UNIVERSITY COLLEGE LONDON

DOCTORAL THESIS

Analyzing the Privacy and Societal Challenges Stemming from the Rise of Personal Genomic Testing

Author:
Alexandros Mittos

Supervisor:
Prof. Emiliano De Cristofaro

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in the
Information Security Group
Department of Computer Science

December 15, 2020

Declaration of Authorship

I, Alexandros Mittos, declare that this thesis titled, "Analyzing the Privacy and Societal Challenges Stemming from the Rise of Personal Genomic Testing" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"Somewhere, something incredible is waiting to be known."

Carl Sagan

Abstract

Progress in genomics is enabling researchers to better understand the role of the genome in our health and well-being, stimulating hope for more effective and cost efficient healthcare. At the same time, the rapid cost drop of genome sequencing has enabled the emergence of a booming market for direct-to-consumer (DTC) genetic testing. Nowadays, companies like 23andMe and AncestryDNA provide affordable health, genealogy, and ancestry reports, and have already tested tens of millions of customers. However, while this technology has the potential to transform society by improving people's lives, it also harbors dangers as it prompts important privacy, and societal concerns.

In this thesis, we shed light on these issues using a mixed-methods approach. We start by conducting a technical investigation of the limitations on privacy-enhancing technologies used for testing, storing, and sharing genomic data. We rely on a structured methodology to contextualize and provide a critical analysis of the current state-of-the-art and we identify and discuss ten open problems faced by the community. We then focus on the societal aspects of DTC genetic testing by conducting two large-scale analyses of the genetic testing discourse focusing on both mainstream and fringe social networks, specifically, Twitter, Reddit, and 4chan. Our analyses show that DTC genetic testing is a popular topic of discussion on all platforms. However, these discussions often include highly toxic language expressed through hateful and racist comments and openly antisemitic rhetoric, often conveyed through memes. Overall, our findings highlight that the rise in popularity of this new technology is accompanied by several societal implications that are unlikely to be addressed by only one research field and rather require a multi-disciplinary approach.

Impact Statement

The research in this thesis revolves around the societal impact of personal genomic testing. It investigates whether the privacy of those who contribute their data to genetic testing services and public initiatives can be adequately protected both in the short- and the long-term, how personal genomic testing is reflected online, and whether it is purposely misused. Thus, this thesis impacts various entities, such as researchers, policy makers, companies, and individuals.

First, we demonstrate that current genome privacy tools developed for testing, storing, and sharing genomic data are unable to adequately protect the privacy of the users, both in the short- and the long-term. Our evaluation shows that the overwhelming majority of proposed techniques aiming to scale up to large genomic datasets need to opt for weaker security guarantees or weaker models. Furthermore, one serious challenge stems from lack of long-term security protection, which is difficult to address as available cryptographic tools are not suitable for this goal. We believe that our work will inspire researchers to find alternative and/or creative ways of addressing some of the non-trivial open problems we identify in this thesis.

The results of this thesis have also real-world implications on the DTC genetic testing industry and policy makers. Specifically, we show that DTC genetic testing is being used by fringe groups online to ingrain and empower genetics-based prejudice and discrimination, and even call for genocide. Considering that platforms like Facebook and Twitter have begun to be held accountable when their services enable harmful behavior [229], we believe that our work will motivate policy makers to address the DTC genetic testing industry and legislate so that these companies consider the potential abuse of their services and attempt to find ways of minimizing this behavior. Finally, it is our hope that this thesis will raise awareness about the potential privacy and societal implications of DTC genetic testing to the general public.

Acknowledgements

I would like to express my deep appreciation to all the people who have contributed to this thesis and supported me throughout this process.

First and foremost, I would like to thank my supervisor, Emiliano De Cristofaro. Both his advice and his actions have been a source of inspiration. I would like to also thank him for pushing me to become the best I can be, for teaching me what is important, and for being available when I needed him to. Furthermore, I have been extremely lucky to collaborate with and learn from a few great researchers. Namely: Prof. Bradley Malin, Prof. Jeremy Blackburn, and Dr. Savvas Zannettou.

I would like to thank everyone in the Information Security Research Group at UCL for creating a fun and stimulating environment. I would also like to thank the people involved in the Privacy & Us group, both experienced researchers and Ph.D. students. Your sincere passion for research has always been contagious and every group meeting inspired me to work harder.

This thesis would not have been possible without the unequivocal support of my family who has never stopped believing in me. Last but not least, I would like to thank Zoe for never giving up on me, for reminding me what is important, and for her unconditional love and support despite my “eccentric” personality.

* This research was supported by the European Union’s Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie “Privacy & Us” project (GA No. 675730).

Contents

List of Figures

List of Tables

List of Abbreviations

BAM	**B**inary **A**lignment **M**ap
CNV	**C**opy **N**umber **V**Variations
DTC	**D**irect **T**o **C**onsuming
FGS	**F**ully **S**equenced **G**enomes
GWAS	**G**enome **W**ide **A**ssociation **S**tudy
PET	**P**rivacy **E**enhancing **T**echnology
RFLP	**R**estriction **F**ragment **L**ength **P**olymorphisms
SAM	**S**equence **A**lignment **M**ap
SNP	**S**ingle **N**ucleotide **P**olymorphism
SNV	**S**ingle **N**uncleotide **V**ariant
STR	**S**hort **T**andem **R**epeats
VCF	**V**ariant **C**all **F**ormat

To Zoe, for keeping me sane in a mad world.

To my parents, for always giving their best.

Chapter 1

Introduction

Recent breakthroughs in biology and bioinformatics have enabled researchers to fully sequence the human genome at a cost of about $1,000 [120] and genotype it (i.e., look for specific markers) for much less [107]. These developments have led to a "genomic revolution" [133] that is taking shape in a number of different contexts. For example, they are paving the way to *personalized medicine*, a concept that enables diagnosis and treatment to be tailored to patients' genetic features, aiming to make healthcare more preventive and effective. They also enable public initiatives to sequence large numbers of genomes and build large biorepositories for research purposes, for example, the All Of Us research program in US [164] and the Genomics England project in UK [91] are sequencing the genomes of, respectively, 1 million and 100 thousand volunteers.

At the same time, this rapid cost drop has led to the birth of *direct-to-consumer* (DTC) genetic testing. Rather than visiting a clinic, customers purchase a collection kit for a few hundred dollars or less, deposit a saliva sample, and mail it back. After a few days, they receive a report with information about their genetic health risks (e.g, their likelihood of developing Parkinson's), wellness (e.g., how well they metabolize lactose), carrier status (e.g., their likelihood of hearing loss), and traits (e.g., their eye color). However, perhaps the most popular service is that of genetic *ancestry* testing [188] which promises to discover one's ancestral roots, building on patterns of genetic variations common in people from similar backgrounds [185]. Nevertheless, these are subject to several limitations and their results differ from provider to provider due to different control groups [162]. As of June 2020, AncestryDNA alone, a company that specializes in genetic ancestry tests, has tested more than 16 million customers [7].

However, while this technology has the potential to transform society by improving people's lives, it also harbors dangers as it prompts important privacy, security, and ethical concerns. Specifically, research has shown that genomic data is hard to anonymize [99, 196] while containing sensitive information related to a variety of factors, such as one's ethnic heritage, their disease predispositions, and other phenotypic traits [84]. Furthermore, the consequences of genomic data disclosure are neither limited in time nor to a single individual, i.e., due to its hereditary nature, an adversary obtaining a victim's genomic data can also infer a wide range of features that are relevant to her close relatives and her descendants. Consequently, disclosing the genomic data of a single individual will also put the privacy of others at risk [191].

At the same time, the increased popularity of DTC genetic testing has been accompanied by media reports of far-right groups using it to attack minorities and prove their genetic "purity" [119, 182] mirroring concerns of a new wave of scientific racism [187]. Recently, white nationalists were taped chugging milk at gatherings to demonstrate the ability of white people to better digest lactose [104]. Meanwhile, statements from US President Donald Trump led Senator Elisabeth Warren to publicly confirm her Native American ancestry via genetic testing, prompting heated debates on the matter [144]. This interest in DTC genetic testing by right-wing communities comes at a time when racism, hate, and antisemitism on online social platforms such as 4chan, Gab, and certain communities on Reddit is on the rise [108].

In this thesis, we study several unexplored *technical* and *societal* challenges stemming from the rise of personal genomic testing. On the technical side, we focus on the limitations of proposed privacy enhancing technologies (PETs) geared for testing, storing, and sharing genomic data and identify ten open research questions. To reconcile privacy with progress in genomics, researchers have initiated investigations into solutions for securely testing and studying the human genome. While this effort has produced a relatively large number of publications on the topic over the last few years, it is also partially operating ahead of the curve, proposing the use of PETs in an envisioned, rather than an existing, setting, and often making assumptions for the future; for example, that cheap, error-free whole genome sequencing will soon be available to citizens, or that individuals will be sequenced at birth so that all genetic tests can be easily and cheaply done via computer algorithms. Furthermore, while millions of genomes are already being sequenced and stored in biorepositories both in the public and the private sector [26], at the moment, the privacy of the clients/donors is protected by legal or policy restrictions and access control, rather than by cryptographic tools.

Meanwhile, on the societal side, the fact that individuals can now learn potentially life-changing results with a few clicks of the mouse, without contacting a medical professional leads to social media platforms attracting discussions, sharing of experiences, and molding of perceptions around genetic testing. However, it is unclear what exactly the genetic testing discourse is about. Furthermore, the fact that racist, misogynistic, and dangerous behavior festers and spreads on the Web at an unprecedented scale, prompts the need for a thorough understanding of how genetic testing tools are being used in online discussions.

1.1 Research Questions and Contributions

The widespread use of genetic testing motivates a number of interesting questions that have received little attention from the research community. Specifically:

RQ1. *Can we leverage PETs to adequately protect genome privacy now or in the near future?* Considering that disclosing the genomic data of a single individual can have a significant negative impact both to them and their relatives, it is important to understand whether the privacy of the clients/donors of genetic testing services is likely to be adequately protected in the near future by cryptographic tools or whether this is impossible due to unassailable obstacles tied to the very nature of the problem.

RQ2. *How do people perceive and feel about DTC genetic testing when discussing it online?* As citizens of most developed countries have now easy and affordable access to a wealth of reports related to their health and cultural heritage, it is important to understand how their experiences mold the discourse on genetic testing. Do people discuss about genetic testing online? If yes, in what context and how do they feel about it?

RQ3. *Is DTC genetic testing being used as a means to promote racist views and ideologies online?* As genetic testing is often a controversial topic that is frequently associated with racism and ethical concerns, it is important to investigate whether these controversies hold truth. Do people use DTC genetic testing results to discriminate against groups of people and/or minorities? If yes, how?

In this thesis, we shed light on these research questions. We start by adopting a methodology to analyze themes of genome privacy research, which we then study across several axes using a set of systematic criteria that span a broad spectrum of properties. This allows us to reason about the challenges the genome privacy community faces and whether these are likely to be addressed in the near future.

Then, we conduct an exploratory, large-scale analysis on Twitter to understand the nature and the prevalence of the online genetic testing discourse. Using 10 keywords related to DTC genetic testing companies and 3 to genomics initiatives, we search and crawl all available tweets containing these keywords posted between January 1, 2015 and July 31, 2017. We collect 302K tweets from 113K users and analyze them content-wise, studying the most common hashtags/URLs and measuring sentiment. This enables us to understand how DTC genetic testing is viewed online.

Finally, we move our attention to two platforms that are only loosely moderated and are often associated with high degrees of toxicity: Reddit and 4chan. We expand our set of keywords by an order of magnitude, aiming to capture as many hateful instances as possible. Specifically, we collect 77K comments from Reddit related to genetic testing from January 1, 2016 to March 31, 2018, and 7K threads from the politically incorrect (/pol/) board of 4chan (consisting of 1.3M posts) from June 30, 2016 to March 13, 2018. Our analysis focuses on instances of hateful and toxic speech, using natural language processing and machine learning tools, including (i) Latent Dirichlet Allocation (LDA) to identify topics of discussion, (ii) word embeddings to uncover words used in a similar context across datasets, (iii) Google's Perspective API to measure toxicity in texts, and (iv) Perceptual Hashing to assess the imagery and memes shared in posts. This allows us to study the connection between DTC genetic testing and online hate.

To this end, the general contributions of this work can be summarized as follows:

1. We identify and discuss ten open research questions faced by the genome privacy community and demonstrate that several of them are unlikely to be addressed organically as they are inherently tied to the unique properties of the human genome.

2. We show that genetic testing is a popular topic of discussion online that is discussed in a variety of contexts and is mostly viewed in a positive light. At the same time, we find evidence of people sharing and discussing screenshots of their ancestry test results, despite the possible privacy implications and we show that the conversation around genetic testing is dominated by users that might have a vested interest in its success.

3. We uncover evidence of genetic testing being misused online by groups adjacent to fringe political agendas ingraining and empowering genetics-based prejudice and discrimination, and even calling for genocide.

1.2 Implications

Access to easy, accurate, and affordable genetic testing has the potential to transform society and improve people's lives. At the same time, the rise in popularity of this new technology is accompanied by several societal implications that are unlikely to be addressed by only one research field.

In fact, our findings demonstrate that a *multi-disciplinary* approach may be essential for addressing the issues pertaining DTC genetic testing. Specifically, this thesis sheds light on two aspects of this conundrum: a) a technical investigation of the technological limitations regarding privacy, and b) a societal investigation of how people perceive and use DTC genetic testing services using large-scale quantitative analyses.

The former focuses the privacy implications of accessing, storing, and sharing one's genomic data. In that respect, the genome privacy community has achieved admirable progress in a short amount of time. Nevertheless, our research shows the existence of several open problems, some of which are unlikely to be addressed organically (e.g., by discovering better and faster cryptographic primitives). Thus, it may be essential to the progress of the field to focus on alternative and/or creative solutions to known problems [222].

The latter makes use the relatively new phenomenon of sharing personal experiences on online social platforms. As this has become the norm for a large portion of the total population, it is important to measure and keep track of the effect DTC genetic testing has on society. Considering how technological innovation often has the potential for societal disruption [37, 208], and in light of recent evidence of people (mis)using DTC genetic testing to promote racist ideologies, it is important to keep using quantitative metrics to understand both the benign and harmful sides of this technology.

1.3 Thesis Outline

The organization of this thesis and its detailed contributions are the following:

Chapter 2 introduces preliminary notions, concepts, and tools that are widely used throughout this thesis. Then, Chapter 3 reviews prior work related to the research performed in this thesis.

In Chapter 4, we critically evaluate the research produced by the genome privacy community in the context of PETs geared for testing, storing, and sharing genomic data across several axes, using a set of systematic criteria that span a broad spectrum of properties. We start by adopting a methodology to analyze themes of genome privacy research using a sample of representative papers. Then, we present a systematization which we rely upon to summarize the critical analysis and guide the examination of 10 key aspects of genome privacy. Finally, aiming to validate and broaden the discussion around the identified challenges, we report on an online-administered survey of genome privacy experts, whom we ask to weigh in on them with respect to their importance and difficulty.

Next, in Chapter 5, we present an exploratory, large-scale analysis of Twitter discourse related to genetic testing. We collect 302K tweets from 113K users, and analyze them along several axes, seeking to understand who tweets about genetic testing, what they talk about, and how they use Twitter for that. We start by presenting a general characterization of our dataset. Then, we analyze the tweets content-wise, studying the most common hashtags/URLs and measuring sentiment. Next, we perform a user-based analysis, looking at the profiles and their location, and assessing whether they are likely to be bots. We also select a random sample of 15K users and analyze their latest 1K tweets to study their interests. Finally, we examine the most negative tweets in our dataset, finding a number of tweets related to racism and hate-speech, as well as fears of privacy and data misuse, and look for instances of users sharing screenshots of their test results.

In Chapter 6, we collect 7K threads consisting of 1.3M posts from the politically incorrect (/pol/) board of 4chan and 77K comments from Reddit related to genetic testing between 2016 and 2018. We then analyze them along several axes relying on several natural language processing, computer vision, and machine learning tools to study the connection between DTC genetic testing and online hate.

Finally, Chapter 7 concludes the thesis with a discussion about the implications of its results, its limitations, as well as potential avenues for future research.

1.4 Collaboration

The content presented in this thesis has been co-authored with other researchers and has been published in Computer Science conferences and journals.

The work of Chapter 4 has been conducted in collaboration with Prof. Bradly Malin, and Prof. Emiliano De Cristofaro, and was published at PoPETS 2019 [153]. Specifically, the research required and the writing of the paper was done by the author of this thesis while Prof. Malin and Prof. De Cristofaro had an advisory and editorial role.

The work presented in Chapter 5 has been conducted along with Prof. Jeremy Blackburn and Prof. Emiliano De Cristofaro and was accepted at ACM TWEB 2020. The data analysis and the writing of the paper was done by the author of this thesis while Prof. Blackburn and Prof. De Cristofaro had an advisory and editorial role.

Finally, the work of Chapter 6 has been conducted in collaboration with Dr. Savvas Zannettou, Prof. Jeremy Blackburn, and Prof. Emiliano De Cristofaro and was published at ICWSM 2020 [154]. The data analysis and the writing of the paper was done by the author of this thesis, except from Figures 6.3 and 6.5 which were produced by Dr. Zannettou. Subsections 6.4.2 and 6.5 were co-written by the author of this thesis and Dr. Zannetou. Specifically, Dr. Zannettou wrote the parts of these subsections that explain how Figures 6.3 and 6.5 were produced. Prof. Blackburn and Prof. De Cristofaro had an advisory and editorial role.

Chapter 2

Background

In this chapter, we introduce preliminary notions, concepts, and tools that are widely used throughout this thesis. We start with a primer on genetic testing focusing on terminologies, contexts, attacks, and legal aspects. Then, we provide descriptions of the social networks we focus on in this thesis and we also describe the tools we use to analyze them.

2.1 Genomics Primer

We now describe important terminologies and contexts discussed in this thesis, as well as common attacks against genome privacy. We also discuss how genome privacy is being protected legally in the EU and USA.

2.1.1 Terminologies

Whole Genome Sequencing. Whole Genome Sequencing (WGS) is the process of determining the complete DNA sequence of an organism. In other words, it is used to digitize the genome of an individual into a series of letters corresponding to the various nucleotides (A, C, G, T). WGS costs are now on the order of $1,000 [157].

Whole Exome Sequencing. Whole Exome Sequencing is the process of sequencing parts of DNA which are responsible for providing the instructions for making proteins.

Genotyping. Sequencing is not the only way to analyze the genome; in fact, in-vitro techniques such as genotyping are routinely used to look for known genetic differences using biomarkers.

SNPs, CNVs, STRs, RFLPs, Haplotypes, and SNVs. All members of the human population share around 99.5% of the genome, with the remaining 0.5% differing due to genetic variations. However, this 0.5% variation does not always manifest in the same regions of the genome. As an artifact, storing how much or where differences occur could lead to uniqueness and identifiability concerns.

A Single Nucleotide Polymorphism (SNP) is a variation at a single position, occurring in 1% or more of a population. SNPs constitute the most commonly studied genetic feature today, as researchers use them to identify clinical conditions, predict the individuals' response to certain drugs, and their susceptibility to various diseases [228].

However, Copy Number Variations (CNVs) [209] and Short Tandem Repeats (STRs) [40] are also becoming increasingly more used. Restriction Fragment Length Polymorphisms (RFLPs) refer to the difference between samples of homologous DNA molecules from differing locations of restriction enzyme sites, and are used to separate DNA into pieces and obtain information about the length of the subsequences.

A haplotype refers to a group of genes of an individual that was inherited from a single parent. Finally, while a SNP refers to variation which is present to at least 1% of a population, a Single Nuncleotide Variant (SNV), is a variation occurring in an individual, without limitations on frequency.

GWAS. Genome Wide Association Study is the process which compares the DNA of study participants to discover SNPs that occur more frequently in people carrying a particular disease.

