importantly, however, nothing in the Court's opinions in *Hiibel* and *Brown* forbids police from engaging in the behavior of briefly stopping someone on the street and asking their name. The Court simply precludes the state from imposing criminal penalties when an individual chooses not to comply. Thus, these cases only foreclose the arbitrary imposition of punishment, not the arbitrary amassing (or utilization) of information, by police.

The debate around DNA databases can be conceived similarly. At one extreme are arguments that, at most, the Constitution ought to monitor closely the government's power to engage in even these small incursions on liberty and informational privacy, without regard to considerations such as the volatility of the pairing between genetics and policing. At the other end of the extreme are arguments that the Constitution ought to monitor closely the government's power to engage in these small incursions into liberty and informational privacy precisely *because* of the volatility of that pairing. But in a world of increasing

176. *See generally* DOUGLAS A. BLACKMON, SLAVERY BY ANOTHER NAME 99 (Anchor Books 2009) (describing the "application of laws written to criminalize [B]lack life"); GILBERT KING, DEVIL IN THE GROVE (2013) (describing use of criminal law, and extralegal action or inaction by criminal justice actors, to enforce White supremacy).

177. 443 U.S. 47 (1979).

178. 542 U.S. 177 (2004).

179. *Brown*, 443 U.S. at 52 ("In the absence of any basis for suspecting appellant of misconduct, the balance between the public interest and appellant's right to personal security and privacy tilts in favor of freedom from police interference."); *Hiibel*, 542 U.S. at 182, 186 (upholding a "stop and identify" statute because the stop had to be predicated on reasonable suspicion).

genetic transparency, both sides of the debate may be missing the fundamental point: what matters is not the government's compulsory collection power and its disparate impact but police power more generally and its discriminatory impact.

CONCLUSION

This Article ultimately pursues two contradictory sets of ideas. On the one hand, it endeavors to quantify more precisely the racial and ethnic composition of DNA databases, in part to directly engage conversations about equity and privacy in forensic genetics. On the other hand, it uses the disclosed data and our own estimates to engage a broader set of questions regarding the measurement of efficiency in DNA database systems, the problematic categories of "race" and "ethnicity" in criminal justice, the optimal structures for storing biometric data, and the myth of genetic privacy generally. In short, this Article painstakingly generates data even while questioning the wisdom and utility of doing so.

At minimum, it feels irresponsible to allow the criminal justice system—with its tainted history of using state power to oppress marginalized populations and with the enduring legacy of that inequity still so manifest today—to be distanced from frank conversations about race and ethnicity in forensic genetics. At the same time, those conversations seem quaint as the era of wholesale genetic transparency approaches, outmoded as the fixed and tidy categories of "race" and "ethnicity" become increasingly suspect, and obtuse as data security increasingly requires diffusion, dispersal, or even outright disavowal of sensitive information. These questions offer no easy answers, but we have endeavored, in this Article, to at least open the door to asking them.