SAM, BAM, FASTQ, and VCF. Fully Sequenced Genomes (FGSs) are typically stored in either SAM, BAM, or FASTQ formats. SAM (Sequence Alignment Map) is a text-based format, which may include additional information such as the reference sequence of the alignment or the mapping quality, BAM is a binary format (in practice, the compressed and lossless version of SAM), while FASTQ is a text-based format which stores nucleotide sequences along with their corresponding quality scores. VCF (Variant Call Format) is another text file format broadly used in bioinformatics for storing gene sequence variations.

Genome Operations. A variety of operations can be performed on genomes. For the purpose of this thesis, we describe three of the most common:

1. *SNP weighted average.* There are several methods to compute the disease susceptibility of an individual, or other genetic testing related analysis that involve SNPs. A common method is based on weighted averaging, where certain weights are applied on SNPs to calculate the result of a test.
2. *Edit distance.* An edit distance algorithm measures the dissimilarity of two strings. Specifically, the edit distance between two strings corresponds to the minimum number of edit operations (i.e., insertion, deletion, substitution) needed to transform one string into the other. In the context of genomics, usually researchers measure the edit distance between a patient and a reference genome.
3. *χ^2 test.* A χ^2 test is a statistical method to test hypotheses. Specifically, a χ^2 test is used to determine whether there is a significant difference between the observed frequencies in a sample against the expected ones if the null hypothesis is true. In the context of genomics, a χ^2 test helps researchers determine whether certain hypotheses are true, e.g., determine whether two or more alleles are associated (linkage disequilibrium).

2.1.2 Contexts

Progress in genomics has led to a "genomic revolution" that is taking shape in a number of different contexts. In this thesis, we focus on two of them, namely, public sequencing initiatives and the private market.

Public Sequencing Initiatives. The promise of improved healthcare has encouraged ambitious sequencing initiatives, aiming to build biorepositories for research purposes. In 2015, the US government announced the Precision Medicine Initiative (now known as the All Of Us Research Program [164]), aiming to collect health and genetic data from one million citizens. Similarly, Genomics England is sequencing the genomes of one hundred thousand patients, focusing on rare diseases and cancer [91].

The rise of data-driven genomics research also prompts the need to facilitate data sharing. In 2013, the Global Alliance for Genomics and Health (GA4GH) was established with an objective to make data sharing between institutes simple and effective [94]. The GA4GH has developed various software, such as the Beacon Network [72], which permits users to search if a certain allele exists in a database hosted at a certain organization, and the Matchmaker Exchange [176], which facilitates rare disease discovery.

Private Market. Progress in genomics has also encouraged the rise of a flourishing private sector market. Several companies operate successfully in the business of sequencing machines (e.g., Illumina), genomic data storage and processing (e.g., Google Genomics), or AI-powered diagnostics (e.g., Sophia Genetics). At the same time, others offer genetic testing *directly* to their customers, without involving doctors or genetics experts in the process. There are now hundreds of DTC genetic testing companies [177], which collectively have amassed tens of millions of customers [1, 7].

2.1.3 Attacks Against Genome Privacy

Sharing genetic findings is crucial for helping biomedical discoveries advance. However, research has shown that genetic data is hard to anonymize [99, 196] prompting important privacy concerns. In that respect, several types of attacks against genome privacy have been demonstrated by the research community.

A few *re-identification* attacks have been proposed whereby an adversary recovers the identity of a target by relying on quasi-identifiers, such as demographic information (e.g., linking to public records such as voter registries), data communicated via social media, and/or search engine records [210]. For instance, Gymrek et al. [99] infer the surnames of individuals from (public) anonymized genomic datasets by profiling short tandem repeats on the Y chromosome while querying genealogy databases.

Also, in *membership inference* attacks, an adversary infers whether a targeted individual is part of a study that is possibly associated with a disease, even from aggregate statistics. Homer et al. [109] do so by comparing the target's profile against the aggregates of a study and those of a reference population obtained from public sources. Wang et al. [223] leverage correlation statistics of a few hundreds SNPs, while Im et al. [118] use regression coefficients.

Shringarpure and Bustamante [196] present inference attacks against Beacon by repeatedly submitting queries for variants present in the genome of the target, whereas, Backes et al. [16] attacks focused on microRNA expressions. More generally, Dwork et al. [75] prove that membership attacks can be successful even if aggregate statistics are released with significant noise. For a comprehensive review of the possible/plausible attacks against genome privacy, we refer the readers to [76].

2.1.4 Legal Aspects

The General Data Protection Regulation (GDPR) [173] came into effect in the EU in May 2018. However, its impacts on genetic testing and genomic research are not yet clear. The *data minimization* principle suggests that the minimum required amount of data should be stored to achieve the intended goal, while the *purpose limitation* principle dictates that researchers should predetermine the scope of the study (Article 5). Under Article 35, genetic data is designated both as sensitive and personal data.

In the US, there is no equivalent of GDPR, however, certain legislation and policy protects the privacy of study participants using indirect means. For example, to access sensitive data in NIH databases, a researcher must first submit a request. More specifically, under the Health Insurance Portability and Accountability Act (HIPAA)[1] and its Privacy Rule[2] genetic information is considered to be "health information" even if it is not clinically significant, and under the Genetic Information and Nondiscrimination Act of 2008 (GINA) it is illegal for employers or health insurers to request genetic information of individuals or of their family members. However, this legislation does not cover the cases of life, disability, and long-term care insurance.

Overall, the HIPAA Privacy Rule was not designed to be a comprehensive health privacy protection system. In the US healthcare system, patients are required to sign an acknowledgment of privacy practices when they seek care which may lead them to the assumption that their health privacy is adequately protected, however, the HIPAA Privacy Rule has numerous exceptions permitting access to individually identifiable health information, such as when it is required by law, to avert a serious threat to health or safety, for specialized government functions including national security, and others [55]. Finally, there are cases of individuals voluntarily making their genomic data public without considering the privacy impact on themselves and their relatives.

2.2 Social Platform Analysis

In this section, we describe the three social networks this thesis focuses on, as well as the tools we use to analyze them.

[1] 42U.S.C.300gg-300gg-2

[2] 45C.F.R.pts.160,162,164

2.2.1 Social Platforms

Our work in Chapters 5 and 6 uses datasets from three social platforms, namely, Twitter, Reddit, and 4chan.

Twitter. Twitter is a social networking and micro-blogging service where users can post short messages known as "tweets." These tweets can, in return, be commented on, retweeted, and liked by other users.

Reddit. Reddit is a social news aggregation and discussion website, where users post content (e.g., images, text, links) that gets voted up or down by other users. Users can add comments to the posts, and comments can also be voted up or down and receive replies. Top submissions appear on the front page, and top comments appear at the top of each post. Content on Reddit is organized in communities created by users called *subreddits* which are usually associated with areas of interest (e.g., movies, sports, politics).

4chan. 4chan is an imageboard website with virtually no moderation. An "Original Poster" (OP) creates a thread by posting an image that may include a message. Content is organized in subcommunities, called boards (as of June 2020, there are 80 of them), with various topics of interest (e.g., video games, literature, etc.). Other users can post in the OP's thread with a message and/or an image. On 4chan, users do not need a registered account to post content.

As of 2020 there are at least 95 Internet social networks [2]. Studying how DTC genetic testing is perceived on all of them would an infeasible task and arguably out of scope. Nevertheless, we choose these three networks for several reasons. First, Twitter and Reddit are extremely popular; as of June 2020, Twitter has more than 330 million active users [239] and Reddit has more than 430M monthly active users and 21B visits, which makes it the fifth most visited site in the US [181]. Second, 4chan has been shown to include a high volume of racist, xenophobic, and hateful content [108] which allows us to study how DTC genetic testing is used in this context.

2.2.2 Tools

To analyze and interpret our datasets, we use the set of tools described below.

Latent Dirichlet Allocation. LDA is a generative statistical model that allows the identifying of topics that best describes a set of documents [29]. In the context of this thesis, it allows us to summarize our collected documents (i.e., posted texts) into a small number of topics consisting of keywords that are representative of the overall themes pertaining them.

Before applying the LDA algorithm we cleaned the data (i.e., removed punctuation and special characters, lower-cased). We did so by using the Python library NLTK.[3] For the LDA process we used the Python library Scikit Learn, specifically its Latent Dirichlet Allocation method.[4]

When using an LDA algorithm there are two important parameters to take into consideration: The *alpha* parameter is a concentration parameter that represents document-topic density. Specifically, higher alpha means that documents are assumed to be made up of more topics which results in more specific topic distribution per document. The *beta* parameter is a concentration parameter that represents topic-word density. Specifically, higher beta means that topics are assumed to made of up most of the words and result in a more specific word distribution per topic. For our implementation, and after experimenting with the *alpha* and *beta* values, we chose to keep Scikit Learn's default values which, for both parameters, equals to $1/number\_of\_topics$. We also used the *batch* learning method. We kept the rest of the parameters on their default options.

While LDA can be a valuable tool for summarizing topics of interest it also suffers from several limitations. Most notably, the number of the extracted topic (usually denoted as *K*) is fixed (i.e., it must be known ahead of time). In this thesis, our *K* varies from 3 to 10 topics depending on the number of documents processed in each analysis.

[3] https://www.nltk.org/

[4] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

Furthermore, LDA cannot capture correlations and assumes that words are exchangeable.

Google's Perspective API. Perspective API is a publicly available tool geared to identify toxic comments [175]. The API returns a value between 0 and 1, pertaining to the likelihood of a text being toxic, i.e., how rude, disrespectful, or unreasonable a comment is likely to be; severely toxic, which is similar to toxicity but only focuses on the "most toxic" comments; and inflammatory, which focuses on texts intending to provoke or inflame.

Perceptual Hashing. Perceptual hash functions extract features from multimedia content and calculate hash values based on them. They can be used to compare two objects by calculating a distance/similarity score between two hash values; the objects are labeled as (perceptually) equal if the distance is below a chosen threshold [166].

Word Embeddings. Word embeddings is the collective name for a set of language modeling and feature learning techniques in the natural language processing (NLP) field where words or phrases from a vocabulary are mapped to vectors of real numbers. In this thesis, we use a specific type of word embeddings called word2vec [151]. Word2vec models are trained on large corpora of text and generate a high-dimensional vector for each word that appears in the corpus; words that are used in similar context also have a closer mapping to the high-dimensional vector space. This allows us to study which words are used in similar contexts.

Chapter 3

Related Work

The work in this thesis revolves around the privacy and societal challenges stemming from the rise of personal genomic testing. In this chapter, we review prior work related to our line of research. Specifically, we start by presenting the related literature on analyzing security and privacy challenges in the context of genomic data. Then, we present prior work on the societal effects of DTC genetic testing, followed by studies of health-related topics on social networks. We also review prior exploratory studies on Twitter, Reddit, and 4chan. Finally, we present prior work on hate speech on social platforms.

3.1 Security & Privacy Challenges of Genetic Testing

In this section, we review prior work analyzing security and privacy challenges in the context of genomic data.

Erlich and Narayanan [76] analyze the different routes that can be used to breach and defend genome privacy. They group plausible and possible attacks into three categories: completion, identity tracing, and attribute disclosure attacks, and discuss the extent to which mitigation techniques (i.e., access control, anonymization, and cryptography) can address them. Ayday et al. [12] summarize the importance of progress in genomics along with the need for preserving the privacy of the users when dealing with their genomic data.

Shi et al. [194] review genomic data sharing, the potential privacy risks, as well as regulatory and ethical challenges. They also categorize tools for protecting privacy into controlled access, data perturbation (specifically in the context of differential privacy), and cryptographic solutions, providing an overview for each category. Also, Wang et al. [224] study the clinical, technical, and ethical sides of genome privacy in the United States, describing available solutions for the disclosure of results from genomic studies along record linkage, and the ethical and legal implications of genome privacy in conjunction to informed consent.

Naveed et al. [161] present an overview of genome privacy work from a computer science perspective, reviewing known privacy threats and available solutions, and discuss some of the known challenges in this area—e.g., that genomic databases are often not under the control of the health system, or that privacy protection might affect utility. They also interview 61 experts in the biomedical field, finding that the majority of them acknowledge the importance of genome privacy research and risks from privacy intrusions.

Aziz et al. [15] categorize genome privacy research in three groups: privacy-preserving data sharing, secure computation and data storage, as well as query or output privacy; they discuss solutions and open problems in those fields, aiming to help practitioners better understand use cases and limitations of available crypto primitives. and compare their performance and security. Finally, Akgün et al. [6] focus on bioinformatics research where private information disclosure is possible. For instance, they consider scenarios about querying genomic data and sequence alignment, while surveying available solutions for each.

3.2 Genetic Testing & Society

We now present studies that have worked toward understanding the societal effects of genetic testing.

Roth and Ivemark [192] interview users to study how ancestry testing affects ethnic and racial identities by conducting 100 interviews with people who have white, black, Hispanic/Latino, Native American, and Asian ancestry. They find instances of consumers not accepting test results, and instead focus on estimates based on social appraisals and aspirations. Overall, they suggest that genetic ancestry testing may reinforce race privilege.

Clayton et al. [56] conduct a meta-analysis of 53 studies involving 47K people around perceptions of genetic privacy, highlighting how survey questions are often phrased poorly, thus leading to possible misinterpretations of the results. They also show that not enough attention was paid to influential factors, e.g., participants' sociocultural backgrounds.

Also, Couldry and Yu [58] discuss how DTC genetic companies, such as 23andMe, influence the public toward sharing their genetic data by claiming that the abundance of data will improve people's lives in the long term, despite a body of work showing that genetic data cannot be securely anonymized [99, 196].

Panofsky and Donovan [172] analyze 70 discussion threads on the far-right website `Stormfront.org`, where at least one user posted ancestry test results. They group posters based on whether they consider their results good or bad and study how other Stormfront users react: if the posters receive "bad news," they tend to question the validity of genetic genealogy science, trying to reinterpret their results to fit their views on races.

In follow-up work [171], they also look at the relationship between citizen science and white nationalists' use of genetic testing, shedding light on how "repair strategies" combine anti-scientific attacks on the legitimacy of these tests and reinterpretations of them in terms of white nationalist histories. Finally, Chow-White et al. [52] examine 2K tweets containing the keyword '23andMe' spanning one week. They calculate their sentiment and find out that the positive tweets outnumber the negative, while users appear overall enthusiastic about the company's services.

3.3 Health in Social Networks

Twitter has been extensively used to study health and health-related issues, e.g., to measure and predict depression.

De Choudhury et al. [63] identify 476 users self-reporting depression, collect their tweets, and study their engagement, emotion, and use of depressive language. By comparing to a control group, they extract significant differences, and build a classifier to predict the likelihood of an individual's depression. Coppersmith et al. [57] study tweets related to various mental disorders, while Paul et al. [174] gather public health information from Twitter, discovering statistically significant correlations between Twitter and official health statistics. Abbar et al. [3] analyze the nutritional behavior of US citizens: they collect 892K tweets by 400K US users using food-related keywords and find that foods match obesity and diabetes statistics, and that Twitter friends tend to share the same preferences in food consumption.

Prasetyo et al. [179] study how social media can effect awareness in health campaigns. Focusing on the Movember charity campaign, they collect more than 1M tweets, using the keyword 'Movember', and uncover correlations between the visitors of the Movember website and popular Twitter users, but none between tweets and donations. Finally, Cavazos-Rehg et al. [43] study drinking behaviors on Twitter: using keywords related to drinking (e.g., drunk, alcohol, wasted), they collect 10M tweets and identify the most common themes related to pro-drinking and anti-drinking behavior.

3.4 Quantitative Studies on Twitter, Reddit, and 4chan

Researchers have extensively studied social networks like Twitter, Reddit, and 4chan, producing a number of exploratory studies of discourse and overall user behavior.

Twitter. Lerman et al. [142] conduct an emotion analysis on tweets from Los Angeles: using public demographic data, they find that users with lower income and education levels, and who engage with less diverse social contacts, express more negative emotions, while people with higher income and education levels post more positive messages.

Chatzakou et al. [47] study the GamerGate controversy[1] on Twitter, collecting a dataset of tweets containing keywords indicating abusive behavior. They compare the characteristics of the related Twitter profiles to a baseline, finding that users tweeting about GamerGate are more technologically savvy and active, and that their tweets are more negative. Burnap et al. [38] study Twitter responses to a terrorist attack occurred in Woolwich in 2013. Using 'Woolwich' as a keyword search, they collect 427K tweets, finding that opinions and emotional factors are predictive of size and survival of information flows.

Reddit. De Choudhury and De [62] look at Reddit conversations about mental health, aiming to understand language attributes of online self-disclosure and factors driving support in online posts. They show that users explicitly share personal information on their mental health, and use Reddit for self-expression, even for seeking diagnosis or treatment information.

Other studies analyze how users behave in specific subreddits. Kasunic et al. [132] focus on a specific subreddit called /r/RoastMe, where users post photos of themselves and invite others to ridicule them. They find that the RoastMe community relies on a specific set of norms, such as highly valuing caustic comments but also being concerned about the potential psychological harm of the participants. Nobles et al. [167] study

[1] `https://en.wikipedia.org/wiki/Gamergate_controversy`

/r/STD to understand how users seek health information on sensitive and stigmatized topics. They find that most posts crowd-source information about non-reportable STDs, focusing on treatment, symptoms, as well as aspects of social and emotional impact.

Flores-Saviaga et al. [82] analyze 16M comments spanning two years to examine the characteristics of political troll communities. They find that /r/The_Donald subscribers spend energy educating their community on certain events and that they use various socio-technical tools to mobilize other subscribers. Finally, Mills [152] compares /r/The_Donald to /r/SandersForPresident, a subreddit broadly supporting the 2016 presidential candidate Bernie Sanders, exploring whether rapidly formed subreddits exhibit collective intelligence. Mills finds that these communities are very effective on pursuing their agendas and that Trump supporters more often tend to clash with other communities and Reddit administrators.

4chan. Bernstein et al. [24] study 5M posts on 4chan's random board (/b/) to examine anonymity and ephemerality on 4chan. They find that most threads expire in less than 5 minutes, while over 90% of the posts are anonymous.

Hine et al. [108] study 8M posts from /pol/ collected over two and a half months. Their content analysis reveals that while most URLs point to YouTube, a non-negligible amount link to right-wing websites. They also find evidence of organized "raids" against YouTube users, where users collectively post hateful comments on videos they disapprove of.

3.5 Online Hate

Researchers have also studied hate speech on mainstream social networks like Twitter [197, 61, 168, 190], Reddit [44, 168], Facebook [23, 69], YouTube [169, 189], and Instagram [112]. Chatzakou et al. [46] explore cyberbullying on Twitter. They rely on a dataset of 1.6M tweets to study the properties that characterize bullying, and build a machine learning classifier which identifies users exhibiting aggressive behavior.

Founta et al. [83] focus on abuse detection. They leverage a deep learning architecture which takes into consideration various features (e.g., metadata of posts, prior posts, account settings, social network, popularity) and propose a tool that is able to capture several facets of abusive behavior, i.e., cyberbullying, hateful and offensive content, and sarcasm.

Zannettou et al. [236] explore how mainstream and fringe online communities on Twitter, Reddit, and 4chan influence each other with respect to disinformation and hateful propaganda. Then, Zannettou et al. [235] detect and study racist and hateful memes, and their propagation, on 4chan, Gab, Reddit, and Twitter. Among other things, they find that racist memes are very common on /pol/ and Gab, and that /pol/ and the /r/The_Donald subreddit are the most influential Web communities with respect to the dissemination of memes.

Then, in [81], Finkelstein et al. study anti-semitism on /pol/ and Gab—a Twitter clone known for attracting users who are banned from Twitter—revealing that anti-semitic content increases in those networks after major political events, such as the "Unite the Right" rally or the 2016 US elections. Also, they leverage word embeddings to identify terminology associated with anti-semitic content.

Chandrasekharan et al. [45] study how Reddit's decision to ban subreddits that violated anti-harassment policy affected hate speech on the platform. They examine 100M posts and comments from two banned subreddits, namely r/fatpeoplehate and r/CoonTown, and measure the generated hate speech by its users before and after the ban. They find that the ban had a positive effect on the platform as the users who continued posting drastically reduced their hate speech usage.

3.6 Discussion

Our literature review yields several findings. Regarding the field of genome privacy, the current collection of surveys review and summarize available solutions. In combination, they complement our work (Chapter 4) and are useful in providing an overview of the state of the art, as well as the different avenues for privacy protection. Some also highlight challenges from non-technical perspectives, including ethics, law, and regulation. However, they are mostly limited to reviews of already completed work by grouping and assessing available solutions without a structured methodology, and therefore, their analysis is usually restricted to the papers they survey and do not apply to the community in general.

Whereas, in Chapter 4, we define systematic criteria and sample relevant papers that are representative of a line of work; this helps us provide a critical analysis of challenges and outlooks. Specifically, we discuss several challenges which have not been identified in the past by previous surveys, including: the lack of long-term confidentiality protection, the over-reliance on non-collusion assumptions, the challenges presented by making strong data representations assumptions, as well as the lack of solutions applicable to current genomics initiatives.

Regarding studying how DTC genetic testing is perceived we find that most of the research in this area relies on qualitative studies examining the societal effects of genetic testing [42, 59, 102, 53, 163], and somewhat lacks quantitative large-scale analysis. To the best of our knowledge, our work in Chapters 5 and 6 is the largest, quantitative measurement study on the trends, themes, and topics of discussion around genetic testing.

Chapter 4

Reconciling Privacy with Progress in Genomics: A PETs Perspective

4.1 Introduction

Facilitated by rapidly dropping costs, genomics researchers have made tremendous progress over the past few years toward mapping and studying the human genome. Today, the long-anticipated "genomic revolution" [133] is taking shape in a number of different contexts, ranging from clinical and research settings to public and private initiatives.

At the same time, this progress also prompts important privacy, security, and ethical concerns. Genomic data is hard to anonymize [99, 196] and contains information related to a variety of factors, including ethnic heritage, disease predispositions, and other phenotypic traits [84]. Moreover, consequences of genomic data disclosure are neither limited in time nor to a single individual; due to its hereditary nature, an adversary obtaining a victim's genomic data can also infer a wide range of features that are relevant to her close relatives as well as her descendants. As an artifact, disclosing the genomic data of a single individual will also put the privacy of others at risk [191].

Motivated by the need to reconcile privacy with progress in genomics, researchers have initiated investigations into solutions for securely testing and studying the human genome. Over the past few years, the genome privacy community has produced a relatively large number of publications on the topic, with several dedicated events including international seminars [101, 115, 92], a workshop [121], and a competition series [117].

At the same time, the community is partially operating ahead of the curve, proposing the use of privacy-enhancing technologies (PETs) in envisioned, rather than existing, settings. In fact, as discussed later on, genome privacy research also makes assumptions for the future, e.g., that cheap, error-free whole genome sequencing will soon be available to private citizens, or that individuals will be sequenced at birth so that all genetic tests can be easily and cheaply done via computer algorithms.

Based on these developments, it is time to take stock of the state of the field. To do so, we conduct a systematic analysis of genome privacy research, aiming to evaluate not only what it has achieved so far, but also future directions and the inherent challenges the field faces. Overall, our work is driven by three main research objectives:

1. Critically review, evaluate, and contextualize genome privacy research using a structured methodology that can be reused in the future to assess progress in the field.

2. Reason about the relevance of the proposed solutions to current public sequencing initiatives as well as the private market.

3. Identify limitations, technical challenges, and open problems faced by the community. In particular, we aim to assess which of these are likely to be addressed via natural progress and research follow-ups and which are inherently tied to the very nature of the problem, involving challenging trade-offs and roadblocks.

Roadmap. With these objectives in mind, we set out to critically evaluate work produced by the genome privacy community across several axes, using a set of systematic criteria that span a broad spectrum of properties. Rather than presenting an exhaustive review of the very large body of work in this field, we adopt a methodology to analyze themes of genome privacy research using a sample of representative papers.

We focus on research relying on PETs in the context of testing, storing, and sharing genomic data. To do so, we retrieve the list of publications in the field from a community managed website (GenomePrivacy.org) while intentionally excluding papers about attacks and risk quantification (Section 4.2). After identifying relevant sub-areas of genome privacy research, we select results that provide a meaningful sample of the community's work in each area (Section 4.3).

Next, we present a systematization which we rely upon to summarize the critical analysis and guide the examination of 10 key aspects of genome privacy (Section 4.4). Finally, in Section 4.5, aiming to validate and broaden the discussion around the identified challenges, we report on an online-administered survey of genome privacy experts, whom we ask to weigh in on them with respect to their importance and difficulty. Overall, our analysis, along with the input from the experts, motivates the need for future work as well as more interdisciplinary collaborations, while pointing to specific technical challenges and open problems; moreover, our methodology can be reused to revisit new results and assess the progress in the field.

Main Findings. Our analysis also helps us draw some important conclusions. First, that the effective use of PETs in the context of genome privacy is often hindered by the obstacles related to the unique properties of the human genome. For instance, the sensitivity of genomic data does not degrade over time, thus prompting serious challenges related to the lack of effective long-term security protection, as available cryptographic tools are not suitable for this goal. Second, we find that the majority of the proposed solutions, aiming to scale up to large genomic datasets, need to opt for weaker security guarantees or weaker models. While it is not unreasonable to expect progress from the community with respect to underlying primitives, it is inherently hard to address the limitations in terms of utility and/or flexibility on the actual functionalities. When combined with assumptions made about the format and the representation of the data, this poses major hurdles against real-life adoption.

On the positive side, we highlight how, in its short lifespan, the genome privacy community has achieved admirable progress. For instance, several tools can already enable genomics-related applications that are hard or impossible to support because of legal or policy restrictions.

4.2 Methodology

In this section, we introduce our systematization methodology. We provide intuition into how we select a representative sample of the community's work; next, we describe the criteria used to systematize knowledge.

4.2.1 Sampling Relevant Work

GenomePrivacy.org. We study research on genome privacy from the point of view of privacy-enhancing technologies (PETs) – specifically, we focus on work using PETs in the context of testing, storing, and/or sharing genomic data. Therefore, we rely on the website `GenomePrivacy.org`, which bills itself as "the community website for sharing information about research on the technical protection of genome privacy and security." In the summer of 2017, we retrieved the 197 articles listed on the site, and grouped them into six canonical categories – Personal Genomic Testing, Genetic Relatedness Testing, Access and Storage Control, Genomic Data Sharing, Outsourcing, and Statistical Research – from which we selected a sample representing the state of the art for each category.

Excluding Attack/Quantification Papers. We excluded work on attacks (reviewed in Chapter 2) and privacy quantifications as our main focus is on the use of PETs. We refer readers to [220] for a comprehensive evaluation of metrics geared to quantify genomic privacy, and to [222] for game-theoretic approaches to quantify protection in genomic data sharing. We also did not include recent proposals to address specific attacks in the context of Beacon [196], e.g., the work by Raisaro et al. [180] or Wan et al. [221], although we note their importance later in Section 4.4.4.

Selection. To select the list of papers used to drive our systematic analysis (Section 4.4), we followed an iterative process. First, the team selected 45 articles considered to represent the state of the art in the six categories, and later added four more during a revision of the paper in March 2018. Since it would be impossible to review and systematize all of them in a succinct and seamless manner, we trimmed down the selection to 25 papers (reviewed in Section 4.3). When deciding whether to include one paper over another, the team preferred papers published in venues that are more visible to the privacy-enhancing technologies community or that have been cited significantly more, as they arguably have a stronger influence on the community over time.

Ultimately, the selection covers four papers in Personal Genomic Testing, three in Genetic Relatedness Testing, four in Access & Storage Control, six in Genomic Data Sharing, six in Outsourcing, and four in Statistical Research. Note that two articles appear in both Personal Genomic Testing and Genetic Relatedness Testing categories. For completeness, in Section 4.3, we also add a citation to the papers from the first list of 49 papers that are not included in the final selection.

Remarks. We stress that we do not aim to analyze all papers related to genomic privacy in the context of PETs; in fact, our selection is meant to be representative of the state of the art for each category, but not of its breadth or depth. Rather, we systematize knowledge around genomic privacy protection mechanisms and critically evaluate it. As a consequence, if we added or replaced one paper with another, the main takeaways would not be considerably altered.

4.2.2 Systematization Criteria

We now discuss the main axes along which we systematize genome privacy work. To do so, we elicit a set of criteria designed to address the research questions posed in the introduction. Inspired by similar approaches in SoK papers [31, 134], we choose criteria aiming to capture different aspects related to security, efficiency, and system/data models while being as comprehensive as possible. These criteria constitute the columns of Table 4.1 (see Section 4.4), where each row is one of the papers discussed below. Specifically, we define the following 9 criteria:

1. Data Type. We capture the type of genomic data used, e.g., some protocols perform computation on full genomes, or other aspects of the genome such as SNPs or haplotypes.

2. Genomic Assumptions. We elicit whether techniques make any assumptions as to the *nature* of the data. For instance, the processing of sequencing data is not perfect and nucleotides (or even sequences thereof) might be misreported or deleted, while others might be inserted unexpectedly. In fact, the error rate percentage across various next-generations sequencers can be as high as 15% [96]. As such, the output of modern Illumina sequencing machines (i.e., FASTQ format[1]) is made of segments of DNA with probabilities associated with the confidence that letters were read correctly. This criterion serves to note which of the proposed methodologies take into consideration, or are particularly affected by this.

3. Storage Location. We study where genomic data is assumed to be stored. We identify three options: (i) a personal device, like a mobile phone or a dedicated piece of hardware which is operated by the user, (ii) the cloud, from which a user can directly obtain her data or allow a medical facility to obtain it, and (iii) institutions (e.g., biobanks and hospitals), which store and are able to process genomic data at will. We refer to the latter as *Data Controllers*, following GDPR's terminology [173].

[1] https://help.basespace.illumina.com/articles/descriptive/fastq-files/

4. Use of Third Parties. We determine the presence of third parties, if any, as well as their nature. For instance, some protocols may involve key distribution centers and semi-trusted cloud storage providers.

5. Long-Term Security. Due to its hereditary nature, the sensitivity of genomic data does not degrade quickly over the years: even access to the genome of a long-deceased individual might still pose a threat to their descendants. Therefore, we look at the underlying building blocks and the computational assumptions in genome privacy tools and analyze whether or not they can realistically withstand several decades of computational advances.

6. Security Assumptions. We study the assumptions made on entities involved, if any. For instance, we consider if third parties are assumed not to collude with any other entities.

7. Methods. We report on the main security tools and methods used (e.g., secure multiparty computation, homomorphic encryption).

8. Privacy Overhead. We broadly quantify the overhead introduced by the privacy defense mechanisms, compared, whenever possible, to non privacy-preserving versions of the same functionality. This is a non-trivial task because each sub-area of genome privacy has different goals and each piece of work in that area does not necessarily solve the exact same problem. Nonetheless, we analyze the complexity of each solution to assess their efficiency in terms of time and storage overhead. We report on the lower and upper values of complexity to emphasize how each solution fares against the non-privacy version of the same functionality. We do so based on the premise that if the technique imposes orders of magnitude higher overhead than the non-privacy-preserving version, then the overhead is considered to be high, and low otherwise.

9. Utility Loss. Finally, we measure the impact of privacy tools on the utility of the system. Such measurements include the overall flexibility of the proposed work in comparison with the intended task. Similar to the privacy overhead criterion, we compare against non-privacy-preserving versions of the same functionality, and quantify utility loss as either low or high.

Remarks. We do not necessarily report on the specific metrics used in the selected papers (e.g., running times) as (i) not all papers provide metrics, and (ii) similar approaches already appear in prior work (see Section 3.1). Rather, the metrics used in the systematization are designed to support a critical analysis of the PETs invoked to protect genome privacy.

4.3 Representative Papers

We now review the papers selected according to the methodology presented in Section 4.2. These papers constitute the rows in Table 4.1.

4.3.1 Personal Genomic Testing

We begin with papers that define privacy-preserving versions of personal genomic tests. These have a variety of uses, including assessments of a person's predisposition to a disease, determining the best course of treatment, and optimizing drug dosage. Typically, they involve an individual and a testing facility, and consist of searching for and weighting either short patterns or single nucleotide polymorphisms (SNPs). In this context, there are two main privacy-friendly models: (1) one assuming that individuals keep a copy of their genomic data and consent to tests so that only the outcome is disclosed and (2) another involving a semi-trusted party that stores an encrypted copy of the patient's genetic information, and is involved in the interactions.

Baldi et al. [18] operate in model (1), supporting privacy-preserving searching of mutations in specific genes. They use authorized private set intersection (APSI) [68], which guarantees that the test is authorized by a regulator ("authorization authority") and pushes pre-computation offline so that the complexity of the online interaction only depends on the handful of SNPs tested. It also ensures that the variants which make up the test are kept confidential, as this may pertain to a company's intellectual property.

Ayday et al. [11] introduce model (2), letting a Medical Center (MC) perform private disease susceptibility tests on patients' SNPs, by computing a weighted average of risk factors and SNP expressions. In this model patients have their genome sequenced once, through a Certified Institution (CI) that encrypts the SNPs and their positions, and uploads them to a semi-trusted Storage and Processing Unit (SPU). The MC computes the disease susceptibility using cryptographic tools, such as homomorphic encryption and proxy re-encryption. Also in model (2) is the work by Naveed et al. [160], whereby the CI encrypts genomes using controlled-functional encryption (C-FE), under a public key issued by a central authority, and publishes the ciphertexts. MCs can then run tests using a one-time function key, obtained by the authority, which corresponds to one specific test and can only be used for that test.

Djatmiko et al. [70] operate in both models (i.e., patients control their data by storing it on a personal device or in the cloud) to support personalized drug dosing (which in this case happens to be Warfarin, a blood thinner). The testing facility retrieves data to be evaluated (using private information retrieval [51]) and processes it while encrypted. The patient then securely computes the linear combination of test weights (using additively homomorphic encryption), and shows the results to the physician.

Personal Genomic Testing – Selected Papers

1. Baldi et al., CCS'11 [18]
2. Ayday et al., WPES'13 [11]
3. Naveed et al., CCS'14 [160]
4. Djatmiko et al., WPES'14 [70]

Additional Papers

See [215], [28], [65], [195], [150]

4.3.2 Genetic Relatedness

We next look at genetic relatedness, i.e., testing to ascertain genealogy or ancestry of individuals. Genealogy tests determine whether two individuals are related (e.g., father and child) or to what degree (e.g., they are n^{th} cousins), while, ancestry tests estimate an individual's genetic "pool" (i.e., where their ancestors come from). These tests are often referred to as part of "recreational genomics", and are one of the drivers of the DTC market (with 23andMe and AncestryDNA offering them at under $100). However, due to the hereditary nature of the human genome, they also raise several privacy concerns [158]. Privacy research in this area aims to support privacy-respective versions of such tests.

Baldi et al. [18] allow two users, each holding a copy of their genome, to simulate *in vitro* paternity tests based on Restriction Fragment Length Polymorphisms (RFLPs), without disclosing their genomes to each other or third-parties, through the use of private set intersection protocols [66]. He et al. [106] let individuals privately discover their genetic relatives by comparing their genomes to others stored, encrypted, in the same biorepository, using fuzzy encryption [73] and a novel secure genome sketch primitive, which is used to encrypt genomes using a key derived from the genome itself. Finally, Naveed et al. [160] rely on C-FE to enable a client to learn certain functions, including paternity and kinship, over encrypted data, using keys obtained from a trusted authority.

The tools above differ in a few aspects. First, [18] assumes individuals obtain and store a copy of their sequenced genome, whereas [106] and [160] operate under the assumption that users will rely on cloud providers. Second, [18] operates on full genomes, while [160] supports SNP profiles obtained from DTC genomics companies, with [106] requiring individuals' haplotypes.

Genetic Relatdness – Selected Papers

1. Baldi et al., CCS'11 [18]
2. He et al., Genome Research'14 [106]
3. Naveed et al., CCS'14 [160]

Additional Papers

See [110], [67], [150]

4.3.3 Access and Storage Control

Next, we discuss results aiming to guarantee secure access to, and storage of, genomic data. Karvelas et al. [131] use a special randomized data structure based on Oblivious RAM (ORAM) [95] to store data while concealing access patterns, using two servers to cooperatively operate the ORAM. Clients can then query data using a third entity who retrieves encrypted data from the ORAM and instructs the servers to jointly compute functions using secure two-party computation [232]. Ayday et al. [10] present a framework for privately storing, retrieving, and processing SAM files where a CI sequences and encrypts patients' genomes, and also creates the SAM files, storing them encrypted in biorepositories. Then, MCs using order-preserving encryption [5] can retrieve data and conduct genetic tests.

Beyond SAM files, genomic data is also stored in BAM (a binary version of SAM) or CRAM files, which allows a lossless compression. Huang et al. [113] introduce a Secure CRAM (SECRAM) format, supporting compression, encryption, and selective data retrieval. SECRAM requires less storage space than BAM, and maintains CRAM's efficient compression and downstream data processing. Finally, Huang et al. [114] focus on long-term security, introducing GenoGuard, a system aiming to protect encrypted genomic data against an adversary who tries to brute-force the decryption key (likely to succeed in 30 years). They rely on Honey Encryption (HE) [128] so that, for any decryption attempt using an incorrect key, a random yet plausible genome sequence is produced. Overall, we find that security issues in this context are not explored in as much depth as other areas.

Access and Storage Control – Selected Papers

1. Karvelas et al., WPES'14 [131]
2. Ayday et al., DPM'14 [10]
3. Huang et al., IEEE S&P'15 [114]
4. Huang et al., Genome Research'16 [113]

Additional Papers

See [215], [130], [28], [41]

4.3.4 Genomic Data Sharing

We now discuss results in the context of genomic data sharing, which is an important aspect of hypothesis-driven research. Consider, for instance, genome wide association studies (GWAS): to elicit robust conclusions on the association between genomic features and diseases and traits, researchers may need millions of samples [39]. Even if sequencing costs continue to rapidly drop, it is unrealistic to assume that research teams can easily gain access to such a large number of records. Yet, though there is an interest in data sharing, these sharing initiatives face several obstacles, as (1) researchers in isolation may be prevented from (or are hesitant to) releasing data, and (2) might only have patients' consent for specific studies at specific institutions. Therefore, privacy-enhancing methods have been proposed to address these issues.

Kamm et al. [129] present a data collection system where genomic data is distributed among several entities using secret sharing. Secure multiparty computation (MPC) is then used to conduct computations on data, privately, supporting secure GWAS across multiple entities, such as hospitals and biobanks. Xie et al. [230] introduce SecureMA, which allows secure meta-analysis for GWAS. (Meta-analysis is a statistical technique to synthesize information from multiple independent studies [77].) Their framework generates and distributes encryption/decryption keys to participating entities, encrypts association statistics of each study locally, and securely computes the meta-analysis results over encrypted data.

Humbert et al. [116] consider the case of individuals willing to donate their genomes to research. They quantify the privacy risk for an individual using a global privacy weight of their SNPs and use an obfuscation mechanism that functions by hiding SNPs. Wang et al. [226] enable clinicians to privately find similar patients in biorepositories. This could be applied, for instance, to find out how these patients respond to certain therapies. In their paper, similarity is defined as the edit distance [159], i.e., the minimum number of edits needed to change one string into another. Using optimized garbled circuits, they build a genome-wide, privacy-preserving similar patient query system. This requires participating parties (e.g., medical centers) to agree on a public reference genome and independently compress their local genomes using a reference genome, creating a Variation Call Format (VCF) file. The edit distance of two genomes can then be calculated by securely comparing the two VCF files.

Jagadeesh et al. [124] enable the identification of causal variants and the discovery of previously unrecognized disease genes while keeping 99.7% of the participants' sensitive information private using MPC. Finally, Chen et al. [48] introduce a framework for computing association studies for rare diseases (e.g., the Kawasaki Disease [135]) over encrypted genomic data of different jurisdictions. They rely on Intel's Software Guard Extensions (SGX), which isolates sensitive data in a protected enclave and allows the secure computation of the results.

In summary, work in this category focuses on a wide range of problems, from GWAS and meta-analysis to edit distance computation. Also, tools primarily build on cryptographic protocols, except for [48], which relies on SGX.

Genomic Data Sharing – Selected Papers

1. Kamm et al., Bioinformatics'13 [129]
2. Xie et al., Bioinformatics'14 [230]
3. Humbert et al., WPES'14 [116]
4. Wang et al., CCS'15 [226]
5. Jagadeesh et al., Science'17 [124]
6. Chen et al., Bioinformatics'17 [48]

Additional Papers

See [207], [240], [14], [225], [222]

4.3.5 Outsourcing

At times, research and medical institutions might lack the computational resources required to store or process large genomic datasets locally. As such, there is increasing interest in outsourcing data computation to the cloud, e.g., using dedicated services like Google Genomics or Microsoft Genomics. However, this requires users to trust cloud providers, which raises security and privacy concerns with respect to data of research volunteers and/or patients. To address these concerns, several solutions have been proposed. Note that this category relates to the processing of genomic data in a public cloud environment, whereas, the previously discussed Access & Storage Control category relates to where and how data is stored, *regardless* of the location.

Chen et al. [49] propose the use of Hybrid Clouds [88], a method that involves both public and private clouds, to enable privacy-preserving read mapping. Read mapping is the process of interpreting randomly sampled sequence reads of the human genome. Their solution involves two stages: a seeding stage where the public cloud performs exact matching on a large amount of ciphertexts, and an extension stage where the private cloud computes a small amount of computations (such as, edit distance) at the genetic locations found by the seeding stage. Yasuda et al. [233] present a somewhat homomorphic encryption scheme (SWHE) for secure pattern matching using Hamming distance. More specifically, in this setting physicians supply patients with homomorphic encryption keys who then encrypt their genomic data and upload them to the cloud. When the physician needs to test whether a certain DNA sequence pattern appears in the patient's genome, the cloud computes the Hamming distance over encrypted DNA sequences and the desired pattern, and sends the (encrypted) result back to the physician.

Cheon et al. [50] also use SWHE to calculate the edit distance of two encrypted DNA sequences, allowing data controllers (e.g., patients) to encrypt their genomic data and upload them to the cloud, which can calculate the edit distance to the reference genome or other encrypted sequences. Lauter et al. [140] introduce a leveled homomorphic encryption scheme (LHE) to securely process genomic data in the cloud for various genomic algorithms used in GWAS, such as Pearson and χ^2 Goodness-of-Fit statistical tests.[2] Usually, computation of these statistics require frequencies or counts but, since their scheme cannot perform homomorphic divisions, [140] have to modify some of these computations to work with counts only.

Kim and Lauter [136] also use SWHE to securely compute minor allele frequencies and χ^2-statistics for GWAS-like applications, over encrypted data, as well as the edit/Hamming distance over encrypted genomic data. Finally, Sousa et al. [203] rely on SWHE and private information retrieval to let researchers search variants of interest

[2] LHE is a fully homomorphic encryption scheme variant that does not require bootstrapping but can evaluate circuits with a bounded depth.

in VCF files stored in a public cloud. Their solution represents an improvement upon the state of the art in terms of efficiency, however, it suffers from high error rates and poor scalability.

Outsourcing – Selected Papers

1. Chen et al., NDSS'12 [49]
2. Yasuda et al., CCSW'13 [233]
3. Lauter et al., LatinCrypt'14 [140]
4. Cheon et al., FC'15 [50]
5. Kim and Lauter, BMC'15 [136]
6. Sousa et al., BMC'17 [203]

Additional Papers

See [34], [231], [241], [93]

4.3.6 Statistical Research

The last category focuses on attempts to address unintended leakage threats from the disclosure of genomic data statistics, e.g., membership inference attacks discussed in Section 2.

A possible defense is through statistical disclosure control, of which differential privacy (DP) is one related approach. DP enables the definition of private functions that are free from inferences, providing as accurate query results as possible, while minimizing the chances for an adversary to identify the contents of a statistical database [74].

Johnson and Shmatikov [127] point out that it is inherently challenging to use DP techniques for GWAS, since these methods output correlations between SNPs while the number of outputs is far greater than that of the inputs (i.e., the number of participants). In theory, it is possible to limit the number of available outputs and provide results with adequate accuracy [25, 80]. In practice, however, this requires researchers to know beforehand what to ask (e.g., the top-k most significant SNPs), which is usually infeasible because finding all statistically significant SNPs is often the goal of the study. To address this issue, [127] define a function based on the exponential mechanism, which adds noise and works for arbitrary outputs. Their mechanism allows researchers to perform exploratory analysis, including computing in a differentially private way: i) the number and location of the most significant SNPs to a disease, ii) the p-values of a statistical test between a SNP and a disease, iii) any correlation between two SNPs, and iv) the block structure of correlated SNPs.

Uhlerop et al. [218] aim to address Homer's attack [109] using a differentially private release of aggregate GWAS data, supporting a computation of differentially private χ^2-statistics and p-values, and provide a DP algorithm for releasing these statistics for the most relevant SNPs. They also support the release of averaged minor allele frequencies (MAFs) for the cases and for the controls in GWAS. Tramèr et al. [214] build on the notion of Positive Membership Privacy [143] and introduce a weaker adversarial model, also known as Relaxed DP, in order to achieve better utility by identifying the most appropriate adversarial setting and bounding the adversary's knowledge. Finally, Backes et al. [17] study privacy risks in epigenetics, specifically, showing that blood-based microRNA expression profiles can be used to identify individuals in a study, and propose a DP mechanism to enable privacy-preserving data sharing.

Statistical Research – Selected Papers

1. Johnson and Shmatikov, KDD'13 [127]
2. Uhlerop et al., JPC'13 [218]
3. Tramèr et al., CCS'15 [214]
4. Backes et al., USENIX"16 [17]

Papers on statistical methods are not reported in Table 4.1 because the systematization criteria do not apply, but, for context, they are discussed in Section 4.4.3.

Additional Papersk

See [234], [242], [126], [225], [198].

4.4 Systematic Analysis

This section reports on the systematic analysis of privacy-enhancing technologies in the genomics context, as they stand today, building on the methodology and the research results discussed in Section 4.2 and 4.3, respectively. We drive our discussion from Table 4.1, which, in addition to providing an overview of the community's work, concisely summarizes the results of the analysis. It further enables a discussion on insights, research gaps, as well as challenges to certain assumptions. In the process, we highlight a list of ten technical challenges.

4.4.1 The Issue of Long-Term Security

The longevity of security and privacy threats stemming from the disclosure of genomic data is substantial for several notable reasons. First, access to an individual's genome allows an adversary to deduce a range of genomic features that may also be relevant for her descendants, possibly several generations down the line. Thus, the sensitivity of the data does not necessarily degrade quickly, even after its owner has deceased.

	Data Type	Genomic Assumptions	Storage Location	Long-Term Security	Third Parties	Security Assumptions	Methods	Privacy Overhead	Utility Loss
Personal Genomic Testing									
Baldi et al. [18]	FSG	Yes	User	No	AA	SH, NC	A-PSI	HSO, LTO	Low
Ayday et al. [11]	SNP	No	Cloud	No	SPU	SH, NC	Paillier, Proxy	HSO, LTO	Low
Naveed et al. [70]	SNP	No	User/Cloud	No	N/CS	SH	Paillier, PIR	LSO, LTO	High
Djatmiko et al. [160]	SNP	No	Cloud	No	CA, CS	SH, NC	C-FE	LSO, LTO	Low
Genetic Relatedness Testing									
Baldi et al. [18]	FSG	Yes	User	No	No	SH	PSI-CA	LSO, LTO	High
He et al. [106]	SNP	No	Cloud	No	CS	SH	Fuzzy	LSO, LTO	High
Naveed et al. [160]	SNP	No	Cloud	No	CA, CS	SH, NC	C-FE	LSO, LTO	Low
Access & Storage Control									
Karvelas et al. [131]	FSG	No	Cloud	No	CS	SH, NC	ElGamal, ORAM	HSO, HTO	Low
Ayday et al. [10]	FSG	No	Cloud	No	CS, MK	SH, NC	OPE	HSO, LTO	Low
Huang et al. [114]	FSG	Yes	Cloud	Yes	CS	SH	HoneyEncr	LSO, HTO	High
Huang et al. [113]	FSG	No	User/Cloud	No	No	SH	OPE	LSO, LTO	Low
Genomic Data Sharing									
Kamm et al. [129]	SNP	No	Cloud	Yes	CS	SH, NC	SecretSharing	LSO, HTO	High
Xie et al. [230]	SNP	No	DataController	No	KDC	SH, NC	Paillier, MPC	LSO, LTO	High
Humbert et al. [116]	SNP	No	Cloud	No	No	SH	DataSuppr	—, LTO	Low
Wang et al. [226]	VCF	No	DataController	No	No	SH	MPC	LSO, LTO	High
Chen et al. [48]	SNP	No	DataController	No	No	SGX	SGX	LSO, LTO	High
Jagadeesh et al. [124]	FSG	No	User	No	No	SH	MPC	Varies*	Low
Outsourcing									
Chen et al. [49]	FSG	No	Cloud	No	CS	SH	Hash	HSO, HTO	Low
Yasuda et al. [233]	FSG	No	Cloud	No	CS	SH	SWHE	LSO, LTO	High
Lauter et al. [140]	SNP	No	Cloud	No	CS	SH	LHE	LSO, LTO	High
Cheon et al. [50]	FSG	No	Cloud	No	CS	SH	SWHE	LSO, HTO	High
Kim and Lauter [136]	FSG	No	Cloud	No	CS	SH	SWHE	LSO, HTO	Low
Sousa et al. [203]	VCF	No	Cloud	No	CS	SH	SWHE, PIR	—, LTO	High

Data Type: FSG: Fully Sequenced Genome, SNP: SNPs, Hap: Haplotypes, VCF: Variation Call Format
Third Parties: CS: Cloud Storage, SPU: Storage & Processing Unit, AA: Authorization Authority, CA: Central Authority, KDC: Key Distribution Center, MK: Masking & Key Manager, No: No Third Party
Security Assumptions: NC: No Collusions, SGX: Software Guard Extensions
Methods: SWHE: Somewhat Homomorphic Encryption, LHE: Leveled Homomorphic Encryption, Fuzzy: Fuzzy Encryption, PSI-CA: Private Set Intersection Cardinality, A-PSI: Authorized Private Set Intersection, C-FE: Controlled Functional Encryption, HoneyEncr: Honey Encryption, OPE: Order-Preserving Encryption, MPC: Secure Multiparty Computation, PIR: Private Information Retrieval, SGX: Software Guard Extensions
Privacy Overhead: LSO: Low Storage Overhead, HSO: High Storage Overhead, LTO: Low Time Overhead, HTO: High Time Overhead
****Varies***: Depends on the Input Size

TABLE 4.1: A systematic comparison of the representative genomic privacy methodologies. The rows represent each work and the columns represent the list of criteria we apply for assessment purposes.

Moreover, the full extent of the inferences one can make from genomic data is still not clear, as researchers are still studying and discovering the relationship between genetic mutations and various phenomena.

These issues also imply that the threat model under which a volunteer decides to donate their genome to science, or have it tested by a DTC company, is likely to change in the future. As a consequence, the need or desire to conceal one's genetic data might evolve. For instance, a family member may decide to enter politics, or a country's political landscape shifts toward supporting racist ideologies aimed to discriminate against members of a certain ancestral heritage.

Inadequacy of Standard Cryptographic Tools. We find that the vast majority of genome privacy solutions rely on cryptographic tools, yet, they are not fit for purpose if long-term security is to be protected. Modern cryptosystems assume that the adversary is computationally bounded, vis-à-vis a "security parameter." Suggestions for appropriate choices of the value for this parameter, and resulting key sizes, are regularly updated by the cryptography community, however, assuming at most the need for security for thirty to fifty years [200]. While this timeframe is more than adequate in most cases (e.g., classified documents get regularly de-classified and financial transactions/records become irrelevant), it may not be in the case of genomic data.

In theory, one could increase key lengths indefinitely, but, in practice, this is not possible for all cryptosystems, e.g., the block and stream ciphers available today are only designed to work with keys up to a certain length, and libraries implementing public-key cryptography also impose a limit on key sizes. Furthermore, flaws in cryptosystems considered secure today may be discovered (as happened, recently, with RC4 or SHA-1), and quantum computing might eventually become a reality [238].

Implications. Naturally, the issue of long-term security affects different genome privacy solutions in different ways. For instance, if genomic information is stored in an encrypted form and processed by a specialized third entity, such as the SPU in [11], then a malicious or compromised entity likely has multiple chances over time to siphon encrypted data off and succeed in decrypting it in the future. This is also the case in settings where biobanks store patients' encrypted SAM files [10] or in the context of secure outsourcing solutions, where genomic information is offloaded and encrypted, to a cloud provider. On the other hand, if encrypted data is only exchanged when running cryptographic protocols, but not stored long-term elsewhere (as in [18, 70, 230]), then the adversary has a more difficult task. Nonetheless, long-term security compromise is still possible, even by an eavesdropping adversary and even if the protocol run is super-encrypted using TLS. In fact, documents leaked by Edward Snowden revealed that the NSA has tapped submarine Internet cables and kept copies of encrypted traffic [148, 35].

Possible Countermeasures. Ultimately, genome privacy literature has not sufficiently dealt with long term security. In fact, only the work by Huang et al. [114] attempts to do so, relying on Honey Encryption to encrypt and store genomic data. Though a step in the right direction, this technique only serves as a storage mechanism and does not support selective retrieval of genomic information, testing over encrypted data, and data sharing. Moreover, it suffers from several security limitations. Specifically, while their solution provides information-theoretic guarantees (and long-term security), their threat model needs to account for possible side-channel attacks. This is because, if the adversary knows some of the target's physical traits (e.g., hair color or gender), then it can easily infer that the decryption key she is using is not the correct one. The authors attempt to address this issue by making their protocol phenotype-compatible for the cases of gender and ancestry, but there are many other traits in the human genome that possess probabilistic genotype-phenotype associations [145] thus making it very hard to fully address.

Cryptosystems providing information theoretic security could help, as they are secure even when the adversary has unlimited computing power. Unfortunately, they require very large keys and do not support the homomorphic properties needed to perform typical requirements for genomic data (e.g., testing or sharing). Work relying on secret sharing (e.g., [129]) is somewhat an exception, in that it can provide information-theoretic guarantees. However, for secret sharing to work, one needs non-colluding entities, which is a requirement that is not always easy to attain (see Section 4.4.2).

P1. Long-Term Security

An individual's genomic sequence does not change much over time, thus, the sensitivity of the information it conveys may not diminish. However, cryptographic schemes used by PETs in genomics guarantee security only for 20-30 years.

4.4.2 Security Limitations

Next, we focus on a number of security assumptions made by some genome privacy protocols.

Semi-honest Adversaries. All papers listed in Table 4.1, as well as the vast majority of genome privacy solutions, consider only semi-honest security. Rare exceptions are represented by possible *extensions* to [19, 18]. This is because solutions in this model are significantly easier to instantiate and yield computation and communication complexities that are orders of magnitude lower than in the malicious model.

However, security in the semi-honest model assumes that the parties do not deviate from the protocol and fails to guarantee correctness (i.e., a corrupted party cannot cause the output to be incorrectly distributed) or input independence (i.e., an adversary cannot make its input depend on the other party's input) [105]. Moreover, in the semi-honest model, parties are assumed to not alter their input. In practice, these requirements impose important limitations on the real-world security offered by genome privacy solutions. Specifically, it might not suffice to ensure that protocols only disclose the outcome of a test to a testing facility or provide hospitals with only information about common/similar patients. Indeed, this makes no guarantees as to whether the contents of the test or the patient information has not been maliciously altered or inflated. Additionally, the privacy layer makes it more difficult and, at times, impossible, to verify the veracity of the inputs.

P2. Malicious Security

Most genome privacy solutions are designed for settings where the adversaries are considered to be honest-but-curious as opposed to malicious, which may impose limitations on real-world security.

Non-Collusion. We also observe that a number of solutions that involve third parties (e.g., for storage and processing encrypted genomic data [11], issuing keys [160], and authorizing tests [18, 160]) assume that these parties do not collude with other entities. Such an assumption has implications of various degrees in different contexts. For instance, [160] assumes that a central authority (CA) is trusted to issue policies (i.e., generating one-time decryption keys, allowing researchers to access a specific genome for a specific case). The CA is expected to be operated by some established entity such as the FDA, so that one can likely assume it has no incentive to misbehave (unless compromised). Similarly, protocols supporting large-cohort research, like the one in [230], involve medical centers with little or no economic incentive to collude, and violate patients' privacy.

On the other hand, in some cases, non-collusion might be harder to enforce, while the consequences of collusion might be serious. For instance, the framework in [11] supports private disease susceptibility tests, and involves three entities: (i) the Patient, (ii) the MC, which administers the tests, and (iii) the SPU, which stores patients' encrypted SNPs. Data stored at the SPU is anonymized. However, if the SPU and MC collude, then the SPU can re-identify patients. Moreover, the MC's test specifics must be considered sensitive (e.g., a pharmaceutical company's intellectual property), otherwise there would be no point in performing *private* testing. This is because one could simply tell the patient/SPU which SNPs to analyze and run the test locally. However, patient and SPU collusion implies that confidentiality of the MC's test would be lost. Also, solutions that assume third-party cloud storage providers do not collude with testing facilities, such as [131], are limited to settings where one can truly exclude financial or law enforcement disincentives to collusion.

P3. Non-Collusion Assumption

Some genome privacy solutions involve a collection of entities. These solutions further assume that the entities do not collude with each other, which may be difficult to enforce or verify.

Trusted Hardware. Other assumptions relate to secure hardware, like SGX, which isolates sensitive data into a protected enclave, thus supporting secure computation of the results, even if the machine is compromised. For instance, [48] relies on secure hardware to enable institutions to securely conduct computations over encrypted genomic data. However, side-channel attacks have been recently demonstrated to be possible [36, 100] and the full extent of SGX security has yet to be explored.

P4. Trusted Hardware

Some genome privacy solutions rely on trusted hardware, such as SGX. However, the security of such hardware is not yet fully understood and side-channel attacks may limit the security of these solutions.

4.4.3 The Cost of Protecting Privacy

Genome privacy research mostly focuses on providing privacy-preserving versions of genomics-related functionalities (e.g., testing, data processing, and statistical research). While some of these functionalities are already in use (e.g., personal genomic tests offered by DTC companies, data sharing initiatives), others do not yet exist, at least in the way the genome privacy community has envisioned them. For instance, some investigations assume that individuals will soon be able to obtain a copy of their fully sequenced genome [18] or that we will be able to create an infrastructure and a market with dedicated providers to store and process genomic data for third-party applications [11, 131]. Table 4.1 attempts to evaluate the overhead incurred by privacy protection on efficiency and scalability, by comparing to that of supporting its functionality in a non privacy-preserving way. Similarly, we measure the loss in utility and flexibility.

Privacy Overhead. We observe that high privacy overhead is linked to the use of expensive cryptographic tools, (e.g., ORAM, Paillier, and SWHE). On the one hand, we can assume that some might become increasingly efficient in the future, thanks to breakthroughs in circuit optimization [202]. Moreover, the efficiency of fully homomorphic encryption has improved several orders of magnitude over the last couple of years [213].

On the other hand, the characteristics of the privacy properties under consideration intrinsically make the problem harder. As a result, it is less likely that efficiency will eventually improve in the foreseeable future. For instance, in personal genomic testing, a basic privacy property is concealing which parts of the genome are being tested. This implies that every single part needs to be touched, even if the test only needs to examine a few positions. Some solutions [11, 18] partially address this issue through means of pre-computation. This is accomplished by encrypting genomic data so that it can be privately searched. However, the ciphertext still needs to be transferred in its entirety. Another example is in the context of genealogy testing, where the goal is to find relatives and distant cousins [106]. Accomplishing this in the encrypted domain requires the presence of a central, non-colluding authority, which, as discussed above, is not always feasible. A similar situation arises in the context of data sharing: while secure two-party computation can efficiently support pairwise privacy-friendly information sharing, these do not scale well to a large number of participating entities.

P5. Privacy Overhead

Some technical genome privacy solutions rely on cryptographic tools (e.g., homomorphic encryption, garbled circuits, or ORAM). These often come with non-negligible computation and communication overheads.

Data Representation. In Table 4.1, we capture the type of data each solution works with. For instance, some protocols operate on SNPs (e.g., [11, 160]), others support FSGs (e.g., [18, 131]). On the one hand, working with FSGs means that researchers and clinicians can consider the genome as a whole, supporting various services, such as research and testing relevant to rare genetic disorders. On the other hand, this might be challenging, especially in the ciphertext domain. For instance, genome sequencing is still not an error-free process: nucleotides are often misread by the sequencing machines, especially when operating at lower costs. Additionally, deletions/insertions of nucleotides are not uncommon and the exact length of the human genome may vary among individuals. Handling with such issues is easier in-the-clear than in the ciphertext domain.

In some cases, solutions like [18] assume simplified formats where the genome is stored and processed as a long vector of nucleotides along with their exact position. Yet, when errors, deletions, or insertions are not identified before encryption, the accuracy of testing will dramatically reduce (testing in [18] requires access to specific positions of a vector containing all nucleotides in the genome, thus, if an unidentified insertion or a deletion occurs, the position would shift and the test would not work). Also, important metadata contained in standard formats (such as, SAM, BAM, and FASTQ) is lost in such a custom representation. (Note that all of the selected papers use *only* the data type reported in Table 4.1 as an input; however, if a tool works with fully sequenced genomes (FSG), it can also support other formats (e.g., one can extract SNPs from an FSG). Finally, a non-negligible portion of genome privacy investigations requires systems to change the way they store and process genomic data, which can create challenging hurdles to adoption.

P6. Data Representation

Some genome privacy solutions make data representation assumptions, e.g., introducing a custom or simplified data format, not taking into account sequencing errors, removing potentially useful metadata.

Utility Loss. Finally, Table 4.1 suggests that, in many instances, the loss in utility, when compared to the corresponding functionality in a non-privacy-preserving setting, is high overall. For instance, this may arise due to data representation assumptions discussed above, or because the functionality needs to be adapted for privacy-enhancing tools to work. Consider that the edit distance algorithm in [50] can only support small parameters (thus, short sequences), while in [140] algorithms like Estimation Maximization need to be modified.

Overall, privacy protection inevitably yields a potential loss in utility, as the exact amount of information that should be disclosed needs to be determined ahead of time and rigorously enforced. As such, a clinician performing a test on the patient's genome loses the freedom to look at whichever data she might deem useful for diagnostic purposes. Similarly, a researcher looking for relevant data in large-cohorts might be limited as to what can be searched. A related consideration can be made with respect to data portability across different institutions. For instance, if a patient's genomic information is encrypted and stored in a hospital's specialized unit [11], and the patient is traveling or visits another medical facility, it may be significantly harder to access and use her data.

P7. Utility Loss (Functionality)

Some genome privacy solutions may result in a loss of utility in terms of the functionality for clinicians and researchers. For instance, privacy tools might limit flexibility, portability, and/or access to data.

The Challenges of Statistical Research. In Section 4.3, we reviewed certain efforts [127, 218, 214, 17] to achieve genome privacy using differentially private mechanisms. The use of DP in the context of statistical research is limited by the inherent trade-off between privacy and utility. DP mechanisms aim to support the release of aggregate statistics while minimizing the risks of re-identification attacks. In this context, every single query yields some information leakage regarding the dataset, and, as the number of queries increases, so does the overall leakage. Therefore, to maintain the desired level of privacy, one has to add more noise with each query, which can degrade the utility of the mechanism. The privacy-vs-utility trade-off is a common theme in DP, although in many settings genomic data can present unique characteristics with respect to its owner, thus further compounding the problem. This challenge is exemplified in a case study by Fredrikson et al. [85], which focused on a DP release of models for personalized Warfarin dosage. In this setting, DP is invoked to guarantee that the model does not leak which patients' genetic markers were relied upon to build the model. They show that, to effectively preserve privacy, the resulting utility of the model would be so low that patients would be at risk of strokes, bleeding events, and even death.

However, in some settings, privacy and utility requirements might not be fundamentally at odds, and could be balanced with an appropriate privacy budget. For instance, [218] show that adding noise directly to the χ^2-statistics, rather than on the raw values, yields better accuracy, while [127] demonstrate that the accuracy of the private statistics increases with the number of patients in the study. Also, Tramèr et al.'s techniques [214] can achieve higher utility than [218, 127] by bounding the adversary's background knowledge. Moreover, it has also shown to be challenging to convince biomedical researchers, who are striving to get the best possible results, to accept a non-negligible toll on utility [150].

P8. Utility Loss (Statistical Research)

Some genome privacy solutions rely on differential privacy, i.e., introducing noise to the data which yields a potential non-negligible loss in utility.

4.4.4 Real-Life Deployment

Relevance to Current Genomics Initiatives. An important aspect of the genome privacy work to date is its relevance to current genomics initiatives and whether solutions introduced can be used in practice, to enhance the privacy of their participants. At the moment, these initiatives deal with privacy by relying on access control mechanisms and informed consent, but ultimately require participants to voluntarily agree to make their genomic information available to any researchers who wish to study it. Surprisingly, we only came across one solution that could be leveraged for this purpose, although it would require infrastructure changes. Specifically, the controlled-functional encryption (C-FE) protocol presented in [160] would allow participants' data to be encrypted under a public key issued by a central authority. This would allow researchers to run tests using a one-time function key, obtained by the authority, which corresponds to a specific test and can only be used for that purpose. This means that the authority would need to issue a different key for each individual, for every request, and for every function. Unfortunately, this is not practical in the context of large-scale research involving millions of participants and hundreds (if not thousands) of researchers. However, there actually is work aiming to address some attacks against *data sharing* initiatives, e.g., membership inference against the Beacon network [196] (see Section 3.1). To address this attack, Raisaro et al. [180] propose that a Beacon should answer positively only if the individuals containing the queried mutation are more than one. They also propose hiding a portion of unique alleles using a binomial distribution and providing false answers to queries targeting them, or imposing a query budget. Wan et al. [221] measure the discriminative power of each single nucleotide variant (SNV) in identifying a target in a Beacon, and flip the top SNVs according to this power, measuring the effects on privacy and utility. However, both solutions make important trade-offs with respect to utility. More specifically, [180] alters or limits the responses of the network, while [221], acknowledging that utility loss is unavoidable, provides ways to calculate (but not solve) this trade-off.

P9. Relevance to Genomics Initiatives

Current genomics initiatives – e.g., All of Us or Genomics England – primarily address privacy by relying on access control mechanisms and informed consent. In the meantime, we lack technical genome privacy solutions that can be applied to such initiatives.

Relevance to Private Market. We also reason about the applicability of PETs to the genomics private market. As a case study, we consider DTC products for health reports and genetic relatedness testing, where the potential drawbacks mentioned earlier (e.g., in terms of utility loss or computational overhead) might be less relevant than in clinical settings.

Today, companies like AncestryDNA and 23andMe provide cheap and seamless services, while investing substantial resources on improving user experience. Moreover, as their customer base grows, they can discover and match a greater number of relatives, as well as increase the accuracy of their models using metadata provided by the users. In return, these companies can profit from monetizing customers' data, e.g., by helping researchers recruit participants for research studies or providing pharmaceutical companies with access to data at a certain price [170]. However, without access to data, their business model is not viable. This is because deploying privacy-preserving testing would require the presence of entities that are willing to operate these services with minimal data monetization prospects. Notably, this is not a new issue in privacy, and similar considerations can be raised about research on privacy-preserving social networking or cloud computing, which has also struggled with adoption.

P10. Relevance to Private Market

Direct-to-consumer genetic testing companies monetize customers' data, and/or use it for research studies. As such, genome privacy solutions for personal genome tests may lack a viable business model.

4.5 Experts' Opinions

The systematic analysis of research using PETs to protect genome privacy led to the identification of a list of ten technical challenges. Aiming to validate and broaden the discussion around them, we sought the viewpoints of experts in the field with respect to their importance and difficulty.

Questionnaire. We designed a short questionnaire, presenting participants with each of the ten challenges (P1–P10 in Section 4.4) and asking four questions for each:

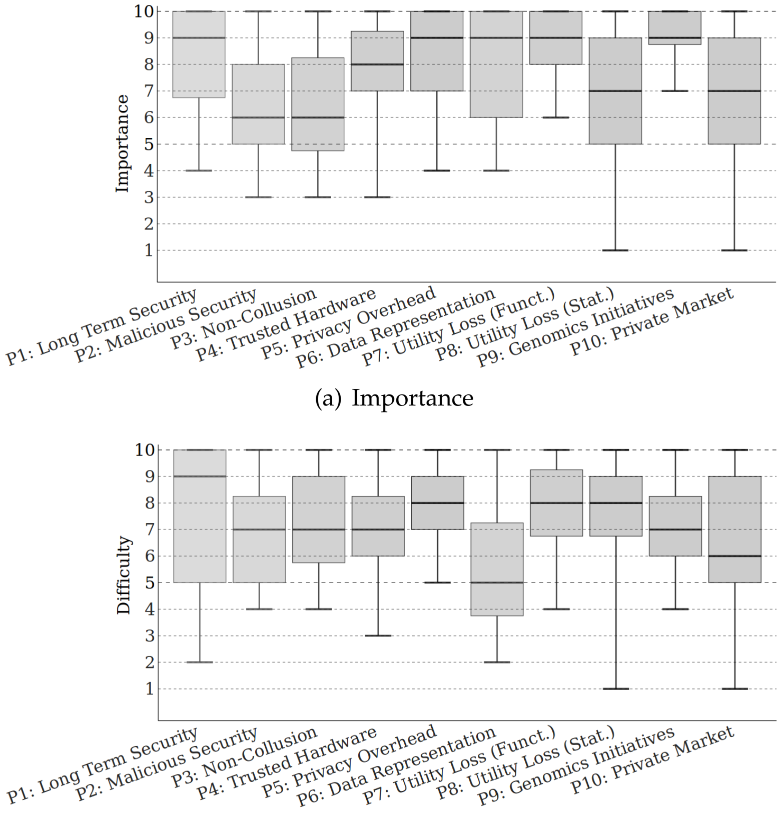
Q1. How important is it to solve this problem?

Q2. How difficult is it to solve this problem?

Q3. What can be done to address this problem?

Q4. How knowledgeable are you in this area?

For Q1-Q2, we used a ten-point Likert scale, with 1 corresponding to "not at all important/difficult" and 10 to "extremely important/difficult." Q3 was a non-mandatory



(a) Importance

(b) Difficulty

FIGURE 4.1: Boxplots of answers to Q1 and Q2 of the online survey.

open-ended question, while Q4 provided three options: unfamiliar, knowledgeable, or expert. The questionnaire took 10-15 minutes to complete.

Participants. To compile a list of genome privacy experts, we again used GenomePrivacy.org, exporting all the authors of the papers listed on the site (262 as of March 2018). We manually inspected each and removed those that appeared to have primarily biomedical or legal backgrounds, or only authored one paper, thus shortening the list to 92 names. Then, we retrieved the email addresses of the authors from their websites and, in April 2018, we sent an email with a request to fill out the questionnaire. After 30 days, we received answers from 21 experts.

The survey was designed to be *anonymous*, i.e., participants were provided with a link to a Google Form and were not asked to provide any personal information or identifiers.

Analysis. Figures 4.1(a) and and 4.1(b) present boxplots of the answers to Q1 (importance) and Q2 (difficulty). For most questions, participants considered themselves experts or knowledgeable; only three identified as unfamiliar for P4: Trusted hardware and P10: (Relevance to) Private Market, two for P8: Utility Loss (Statistical) and P9: (Relevance to) Genomics Initiatives, and one for P3: Non-Collusion, P5: Privacy Overhead, and P6: Data Representation.

Even though the questionnaire was administered with experts in the field, its limited sample and scope does not allow us to perform a quantitative analysis in terms of statistical hypothesis testing with respect to each problem. However, the responses obtained, along with the open-ended questions, do offer several interesting observations. In fact, we found that participants seemed to consider most of the problems to be quite important overall. Looking at the median (M), we find that $M \geq 7$ for eight out of ten problems, and $M=6$ for the other two: P2: Malicious Security and P10: Private Market.

Most Important Problems. Three problems stood out as the most important, with $M=9$. In particular, P1: Long-term Security received an average score of $\mu=8.23\pm2.02$, with 6 out of the 7 self-reported 'experts' in this area giving it a 10. High scores were also obtained by P7: Utility Loss (Functionality), with $\mu=8.85\pm1.16$, and P9: Genomics Initiatives ($\mu=9.09\pm0.97$, the highest). The importance of these problems was confirmed when normalizing scores by the average response of each participant, which let us reason about the "relative" importance; again, P1 and P7 stood out.

When asked about how to solve the problem of long-term security, the experts provided a wide range of ideas, thus confirming the need for future work that takes different approaches than the current ones. For instance, some experts raised the availability of honey encryption [114], which, however, is limited in scope and security (see Section 4.4.1). Others proposed the use of gene manipulation techniques like CRISPR [141], blockchain-based solutions, and post-quantum cryptography. Specifically, one participant said: "use post-quantum or information-theoretic solutions. In theory, we know how to do this (just exchange primitives with PQ-secure ones), but one needs to check security and performance issues". Others focused on transparency of data access and management rather than confidentiality, stating, e.g., that "maybe cryptography is not the answer. Perhaps setting up an environment with different ways of controlling how the data is managed in order to provide more transparency." An expert proposed re-encrypting the data, while another suggested the use of large parameters: "One option is to use particularly large parameters, but this will degrade performance. Of course we can't know what improved attacks are coming up." Finally, a participant responded that this issue is not critical at the moment although it may be in the future, stating that "despite of the issues of impact to future generations, I do not consider this a critical factor in todays operations," while another simply responded that there is not much the community can do to address this issue.

Whereas, when asked about utility loss with respect to functionality, there was some agreement that the community should start to carefully design specific solutions, rather than generic ones, and validate them with clinicians and researchers *before* developing them. For example, one participant mentioned: "we can develop fundamental and flexible evaluations primitives for various applications. Some solution seems to be the best for one case but there may be a better solution to other applications", while another suggested to "work on practical use cases by collaborating with, e.g., medical researchers." Further, an expert suggested to work on hybrid solutions: "The solution might have to be multi-layered with crypto, access control policy, etc. components", while another focused on the users, suggesting to "educate humans to change their ways."

Disagreements. We also found that two problems, P8: Utility (Statistical) and P10: Private Market attracted very *diverse* scores, yielding a variance of, resp., 5.85 and 6.80.

This was confirmed by the participants' open-ended answers: four participants (self-identified as knowledgeable) rejected the use of differentially private mechanisms in clinical settings, while also providing low importance scores. When asked how the community can address this challenge, some explained that, in certain cases, the utility loss can be manageable (e.g., "A comprehensive, experiment-driven analysis for feasible epsilon and delta values can pave the way towards more formally founded utility results for DP"), while another suggested that the utility loss is usually high because the proposed solutions are generic instead of specialized for a specific task. A couple of participants did not recognize the issue of relevance to the private market as particularly important since they found privacy in this context to be fundamentally at odds with the DTC market, while others gave high scores for the very same reason.

Other Problems. Among other observations, we found that P6: Data Representation, although considered important (M=9), was also considered the easiest to address. Specifically, it was suggested that solutions should be designed to better handle *real* data, and that the community should focus on harmonizing existing schemes and improve interoperability. Although this might be feasible, it once again highlights the need for truly interdisciplinary work. By contrast, regarding the scarcity of solutions protecting against malicious adversaries (P2) or not relying on non-collusion assumptions (P3), we found that these were less important to the participants, both due to their confidence in advances in cryptography research but also because they felt these might be reasonable assumptions in the medical setting.

Take-Aways. In summary, the 21 genome privacy experts who responded to our survey evaluated the ten challenges identified through our systematization, helping us put them in context and providing feedback on possible avenues to mitigate them. On the one hand, some of the issues are likely to be eventually mitigated via natural research progress (e.g., cryptographic tools geared to provide malicious security might become more efficient), optimization techniques (e.g., supporting specific functionalities in standardized settings), advances in new technologies (e.g., trusted hardware might reduce computation overheads), and/or inter-disciplinary collaboration (e.g., by improving interoperability). On the other hand, however, we do not have a clear grasp as to how to tackle some of the other challenges that are inherently tied to the very nature of genome privacy. For instance, the issue of long-term security cannot be effectively addressed using standard approaches; similarly, the utility loss stemming from hiding data might be hard to overcome with existing cryptographic and/or differential privacy techniques.

4.6 Discussion

We introduced and applied a structured methodology to systematize knowledge around genome privacy, focusing on defensive mechanisms offered by privacy-enhancing technologies (PETs). We selected a representative sample of the community's work and defined systematic criteria for evaluation. We compiled a comparison table which guided a critical analysis of the body of work and helped us identify ten technical challenges/open problems. We also asked genome privacy experts to weigh in on how important and difficult they are to address and to provide ideas as to how to mitigate them. Overall, our analysis serves as a guideline for future work, while our methodology can be reused by other researchers to revisit new results and assess the progress in the field.

In short, we found that using PETs to protect genome privacy may be hindered by the obstacles related to the unique properties of the human genome. For example, the sensitivity of genome data does not degrade over time; as a consequence, one serious challenge stems from lack of long-term security protection, which is hard to address as available cryptographic tools are not suitable for this goal. We also observed that the overwhelming majority of proposed techniques, aiming to scale up to large genomic datasets, need to opt for weaker security guarantees or weaker models. While it is not unreasonable to expect progress from the community with respect to underlying primitives, it is inherently hard to address the limitations in terms of utility and/or flexibility on the actual functionalities. When combined with assumptions made about the format and the representation of the data under analysis, this might pose major hurdles against real-life adoption.

These hurdles are further compounded by the interdependencies between some of the criteria and the categories discussed. For instance, the use of cloud storage for genomic data implies the existence of a third party, and as such, the improvement in usability may be overshadowed by security limitations (Section 4.4.2). Furthermore, the solutions proposed in the Access & Storage Control category may have a direct effect on every category as functionalities like secure storage and selective retrieval are crucial parts of any complete framework, further highlighting the importance of interoperability.

Nevertheless, in its short lifespan, the genome privacy community has achieved admirable progress. Indeed, a significant portion of research can already be used to enable genomics-related applications that are hard or impossible to support because of legal or policy restrictions. For instance, genetic and health data cannot easily cross borders, which makes international collaborations very challenging. In this context, mechanisms provably guaranteeing privacy-friendly processing of genomic data may alleviate these restrictions and enable important research progress, and we hope to see more pilot studies along these lines in the near future. In fact, some initiatives have started to provide encouraging results, e.g., a trial conducted at the CHUV hospital in Lausanne (Switzerland) to encrypt genetic markers of HIV-positive patients and let doctors perform tests in a privacy-preserving way [150].

Chapter 5

Analyzing Genetic Testing Discourse on the Web Through the Lens of Twitter

5.1 Introduction

In 1990, the Human Genome Project was kicked off with the goal of producing the first complete sequence of a human genome; at a cost of almost $3 billion, it was completed 13 years later [139]. Since then, costs have dropped at a staggering rate: by 2006, high-quality sequencing of a human genome cost $14 million, and, by 2020, private individuals could have their genomes sequenced for about $1,000 [120]. This rapid progress is paving the way to *personalized medicine*, a concept advocating for diagnosis and treatment to be tailored to patients' genetic features, aiming to make healthcare more preventive and effective [9]. It also enables *public initiatives* to sequence large numbers of genomes and build large bio-repositories for research purposes; for instance, the Precision Medicine research program in US (now called All Of Us) or the Genomics England project in UK are sequencing the genomes of, respectively, 1M and 100K volunteers.

Moreover, a number of companies have entered the flourishing market of *direct-to-consumer* (DTC) genetic testing. Rather than visiting a clinic, customers purchase a collection kit for a few hundred dollars or less, deposit a saliva sample, and mail it back; after a few days, they receive a report with information about genetic health risks (e.g., susceptibility to Alzheimer's), wellness (e.g., lactose intolerance), and/or ancestry and genealogy information. Today, there are hundreds of DTC companies – naturally, some more reputable than others [178] – including 23andMe (which provides reports on carrier status, health, and ancestry) and AncestryDNA (which focuses on genealogy and ancestry). As of June 2020, 23andMe and AncestryDNA have, respectively, tested more than 12M and 16M customers [1, 7].

Traditionally, health-related issues were communicated to patients primarily by doctors and clinicians, however, the advent of direct-to-consumer genetic testing changes this substantially. Individuals can now learn potentially life-changing results with a few clicks of the mouse, without contacting a medical professional. Also, as results are delivered electronically, they are more easily shared with others. Affordable DTC products and participatory sequencing initiatives make genetic testing increasingly more accessible and available to the general population. Like with other aspects of digital health, this leads to social media attracting discussions, sharing of experiences, and molding of perceptions around genetic testing, thus becoming a key platform for related news and marketing efforts. However, while the research community has analyzed in great detail the interlinked relationship between health and social networks such as Twitter, to the best of our knowledge, genetic testing discourse on social media has not been adequately studied.

To this end, we set to address a few open questions: 1) What are the tweets related to genetic testing really about? 2) Which accounts are particularly active in tweeting and what do they talk about? 3) Is the discussion about genetic testing dominated by certain keywords, themes, or companies? 4) What is the overall sentiment and what topics relate to more negative sentiment? We focus on Twitter due to its popularity and the relatively ease of collecting data.

Aiming to answer these questions, we present an exploratory, large-scale analysis of Twitter discourse related to genetic testing. Starting from 10 keywords related to DTC genetics companies and 3 to genomics initiatives, we search and crawl all available tweets containing these keywords that were posted between January 1, 2015 and July 31, 2017. We collect 302K tweets from 113K users, and analyze them along several axes, seeking to understand who tweets about genetic testing, what they talk about, and how they use Twitter for that. Specifically, after presenting a general characterization of our dataset, we analyze the tweets content-wise, studying the most common hashtags/URLs and measuring sentiment. Next, we perform a user-based analysis, looking at the profiles and their location, and assessing whether they are likely to be bots. We also select a random sample of 15K users and analyze their latest 1K tweets to study their interests. Note that, as a substantial chunk of tweets turns out to be about DTC companies 23andMe and AncestryDNA, we present a few case studies focused on them. Finally, we examine the most negative tweets in our dataset, finding a number of tweets related to racism and hate-speech, as well as fears of privacy and data misuse, and look for instances of users sharing screenshots of their test results.

Overall, our study leads to a few interesting observations:

1. Users tweeting about genetic testing seem overall interested in digital health and technology, although the conversation is often dominated by those with a vested interest in its success, e.g., specialist journalists, medical professionals, entrepreneurs, etc.

2. The two most popular DTC companies, 23andMe and AncestryDNA, also generate the most tweets. However, although 23andMe had half the customers at the time of measurement, it produced almost 5 times more tweets, which is also due to controversy around their failure to get FDA approval in 2015.

3. Sentiment around initiatives is positive, with coverage boosted by mainstream

	Tweets	Users	RTs	Likes	Official	Media	Quotes	Hashtags	URLs	Top 1M
23andMe	132,597	64,014	72,848	149,897	1.31%	6.14%	3.49%	27.23%	68.68%	75.40%
AncestryDNA	29,071	16,905	16,266	47,249	7.08%	8.79%	2.69%	54.29%	75.50%	49.68%
Counsyl	3,862	1,834	2,716	4,255	3.49%	6.98%	4.64%	44.01%	83.94%	74.97%
DNAFit	2,118	844	1,336	2,508	15.34%	18.74%	5.37%	57.22%	78.94%	79.18%
FamilyTreeDNA	2,794	1,205	1,196	3,111	4.36%	19.97%	6.62%	34.18%	36.47%	69.21%
FitnessGenes	2,142	773	908	2,809	16.29%	18.47%	9.40%	44.53%	56.76%	71.28%
MapMyGenome	1,568	704	4,488	3,726	15.30%	13.13%	4.99%	53.63%	80.35%	64.30%
PathwayGenomics	1,544	579	1,968	2,521	2.13%	18.51%	6.11%	61.01%	76.55%	68.12%
Ubiome	14,420	6,762	9,223	13,991	2.71%	4.37%	2.85%	27.85%	73.28%	64.19%
VeritasGenetics	1,292	497	1,443	2,526	6.65%	17.07%	17.07%	46.13%	58.28%	71.95%
Genomics England	7,009	1,863	19,772	18,756	19.68%	17.80%	11.58%	61.19%	69.18%	48.82%
Personalized Medicine	20,302	4,631	19,085	15,514	–	6.93%	7.55%	99.93%	87.42%	71.98%
Precision Medicine	83,329	13,012	118,043	128,303	–	8.56%	10.41%	99.88%	83.39%	77.16%
Total	302,048	113,624	269,292	395,166	2.26%	7.75%	5.92%	56.54%	74.77%	71.80%
Baseline	163,260	131,712	282,063,006	486,960,753	–	41.20%	12.07%	23.48%	45.49%	89.57%

TABLE 5.1: Our keyword dataset, with all tweets from January 1, 2015 to July 31, 2017 containing keywords related to genetic testing.

news and announcements (e.g. President Obama's) and neutral for DTC companies, although with a few strongly opinionated users.

4. There is a clear distinction in the marketing efforts undertaken by different companies, which naturally influence the type and the nature of users' engagement; e.g., we find the the promotional hashtag #sweepstakes in 1 out 8 tweets containing the keyword AncestryDNA.

5. We find a limited presence of social bots, with some keywords attracting a different degree of automated publishing, as some topics seem to be more popular among individuals than others.

6. We find evidence of groups using genetic testing to push racist and anti-semitic agendas, and of users expressing concerns about privacy and data protection.

7. A non-negligible amount of users share and discuss screenshots of their ancestry test results, despite the possible privacy implications.

5.2 Dataset

We now describe our methodology for collecting the dataset.

Genetic Testing Keywords. We start from a list of 36 DTC genetic testing companies compiled by the International Society of Genetic Genealogy [123]. We use each company's name as a search keyword; if the search returns less than 1,000 tweets, we discard it. In the end, we collect tweets for 10 companies: 23andMe, AncestryDNA, Counsyl, DNAFit, FamilyTreeDNA, FitnessGenes, MapMyGenome, PathwayGenomics, Ubiome, and VeritasGenetics. We opt for keywords not separated by spaces (e.g., VeritasGenetics) rather than quoted search (e.g., "Veritas Genetics") since we notice that companies are primarily discussed via hashtags or mentions, and because Twitter's search engine, at the time of the collection, did not provide exact results with quotes.[1]

Besides tweets related to for-profit companies, we also want to study discourse related to public sequencing initiatives and related concepts. Thus, we select three more keywords: PrecisionMedicine, PersonalizedMedicine, and GenomicsEngland. Personalized Medicine aims to make diagnosis, treatment, and care of patients tailored and optimized to their specific genetic makeup. Precision Medicine conveys a similar concept, but also refers to the initiative sequencing the genome of 1M individuals announced by President Obama in 2015 to understand how a person's genetics, environment, and lifestyle can help determine the best approach to prevent or treat disease [90]. Genomics England is a similar UK initiative with 100K volunteers, primarily focusing on cancer and rare disease research. Once again, we search for keywords not separated by spaces (e.g., PrecisionMedicine) since these concepts are mostly discussed via hashtags and because of the incorrectness of the search engine.

Twitter Dataset. We use a Python library called Tweepy [216] that is a wrapper of Twitter's official API to collect all posted tweets from January 1, 2015 to July 31, 2017 returned as search results using the 10 DTC keywords and the 3 keywords related to genomics initiatives. Our script collects, for each tweet, its content, the username, date

[1] For instance, when searching for tweets using the quoted search "Family Tree DNA," we expect to only get the tweets that include these exact words, in that order, including the spaces. However, we notice that the following tweet appears: "Celebrate Valentine's Day & give the gift of Family Finder for only $59. Sale ends February 14th! `http://familytreedna.com`"

and time, the number of retweets and likes, as well as the URL of the tweet. It also visits the profile of the users posting each tweet, collecting their location (if any), the number of followers, following, tweets, and likes. Overall, we collect a total of 191K tweets from 94K users for the 10 DTC companies and 111K from 19K users for the 3 initiatives, as summarized in Table 5.1. Note that to download this information one needs the prior consent of Twitter. When applying for an API key, Twitter asks the applicant to explain their reasons for doing so. We describe in our form our intended methodology and our application is accepted. We also collect a set of 163,260 random English tweets, from the same January 2015 to July 2017 period (approx. 170 per day), which serves as a baseline set for comparisons. This set originates from a pre-existing dataset of tweets collected using Twitter's "Sample Stream" API which returns a random sample of all public tweets.[2]

We remark that the keyword search returns accounts that match that keyword, e.g., tweets including 23andMe, #23andMe, or @23andMe, but also those posted by the @23andMe account. For consistency, we discard the latter, analyzing them separately when relevant. Note that our dataset includes tweets from users who discuss their opinions on genetic testing, but also blog posts, ads, and news articles. As our goal is to discover how genetic testing is reflected through the lens of Twitter, we choose not to discard any of the above subsets in an attempt to "clean" the dataset, or to focus only on certain kinds of profiles. We also make our Twitter dataset available to the public.[3]

5.3 Analyzing the Genetic Testing Discourse

In this section, we present the results of our exploratory, large-scale analysis of the genetic testing discourse online on Twitter. To do so, we study the dataset presented in

[2] See https://developer.twitter.com/en/docs/tweets/sample-realtime/

[3] https://github.com/amittos/genetic-testing-twitter-dataset

Section 5.2 on multiple axes, such as the most used hashtags and URLs, the sentiment, and the nature of the users who tweet about genetic testing.

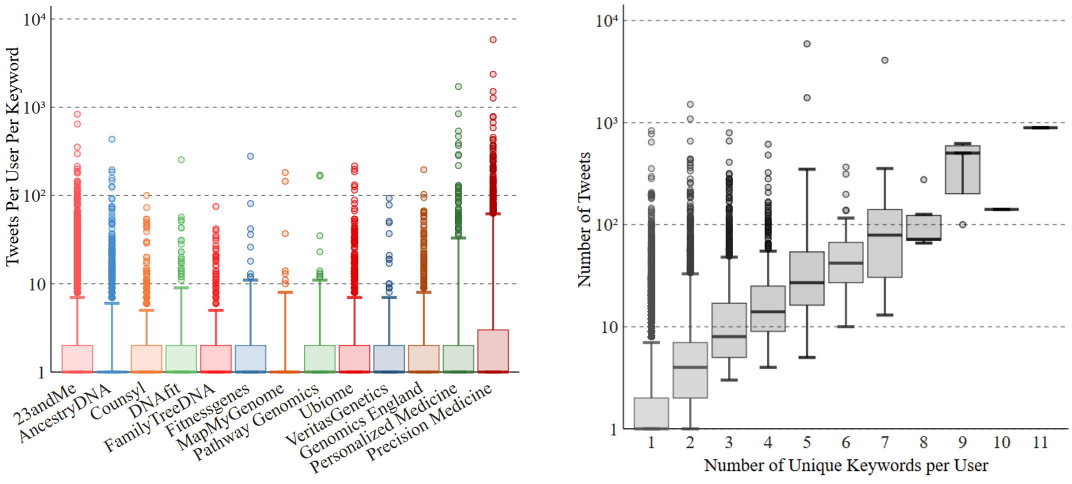
5.3.1 General Characterization

We start by presenting a general characterization of the tweets in our dataset. Simple statistics of our keyword-based dataset are reported in Table 5.1. From left to right, the table lists the total number of tweets, unique users, retweets, and likes for each of the 13 keywords and the random baseline. We also quantify the percentage of tweets made by the official accounts of each company or initiative, as well as the percentage of tweets including media (images and videos), quoted tweets, hashtags, and URLs, and how many of them are in the Alexa Top 1M.[4]

Number of tweets. 23andMe is by far the most popular keyword, with one order of magnitude more tweets than any other company (130K in total, around 140/day, from 64K distinct users); AncestryDNA is a distant second (30K tweets from 16.9K users). Given their large customer bases, this should not come as a surprise. However, it *is* surprising that 23andMe has 4.6 times as many tweets as AncenstryDNA even though AncestryDNA, at the time of measurement, had over twice the number of customers of 23andMe. The least popular companies are MapMyGenome, PathwayGenomics, and VeritasGenetics, with less than 2K tweets each over our 2.5 year collection period. Among the initiatives, Precision Medicine generates a relative high number of tweets (83K from 13K users), much more so than Personalized Medicine (20K tweets).

Tweets per user. For each keyword, we also measure the number of tweets per user (see Figure 5.1(a)). We find that the median for every keyword is 1; i.e., 50% of users tweet about a given DTC company or initiative only once. However, we do find differences in the outliers for different keywords. For instance, there are several highly engaged

[4] https://www.alexa.com/topsites



(a) (b)

FIGURE 5.1: Number of tweets (a) per keyword, and (b) per user as a function of the number of unique keywords they tweeted about. Note the log scale in y-axis.

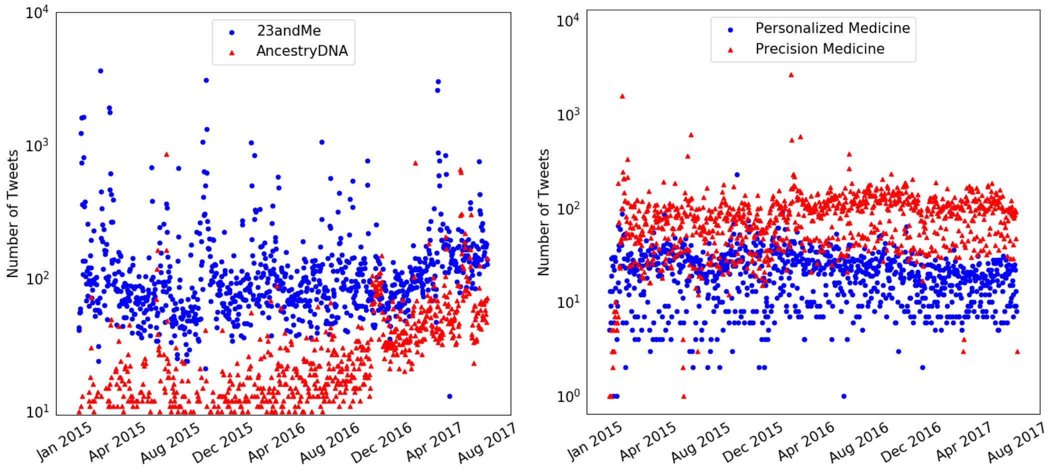
users tweeting about Personalized Medicine and Precision Medicine. Manual examination of these users indicates that most of them are medical researchers and companies actively promoting the initiatives as hashtags. The presence of these heavily "invested" users becomes more apparent when we look at the number of tweets as a function of the number of unique keywords a user posts about, as plotted in Figure 5.1(b): 95% of them post about only one keyword, and those that post in more than one tend to post *substantially* more tweets about genetic testing in general; in some cases, orders of magnitude more tweets.

We also find differences between tweets about DTC genetic testing companies and those about genomics initiatives. The majority of the latter come from a smaller set of users compared to the former, i.e., a few very dedicated users drive the discussion about genomics initiatives. This is clear from Figure 5.1(a), which plots the number of tweets per user for each keyword: Personalized/Precision Medicine have more outliers than most of the DTC genetic companies (although the median for all keywords is 1). We also find these tweets are more likely to contain URLs (87% and 83% of tweets, respectively) than most companies, and even more so when compared to the baseline (45%).

This suggests that tweets about these topics often include links to news and/or other external resources. Only around 50% of URLs linked from tweets related to Genomics England or AncestryDNA are in the Alexa top 1M, compared to 60–75% for other keywords. For Genomics England, this is due to many URLs pointing to genomicsengland.co.uk itself. For AncenstryDNA, whose official site at ancestry.com *is* in the top 1M, it appears to be due a large number of marketing URLs tweeted along with the keyword; which we discuss later on.

Retweets and Likes. The total number of retweets and likes per tweet in the baseline is substantially higher than for tweets related to genetic testing due to outliers, i.e., viral tweets or tweets posted by famous accounts (e.g., a tweet by @POTUS44 on January 11, 2017 has 875,844 retweets and 1,862,249 likes). However, the median for retweets and likes in the baseline dataset mirrors that of tweets in our keywords dataset, with values between 0 and 1. Note that, although the number of retweets and likes per tweet could be influenced by how old the tweets are, this is not really the case in our dataset. Starting in late-August 2017, we collect tweets posted up to July 2017. This allows ample time to capture likes and retweets since preivous work [138] indicates that 75% of retweets happen within 24 hours and 85% happen within a month.

Official accounts tweets. We also look at the tweets including a given keyword (e.g., Ubiome) made by the corresponding official account (e.g., @Ubiome). There are no official accounts for Personalized and Precision Medicine, however, the Precision Medicine initiative is now called All Of Us and has a Twitter account (created in February 2017) that has posted only a few tweets (224 as of April 4, 2018), so we do not consider it. The percentage of tweets made by the official accounts of most companies including the name of the company as keyword is unsurprisingly very low (e.g., 1% for 23andMe). However, it is higher for others (e.g., 15% for DNAfit, Fitnessgenes, and



(a) 23andMe and AncestryDNA

(b) Personalized/Precision Medicine

FIGURE 5.2: Number of tweets per day. Note the log scale in y-axis.

MapMyGenome), due to the fact that these companies actually add their names in their tweets as a hashtag (e.g., #AncestryDNA). In fact, we find that hashtags are used quite predominantly for several DTC keywords, in some cases 40% of tweets have hashtags vs 23% for baseline tweets.

Temporal analysis. Finally, we analyze how the volume of tweets changes over time. In Figure 5.2, we plot the number of tweets per day in our dataset (between Jan 1, 2015–July 31, 2017) for the two most popular companies (23andMe/AncestryDNA) and the two most popular genomics initiatives (Personalized/Precision Medicine). On average, there are 145 and 30 tweets per day for 23andMe and AncestryDNA keywords, respectively. While the former is relatively constant, the latter increases steadily in 2017 (Figure 5.2(a)). This may be the result of AncestryDNA's aggressive promotion strategies (see Section 5.3.2). We also find a number of outliers for 23andMe, mostly around Feb 20 and Oct 19, 2015, and Apr 6, 2017, which are key dates related to 23andMe's failure to get FDA approval for their health reports in 2015, then obtained in 2017 [78]. In fact, 20K/132K 23andMe tweets are posted around those dates. As for Personalized and Precision Medicine (Figure 5.2(b)), the volume of tweets stays relatively flat. There

		– Without Official Accounts –		– Only Official Accounts –	
Keyword	**WH**	**Top 3 Hashtags**	**KH**	**Top 3 Hashtags**	**KH**
23andMe	27.09%	dna (3.58%), genetics (2.07%), tech (1.96%)	12.46%	23andMestory (6.67%), genetics (6.35%), video (5.19%)	9.74%
AncestryDNA	75.48%	sweepstakes (12.38%), dna (4.90%), genealogy (4.86%)	25.94%	dna (11.74%), ancestry (5.92%), familyhistory (5.07%)	46.88%
Counsyl	45.24%	getaheadofcancer (2.64%), cap (1.93%), medical (1.94%)	3.08%	acog17 (6.18%), womenshealthweek (5.15%), teamcounsyl (5.15%)	0%
DNAFit	55.30%	diet (4.19%), fitness (3.72%), crossfit (3.54%)	22.91%	dna (5.33%), fitness (3.71%), generictogenetic (3.48%)	40.37%
FamilyTreeDNA	29.31%	dna (14.24%), genealogy (13.42%), ancestryhour (3.18%)	10.86%	geneticgenealogy (5.55%), ftdnasuccess (4.44%), ftdna (3.33%)	56.66%
FitnessGenes	72.19%	startup (5.93%), london (5.73%), job (5.59%)	18.22%	fitness (5.85%), dna (4.32%), gtsfit (2.79%)	45.29%
MapMyGenome	54.98%	shechat (7.94%), appguesswho (5.32%), genomepatri (4.22%)	15.80%	genomepatri (7.28%), knowyourself (4.04%), genetics (2.02%)	0%
PathwayGenomics	55.85%	coloncancer (6.91%), genetictesting (3.29%), cancer (2.85%)	3.34%	dnaday16 (9.67%), ashg15 (9.67%), health (3.22%)	19.35%
Ubiome	28.57%	microbiome (13.23%), tech (2.14%), vote (2.07%)	6.61%	microbiome (24.48%), bacteria (4.76%), meowcrobiome (2.72%)	6.12%
VeritasGenetics	57.16%	brca (3.92%), genome (3.62%), genomics (3.32%)	4.22%	brca (11.82%), liveintheknow (11.82%), wholegenome (10.75%)	0%
Genomics England	62.05%	genomes100k (14.84%), genomics (7.72%), raredisease (5.24%)	1.77%	genomes100k (32.45%), raredisease (19.49%), genomics (18.71%)	0%
Personalized Medicine	–	precisionmedicine (22.74%), genomics (9.77%), pmcon (8.37%)	–	–	–
Precision Medicine	–	genomics (6.70%), personalizedmedicine (5.49%), cancer (4.89%)	–	–	–

TABLE 5.2: Top 3 hashtags for each keyword, along with the percentage of tweets with at least a hashtag (WH) as well as that of of "keyword hashtags" (KH), e.g., #23andMe.

are outliers for Precision Medicine too, e.g., 2,628 tweets on February 25, 2016, when the White House hosted the Precision Medicine Initiative summit [103].

Discussion. Overall, our characterization shows that highly engaged users drive the discussion around public genomics initiatives, which is particularly influenced, at least in terms of volumes, by important announcements such as the one made by President Obama. As for direct-to-consumer (DTC) genetic testing, the conversation is, as expected, dominated by the two most popular companies: 23andMe and AncestryDNA. However, it is interesting that the former generates 4 times more tweets even though the latter has more than twice the customers. Some of this "popularity" seems to be due to 23andMe's controversy around FDA approval. We also find a non-negligible use of hashtags, possibly used for promotion and marketing efforts, and that a lot of tweets include URLs to popular domains, indicating that they are used to disseminate news and links to external resources. This warrants further exploration, thus, we perform hashtag and URL analysis in the next section.

5.3.2 What Are The Tweets About?

Next, we analyze the content of the tweets related to genetic testing, studying hashtags and URLs included in them and performing a simple sentiment analysis.

Hashtag Analysis. In Table 5.3, we report the top three hashtags for every keyword,

		– Without Official Accounts –		– Only Official Accounts –	
Keyword	**WH**	**Top 3 Hashtags**	**KH**	**Top 3 Hashtags**	**KH**
23andMe	27.09%	dna (3.58%), genetics (2.07%), tech (1.96%)	12.46%	23andMestory (6.67%), genetics (6.35%), video (5.19%)	9.74%
AncestryDNA	75.48%	sweepstakes (12.38%), dna (4.90%), genealogy (4.86%)	25.94%	dna (11.74%), ancestry (5.92%), familyhistory (5.07%)	46.88%
Counsyl	45.24%	getaheadofcancer (2.64%), cap (1.93%), medical (1.94%)	3.08%	acog17 (6.18%), womenshealthweek (5.15%), teamcounsyl (5.15%)	0%
DNAFit	55.30%	diet (4.19%), fitness (3.72%), crossfit (3.54%)	22.91%	dna (5.33%), fitness (3.71%), generictogenetic (3.48%)	40.37%
FamilyTreeDNA	29.31%	dna (14.24%), genealogy (13.42%), ancestryhour (3.18%)	10.86%	geneticgenealogy (5.55%), ftdnasuccess (4.44%), ftdna (3.33%)	56.66%
FitnessGenes	72.19%	startup (5.93%), london (5.73%), job (5.59%)	18.22%	fitness (5.85%), dna (4.32%), gtsfit (2.79%)	45.29%
MapMyGenome	54.98%	shechat (7.94%), appguesswho (5.32%), genomepatri (4.22%)	15.80%	genomepatri (7.28%), knowyourself (4.04%), genetics (2.02%)	0%
PathwayGenomics	55.85%	coloncancer (6.91%), genetictesting (3.29%), cancer (2.85%)	3.34%	dnaday16 (9.67%), ashg15 (9.67%), health (3.22%)	19.35%
Ubiome	28.57%	microbiome (13.23%), tech (2.14%), vote (2.07%)	6.61%	microbiome (24.48%), bacteria (4.76%), meowcrobiome (2.72%)	6.12%
VeritasGenetics	57.16%	brca (3.92%), genome (3.62%), genomics (3.32%)	4.22%	brca (11.82%), liveintheknow (11.82%), wholegenome (10.75%)	0%
Genomics England	62.05%	genomes100k (14.84%), genomics (7.72%), raredisease (5.24%)	1.77%	genomes100k (32.45%), raredisease (19.49%), genomics (18.71%)	0%
Personalized Medicine	–	precisionmedicine (22.74%), genomics (9.77%), pmcon (8.37%)	–	–	–
Precision Medicine	–	genomics (6.70%), personalizedmedicine (5.49%), cancer (4.89%)	–	–	–

TABLE 5.3: Top 3 hashtags for each keyword, along with the percentage of tweets with at least a hashtag (WH) as well as that of of "keyword hashtags" (KH), e.g., #23andMe.

Keyword	Without Official Accounts	Only Official Accounts
23andMe	23andMe.com (7.33%), techcrunch.com (3.09%), fb.me (2.48%)	23me.co (50.88%), 23andMe.com (21.13%), instagram.com (5.40%)
AncestryDNA	journeythroughhistorysweeps.com (15.18%), ancestry.com (13.94%), ancstry.me (6.67%)	ancstry.me (74.11%), youtube.com (3.27%), ancestry.com.au (2.88%)
Counsyl	techcrunch.com (8.42%), businesswire.com (5.30%), bioportfolio.com (4.46%)	businesswire.com (14.78%), counsyl.com (13.91%), medium.com (5.21%)
DNAFit	fb.me (15.81%), instagram.com (14.65%), dnafit.com (2.99%)	fb.me (11.74%), dnafit.com (10.52%), dnafit.gr (2.83%)
FamilyTreeDNA	familytreedna.com (11.31%), myfamilydnatest.com (4.28%), fb.me (4.17%)	familytreedna.com (76.56%), abcn.ws (3.12%), instagram.com (1.56%)
FitnessGenes	instagram.com (14.77%), fitnessgenes.com (8.48%), workinstartups.com (6.29%)	fitnessgenes.com (31.11%), instagram.com (4.44%), pinterest.com (4.44%)
MapMyGenome	yourstory.com (11.84%), owler.us (11.44%), mapmygenome.in (9.18%)	mapmygenome.in (42.12%), youtu.be (14.35%), indiatimes.com (3.70%)
PathwayGenomics	paper.li (11.96%), atjo.es (10.82%), pathway.com (3.31%)	pathway.com (23.07%), nxtbook.com (3.84%), drhoffman.com (3.84%)
Ubiome	techcrunch.com (9.30%), bioportfolio.com (4.83%), ubiomeblog.com (4.21%)	ubiomeblog.com (34.32%), igg.me (26.07%), ubiome.com (6.60%)
VeritasGenetics	veritasgenetics.com (10.97%), technologyreview.com (5.01%), buff.ly (2.30%)	veritasgenetics.com (75.67%), biospace.com (1.35%), statnews.com (1.35%)
Genomics England	genomicsengland.co.uk (33.85%), youtube.com (1.98%), buff.ly (1.64%)	genomicsengland.co.uk (98.03%), peoplehr.net (0.58%), campaign-archive1.com (0.21%)
Personalized Medicine	instagram.com (8.78%), myriad.com (2.54%), buff.ly (2.32%)	–
Precision Medicine	buff.ly (2.92%), instagram.com (2.27%), nih.gov (1.87%)	–
Baseline	instagram.com (4.18%), fb.me (3.44%), youtu.be (2.72%)	–

TABLE 5.4: The top 3 domains per keyword, without official accounts and only considering the official accounts.

while differentiating between tweets made by regular users and those by official accounts. We also quantify the percentage of tweets with at least one hashtag (WH) and that of tweets including the keyword as a hashtag (KH), e.g., #23andMe.

We find a few unexpected hashtags among the DTC tweets, e.g., #sweepstakes (AncestryDNA), #startup (Fitnessgenes), #vote (Ubiome), #shechat, and #appguesswho (MapMyGenome). AncestryDNA's top hashtag, #sweepstakes (12%), is related to a marketing campaign promoting a TV series, "America: Promised Land." There are 3.5K tweets, from distinct users, with the very same content (most likely due to a "share" button): "I believe I've discovered my @ancestry! Discover yours for the chance to win an AncestryDNA Kit. #sweepstakes journeythroughhistorysweeps.com." We also find hashtags like #feistyfrugal and #holidaygiftguide in the AncestryDNA top 10 hashtags, which confirms how AncestryDNA uses Twitter for relatively aggressive marketing campaigns. Moreover, in the Fitnessgenes tweets, we find hashtags like #startup, #london, and #job due to a number of tweets advertising jobs for Fitnessgenes, while #shechat appears in tweets linking to an article related to women in business about MapMyGenome's founder.

By contrast, top hashtags for official accounts' tweets are closer to their main expertise/business. Similarly, those for genomics initiatives are pretty much always related to genetic testing, and this is actually consistent besides top 3. (The top 10 hashtags include, e.g., #digitalhealth, #genetics, and #lifestylemedicine). Finally, the percentage of tweets with the keyword appearing as a hashtag (KH), range from 12% for 23andMe to 25% for AncestryDNA even when excluding official accounts, which might be the by-product of promotion campaigns. When looking at tweets by official accounts KH values go up for some companies, e.g., AncestryDNA heavily promotes their brand using hashtags (46% KH).

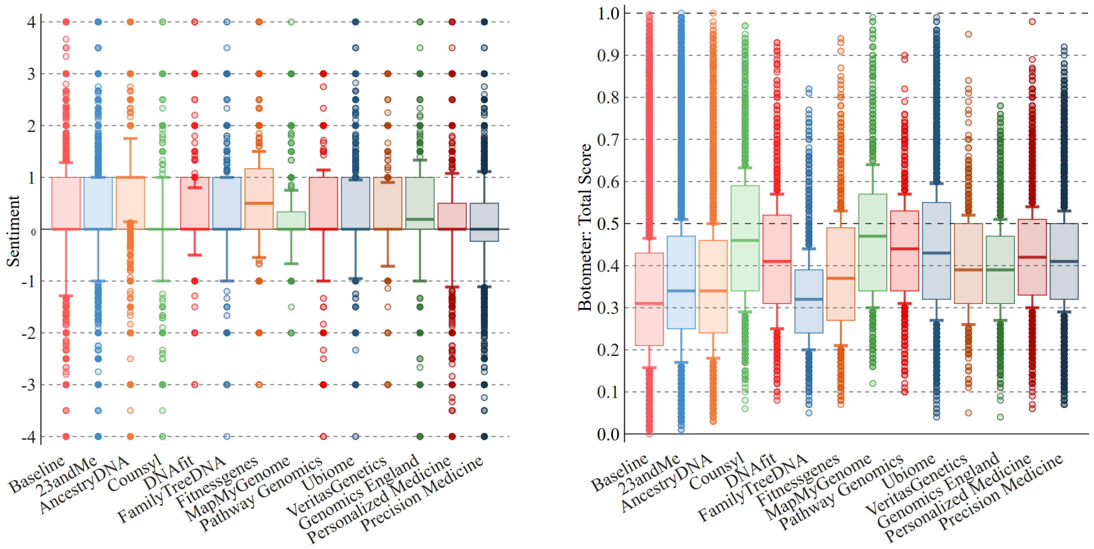
URL Analysis. We also analyze the URLs contained in the tweets of our dataset. Recall that the ratio of tweets containing URLs, as well as the percentage of those in the Alexa top 1M domains, are reported in Table 5.1. Once again, we distinguish between tweets from the official accounts and report the top 3 (top-level) domains per keyword in Table 5.4. If we discover URL shortener services in our dataset (e.g., `bit.ly`, `goo.gl`, `TinyURL`, `ow.ly`) we "unshorten" the URLs and use those in our analysis instead. We also note that all reported cases of `fb.me` lead to Facebook posts.

Among the top URLs shared by the official accounts, we find, unsurprisingly, their websites, as well as others leading to other domains owned by them, e.g., `23me.co`, `ancestry.com.au`, and `ancstry.me`. A few companies also promote news articles about them or related topics, e.g., top domains for Counsyl and MapMyGenome include `businesswire.com` and `indiatimes.com`, while DNAfit seems more focused on social media with its top domain being Facebook. As discussed previously, the domain `journeythroughhistorysweeps.com` appears frequently in AncestryDNA tweets. Then, note that `techcrunch.com`, a blog about technology, appears several times, as it often covers news and stories about genetic testing. We also highlight the presence of `owler.us`, an analytics/marketing provider sometimes labeled as potentially harmful by Twitter, as one of the top domains for MapMyGenome. Finally, for genomics initiatives, we notice `buff.ly`, a social media manager, suggesting that interested users appear to be extensively scheduling posts, thus potentially being more tech-savvy.

Sentiment Analysis We perform sentiment analysis using SentiStrength [212], which is designed to work on short texts. The tool outputs two scores, one positive, in $[1, 5]$, and one negative, in $[-1, -5]$. We calculate the sum value of the positive+negative scores for every tweet, then, collect *all* tweets with that keyword from the *same* user, and output the mean sentiment score.

In Figure 5.3(a), we report the distribution of sentiment across the different keywords. The vast majority of tweets have neutral sentiment, ranging from 0 to 1 scores. We run pair-wise two-sample Kolmogorov-Smirnov tests on the distributions, and in most cases reject the null hypothesis that they come from a common distribution at $\alpha = 0.05$. However, we are *unable* to reject the null hypothesis when comparing the baseline dataset to the PathwayGenomics dataset ($p = 0.77$) and when comparing DNAfit to Ubiome ($p = 0.34$). In general, the genomics initiatives, and in particular Personalized Medicine and Precision Medicine, have many outliers compared to most DTC genetic companies, suggesting more users who reveal strong feelings for or against these concepts. Genomics England, however, has a median above zero, indicating generally positive sentiment. Tweets about Counsyl are very neutral, while Ubiome tweets seem to be the most positive.

Discussion Our content analysis yields a few interesting findings. A large part of the genetic testing discourse appears to be generated from news and technology websites, and from tech-savvy users who rely on services to schedule social media posts. Also,



(a) Sentiment Analysis

(b) Botometer

FIGURE 5.3: Sentiment and Botometer scores of the keyword dataset.

sentiment around DTC companies is overall neutral, but positive for the genomics initiatives, however, tweets about DTC companies include a lot of strongly opinionated users (both positive and negative); we further explore tweets with high negative score in Section 5.4. Finally, tweets related to genetic testing not only contain a significantly higher number of hashtags than a random baseline, but they are also used for promotion. In general, we find several social media marketing strategies at play, with some companies employing traditional giveaways, others promoting mainly third-party articles about the company, and others focusing their efforts across multiple social media platforms. For instance, AncestryDNA is quite active in this context, with one particular hashtag (#sweepstakes) found in 1 out of 8 AncestryDNA tweets. This has a significant impact on how "regular" users engage in tweeting about genetic testing, which we further analyze next.

5.3.3 Who Tweets About Genetic Testing?

In this subsection, we shed light on the accounts tweeting about genetic testing. After a general characterization of the profiles, we look for the presence of social bots [219]. Then, we select a random sample of users tweeting about the two most popular DTC companies and analyze their latest 1,000 tweets to understand their interests. We start by analyzing the profiles tweeting about genetic testing: in Figure 5.4, we plot the distribution of the number of their followers, following, likes, and tweets.

Followers. Accounts tweeting about genomics initiatives have a median number of followers similar to baseline, while for the DTC companies the median is always lower, except for Counsyl, MapMyGenome, PathwayGenomics, and VeritasGenetics (see Figure 5.4(a)). Also considering that, for these four companies, there is a relatively low number of unique users (see Table 5.1), we believe accounts tweeting about them are fewer but more "popular." There are fewer outliers than the baseline, which is not surprising since we do not expect many mainstream accounts to tweet about genetic testing. Some outliers appear for 23andMe and AncestryDNA, which, upon manual examination, turn out to be Twitter accounts of newspapers or known technology websites, reflecting how the two most popular companies also get more press coverage.

Following. The median number of 'following' (i.e., the accounts followed by the users in our dataset) is usually higher than baseline for DTC companies but similar for genomics initiatives (Figure 5.4(b)). This suggests that users interested in DTC genetic testing might want to get more information off Twitter and/or from more accounts.

Likes. We then measure the number of tweets each profile has liked (Figure 5.4(c)). This measure, along with the number of tweets, depicts, to a certain extent degree, a level of engagement. We find that, for all keywords, profiles like fewer tweets than baseline users. There is one interesting outlier for 23andMe (@littlebytesnews), who liked more than 1M tweets; this is likely to be a bot, as also confirmed by Botometer [219]. Also, FamilyTreeDNA appears to have users liking more tweets than others. However, these accounts appear not to be bots, as we discuss later.

Tweets. We also quantify the number of tweets each account posts (Figure 5.4(d)). As with the number of likes, users in our datasets are less "active" than baseline users. There are interesting outliers above 1M tweets, which are due to social bots. We also find more tweets from Counsyl's users, seemingly mostly due to a large number of profiles describing themselves as "promoters" of science/digital life, technology enthusiasts, and/or influencers. Finally, users tweeting about genomics initiatives appear to be even less active, with a lower median value of tweets than the rest. Also considering that these users tweet more about the same keyword (as discussed in Section 5.3.1) but follow more accounts, we believe that they are more *passive* than the average Twitter user, using Twitter to get information but actively engaging less than others.

Geographic Distribution. We then estimate the geographic distribution of the users via the location field in their profile. Note that i) the corpus is biased as the used keywords are in English (e.g. Precision/Personalized Medicine), and ii) the location is self-reported, and users use it in different ways, adding their city (e.g., Miami), state (e.g., Florida), and/or country (e.g., USA). In some cases, entries might be empty (7.5% of the tweets in our dataset), ambiguous (e.g., Paris, France vs Paris, Texas), or fictitious (e.g., "Hell"). Nevertheless, as done in previous work [147], we use this field to estimate where most of the tweets are coming from. We use the Google Maps Geolocation API, which allows us to derive the country from a text containing a location.[5] The API returns an error for 6.6% of the profiles, mostly due to fictitious locations.

We find that the top 5 countries in our dataset are mostly English-speaking ones: 69.1% of all profiles with a valid location are from the US, followed by the UK (8.6%),

[5] https://developers.google.com/maps/documentation/geolocation

Canada (4.5%), India (2.1%), and Australia (1.4%). We then *normalize* using Internet-using population estimates [122], and plot the resulting heatmap, with the top 50 countries, in Figure 5.5. The maximum value is obtained by the US (i.e., 0.000254 users per Internet user), with 72.8K unique users, out of an estimated Internet population of 286M, posting tweets in our dataset. This suggest that US users dominate the conversation on genetic testing on Twitter.

We also perform a geolocation analysis broken down to specific keywords. Unsurprisingly, the top country of origin for Genomics England is the UK, as it is for DNAfit, which is based in London. Similarly, the top country for India-based company MapMyGenome tweets is India. Overall, we find that tweet numbers are in line with the countries where the DTC companies are based or operate – e.g., 23andMe health reports are available in US, Canada, and UK, while AncestryDNA also operates in Australia – as well as where the genomics initiatives are taking place.

Social Bot Analysis Next, we investigate the presence of social bots in our datasets, using the Botometer (`botometer.iuni.iu.edu`), a tool that, given a Twitter handle, returns the probability of it being a "social bot," i.e., an account controlled by software, algorithmically generating content and establishing interactions [219].

In Figure 5.3(b), we plot the distribution of Botometer scores for all keywords. We compare the distributions using pairwise 2 sample KS tests, and reject the null hypothesis at $\alpha = 0.05$ for all datasets *except* Counsyl and MapMyGenome ($p = 0.29$), DNAfit and VeritasGenetics ($p = 0.17$) and PrecisionMedicine and VeritasGenetics ($p = 0.10$). We also find that all median scores are higher than the baseline (between 0.35 and 0.5 vs 0.3). This is not entirely surprising since we expect many blogs, magazines, and news services covering genetic testing, and these are likely to get higher scores than individuals since they likely automate their activities. However, about 80% of the accounts in our dataset have scores lower than 0.5 and 90% lower than 0.6 (i.e., it is unlikely they are bots). We also find the two most popular keywords, 23andMe and AncestryDNA, as well as FamilyTreeDNA, somewhat stand out: accounts tweeting about them get the lowest Botometer scores. Although for FamilyTreeDNA this might be an artifact of the relatively low number of tweets (2K users), the scores suggest there might be more interaction/engagement from "real" individuals and/or fewer tweets by automated accounts about 23andMe and AncestryDNA.

We then look at accounts with Botometer scores *above* 0.7, finding that, for most DTC keywords, they account for 3–5% of the users; not too far from the baseline (2%) and the genomics initiatives (1.5–2%). Counsyl and MapMyGenome have more than 10% of users with scores above 0.7. We also quantify *how many* tweets are posted by (likely) social bots: almost 15% of all PathwayGenomics tweets come from users with score 0.7 or above (4.5% of all users), while for all other keywords social bots are not responsible for a substantially high number of tweets in our datasets.

5.3.4 What Do Users Tweet About Otherwise?

We then focus on the users tweeting about the two most popular companies – i.e., 23andMe and AncestryDNA – and study their last 1K tweets, aiming to understand the characteristics of the accounts who show interest in genetic testing. We only do so for 23andMe and AncestryDNA as these companies have the highest numbers of tweets and users, and thus, are more likely to lead to a representative and interesting sample.

Data. We select a random 20% sample of the users who have posted at least one tweet with keywords 23andMe/AncestryDNA (resp., 12.2K/64K and 3.3K/16.9K users) and, using the same methodology we describe in Section 5.2, we crawl their latest 1K tweets if their account is still active.[6] This yields a dataset of 12M tweets, outlined in Table 5.5. For comparison, we also get the last 1K tweets of a random sample of 5K users from

[6] We find 575 and 61 inactive accounts, respectively, for 23andMe and AncestryDNA.

	Tweets	**Users**	**RTs**	**Likes**	**Hashtags**	**URLs**	**Top 1M**
23andMe	9,534,302	12,227	9,077,066	3,501,053	24.40%	63.62%	81.43%
AncestryDNA	2,466,443	3,320	1,399,804	22,001,065	34.21%	63.64%	78.86%
Total	12,000,745	15,547	10,476,870	25,502,118	26.41%	63.62%	80.89%
Baseline	4,208,967	5,035	139,551,104	342,052,546	17.47%	41.24%	88.41%

TABLE 5.5: Summary of the users' tweets dataset, with last 1K tweets of a 20% sample of 23andMe and AncestryDNA users.

the keyword dataset's baseline users. Note that statistics in Table 5.5 refer to the latest 1K tweets of the user sample, while those in Table 5.1 to tweets with a given keyword.

The numbers of retweets and likes per tweet are, once again, lower than the baseline. However, users tweeting about AncestryDNA receive, for their last 1K tweets, one order of magnitude more likes than those tweeting about 23andMe. Moreover, we observe relatively high percentages of tweets with hashtags (63%) and URLs (around 80%). How far back in time the 1,000th tweet appears varies across users, depending on how often they tweet. We measure the time between the most recent and the 1,000th tweet, and find that baseline users are more "active" than the users who have tweeted about 23andMe and AncestryDNA, in line with what discussed previously. In particular, AncestryDNA users appear to post less: for half of them, it takes at least 359 days to tweet 1K tweets compared to 260 for the baseline and 287 for 23andMe.

Hashtag analysis. We conduct a hashtag analysis on tweets in Table 5.5. In Table 5.6, we report the top 10 hashtags of the users' last 1K tweets. For 23andMe, we find several hashtags related to health in the top 10; also considering that the top 30 include #pharma, #cancer, and #biotech, it is likely that users who have shown interest in 23andMe are also very much interested in (digital) health, which is one of the primary aspects of 23andMe's business. This happens to a lesser extent for AncestryDNA results: while top hashtags include #genealogy (4th), they also include #giveaway, #sweepstakes, #win, #ad, #promotion, #perduecrew, and #contest, suggesting that these users are rather interested in promotional products. This is line with our earlier observation that AncestryDNA extensively uses advertising and marketing campaigns on

23andMe	AncestryDNA	Baseline
tech (1.07%)	giveaway (3.31%)	gameinsight (0.55%)
news (1.06%)	sweepstakes (2.01%)	trecru (0.34%)
health (0.58%)	win (2.01%)	btsbbmas (0.33%)
business (0.48%)	genealogy (1.01%)	nowplaying (0.30%)
healthcare (0.43%)	tech (0.63%)	android (0.28%)
digitalhealth (0.40%)	ad (0.51%)	androidgames (0.27%)
startup (0.39%)	entry (0.51%)	ipad (0.26%)
socialmedia (0.34%)	promotion (0.48%)	trump (0.24%)
viral (0.34%)	perduecrew (0.47%)	music (0.21%)
technology (0.34%)	contest (0.44%)	ipadgames (0.20%)

TABLE 5.6: The top 10 hashtags of the users' tweets dataset.

Twitter.

URL analysis. In Table 5.7, we report the top 5 domains of the three sets. Over the last 1K tweets, users tweeting about 23andMe and AncestryDNA share a substantial number of links to `techcrunch.com`, a popular technology website; i.e., users who have tweeted at least once about these companies have an interest about subjects related to new technologies. In fact, the top 10 list of 23andMe's set of tweets also include `lnkd.in`, `mashable.com`, and `entrepreneur.com`. For AncestryDNA, we find `wn.nr`, another website related to contests and sweeps. There are thousands of tweets like "Enter for a chance to win a $500 Gift Card! wn.nr/DRRrZq #MemorialDaySweeps #Entry". We also note the presence of `woobox.com`, a marketing campaign website, responsible for organizing giveaways, as well as `giveaway.amazon.com`, an Amazon site organizing promotional sweepstakes. Botometer scores indicate these accounts are not actually bots, hence this might be related to the fact that AncestryDNA, through their marketing campaigns, attract Twitter users who are generally active in looking for deals and sweeps.

5.4 Case Studies

In this section, we take a closer look at "negative" tweets, following the sentiment analysis methodology presented previously. We also investigate the presence of users who

23andMe	AncestryDNA	Baseline
fb.me (4.00%)	instagram.com (6.78%)	fb.me (5.85%)
instagram.com (3.06%)	fb.me (5.48%)	instagram.com (4.42%)
youtu.be (2.18%)	techcrunch.com (4.42%)	youtu.be (2.94%)
buff.ly (2.17%)	youtu.be (4.04%)	twittascope.com (0.58%)
techcrunch.com (1.53%)	wn.nr (1.79%)	tmblr.co (0.56%)
lnkd.in (1.02%)	woobox.com (1.51%)	buff.ly (0.54%)
mashable.com (0.65%)	giveaway.amazon.com (1.17%)	fllwrs.com (0.40%)
entrepreneur.com (0.63%)	buff.ly (1.08%)	gigam.es (0.33%)
nyti.ms (0.62%)	swee.ps (0.80%)	soundcloud.com (0.32%)
reddit.com (0.55%)	twittascope.com (0.41%)	vine.co (0.30%)

TABLE 5.7: The top 10 domains of the users' tweets dataset.

post their genetic test results.

5.4.1 Instances of Racism

We select all tweets with genetic testing keywords from users who yield a total sentiment score below -3, obtaining 3,605 tweets from 3,209 unique users. We then manually examine those with keywords 23andMe or AncestryDNA (1,725 and 167, respectively), and find several of them containing themes related to racism, hate, and privacy fears.

In particular, the "ethnic" breakdown provided by ancestry reports[7] seems to spur several instances of negative-sentiment tweets associated with racism and disapproval of multi-cultural/multi-ethnic values. For instance, a user with more than 3K followers self-describing as a "Yuge fan for Donald Trump", tweets: "Get this race mixing shit off my time line!!" (March 23, 2017) in response to a 23andMe video about ancestry. Another posts: "I wanna do that 23andMe so bad! I'm kinda scared what my results will be tho lmao I'm prob like half black tbh"(January 13, 2017), and gets a response: "I was too just do it and never tell anyone if you're a halfbreed haha". Also, a user identifying as 'American Fascist' tweets: "I'd like to get the @23andMe kit but, I'm worried about the results. Just my luck, I'd have non-white/kike ancestors. #UltimateBlackpill" (May 30, 2017).

[7] E.g., https://permalinks.23andMe.com/pdf/samplereport_ancestrycomp.pdf

Although we conduct an in-depth analysis of genetic testing related racism on the Web in Chapter 6, we attempt to assess whether it may be systematic on Twitter, e.g., appearing also in tweets not scored as negative by SentiStrength. To this end, we search for the presence of hateful words using the hatebase.org dictionary, a crowd-sourced list of 1K terms that indicate hate when referring to a third person, removing words that are ambiguous or context-sensitive, as done by previous work [108]. Naturally, this is a best-effort approach since hateful terms might be used in non-hateful contexts (e.g., to refer to oneself), or, conversely, racist behavior can occur without hate words. Also, Twitter might be removing tweets with hate words as claimed in their hateful conduct policy.[8] Nonetheless, we do find instances of hate speech, e.g., anti-semitic tweets such as: "as long as there are khazar milkers to cause people to demand my 23andMe results, i will always be here to shitpost" (November 19, 2016), or "@*** i would be pleased if you posted your 23andMe so i can confirm your khazar milkers are indeed genuine" (December 23, 2016).

Note that "Khazar milkers" refers to an anti-semitic theory on the origin of Jewish people from the 1900s [79]. In a nutshell, it posits that Ashkenazi Jews are not descendant from Israelites, but from a tribe of Turkic origin that converted to Judaism. 23andMe issued ancestry reports that suggested Ashkenazi Jews in a given haplogroup were descendant from a single Khazarian ancestor. Understanding the ancestry of Jewish people has been of interest to the genetics community for years, and the Khazar theory has been refuted repeatedly [22]. Nonetheless, the alt-right has exploited it to corroborate their anti-semitic beliefs [183], and incorporate it into their collection of misleading/factually incorrect talking points. In particular, "khazar milkers" was allegedly coined by the "@***" user mentioned above, and is used to imply a sort of succubus quality of Jewish women.

[8] https://support.twitter.com/articles/20175050

5.4.2 Privacy Concerns

We also identify, among the most negative tweets, themes related to fears of privacy violation and data misuse. Examples include "Is it me? Does the idea of #23andMe seem a bit sinister? Do they keep the results? Who owns the results? Who owns 23andMe?"(January 1, 2016), "Same thing with 23andMe and similar companies. Indefinitely stored data with possible sinister future uses? #blackmirror"(November 13, 2016), and "Why does this scare the hell out of me? How can our privacy ever be assured?" (February 27, 2016). Searching for 'privacy' and 'private' in our keyword dataset returns 1,991 tweets, mostly from 23andMe and Precision Medicine (1.1K and 625, respectively), which we proceed to examine both manually and from a temporal point of view (i.e., measuring daily volumes). Overall, we find that privacy in the context of genetic testing appears to be a theme discussed recurrently on social media and a concern far from being addressed. This is not entirely unexpected, considering that both the DTC market and the genomics landscape are evolving relatively fast, with regulation and understanding of data protection as well as informed consent often lagging behind, as also highlighted in prior work [149, 64, 178].

Interestingly, one of the peaks in tweets related to 23andMe and privacy occurs on October 19, 2015 (with 152 tweets). As discussed in Section 5.3.2, this a relevant date with regards to the FDA revoking their approval for 23andMe's health reports, which yields a peak in 23andMe tweets overall. However, the FDA ruling had nothing to do with privacy, yet, it put 23andMe in the spotlight, possibly causing privacy concerns to resurface. In fact, privacy and 23andMe discussions periodically appear in our dataset, even beyond tweets with negative sentiment, e.g., "I want to do #23andMe but don't want a private company owning my genetic data. Anyone heard of any hacks to do it anonymously?" (July 13, 2017), "@23andMe ur privacy policy describes how there is no privacy. How about u not share any data at all. I pay u and u send the results. Period" (December 8, 2015), "Should we be concerned about data collection and privacy with direct to consumer DNA testing companies like 23andMe?" (April 19, 2017).

5.4.3 Users Sharing Test Results

Finally, we investigate the presence of users who post their genetic test results, aiming to estimate their number and shed light on their profiles. Given their popularity, we only do so for 23andMe and AncestryDNA. Among other things, we believe this is important because health/ancestry reports may contain sensitive information about the individuals taking these tests, including their predisposition to diseases and their ethnic heritage [13].

Methodology. Finding all tweets that may include test results is difficult, and arguably out of scope, thus, we focus on *screenshots* of genetic test results as we anecdotally find a non-negligible number in our dataset. These are almost exclusively ancestry results, even though 23andMe also provides health reports. We start from the 4.5K/1.5K 23andMe/AncestryDNA tweets with images, but, since they are too many to be all manually examined, we use the following methodology. First, we build two sets with 100 "ground truth" images with screenshots of ancestry test results, one each for 23andMe and AncestryDNA, then, we use Perceptual Hashing [137] to find similar images that are likely screenshots of test results too, and manually check them to exclude false positives.[9]

Overall, our approach likely yields a conservative estimate, nonetheless, it constitutes a best-effort approach to identify and analyze tweets including test results. We obtain 366 and 204 images for, respectively, 23andMe and AncestryDNA. Upon manual examination, we find and remove 58 (16%) and 26 (13%) false positives. Thus, we estimate a lower bound of 0.23% and 0.60% of 23andMe and AncestryDNA tweets containing ancestry test screenshots (and 3.40% and 5.15% of tweets with pictures).

[9] Perceptual hash functions extract features from multimedia content and calculate hash values based on them. They can be used to compare two objects by calculating a distance/similarity score between two hash values; the objects are labeled as (perceptually) equal if the distance is below a chosen threshold [237]. We set the pHash distant threshold to 17 since it produces the best results in our setting.

Tweets content. We manually examine the 486 tweets with screenshots of ancestry test results, finding that users often appear to be somewhat enthusiastic about their experience. In some cases, we note a feeling of "relief", sometimes expressed in a humorous way, when the results show they are predominantly "white": about 10% of tweets with screenshots include the word white. Examples include: "23andMe confirms: I'm super white", "Got my @23andMe results back today. I'm super white. Lie, rice on a paper plate with a glass of milk in a snowstorm, white."

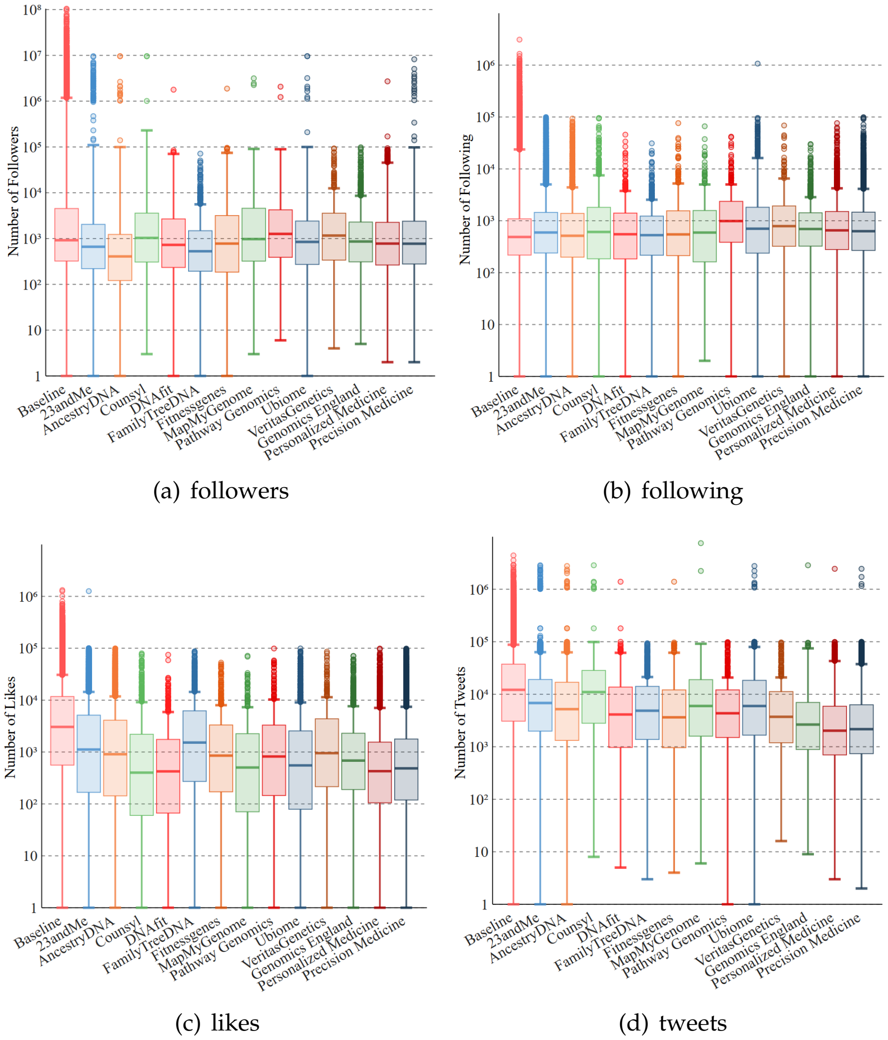
User Analysis. We also crawl the last 1,000 tweets of the 308 users who have posted screenshots with test results, and analyze them as done for the random sample discussed in Section 5.3.3. We find that their most commonly used hashtags are indeed related to genetic testing, confirming that the users who do genetic tests are actually generally interested in the subject. It is also interesting to find #maga to be the second most common hashtag for users who post 23andMe results (appearing in 431 tweets). Then, looking at the top domains shared by these users, we do not observe surprising difference compared to those from a random sample users (see Table 5.7), although we find more social media services like Instagram and Facebook.

5.5 Discussion

We presented a large-scale analysis of Twitter discourse related to genetic testing. We examined more than 300K tweets containing 13 relevant keywords as well as 12M tweets posted by more than 100K accounts that have shown interest in genetic testing. We found that the discourse related to genetic testing is often influenced by news and technology websites, and by a group of tech-savvy users who are overall interested in tech and digital health. Overall, users tweeting about genetic testing are mostly in the US and other English-speaking countries, while we do not find evidence of extensive influence of social bots.

However, the broad conversation seems to be dominated by users that might have a vested interest in its success, e.g., specialist journalists, medical professionals, entrepreneurs, etc. This is particularly evident in the tweets related to genomics initiatives, which are mainly discussed by highly engaged users and which are influenced, at least in terms of volume, by important announcements. Moreover, we noticed that the two most popular DTC companies, 23andMe and AncestryDNA, also generate the most tweets, however, although 23andMe has half the customers, it produces almost 5 times more tweets, also due to controversy around their failure to get FDA approval in 2015. We also observed a clear distinction in the marketing efforts undertaken by different companies, which ends up influencing users' engagement on Twitter.

Our work is particularly timely as genetic testing and genomics initiatives are increasingly often associated to ethical, legal, and societal concerns [87]. In this context, our analysis sheds light not only on who tweets about genetic testing, what they talk about, and how they use Twitter, but also on groups utilizing genetic testing to push racist agendas and users expressing privacy concerns. We also found a number of enthusiastic users who broadcast their test results through screenshots notwithstanding possible privacy implications.



(a) followers

(b) following

(c) likes

(d) tweets

FIGURE 5.4: Boxplots with statistics per user profile (note the log-scale in y-axis).

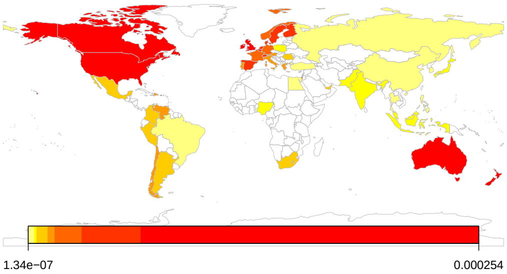


FIGURE 5.5: Geolocation of Twitter profiles, normalized by Internet using population per country.

Chapter 6

Measuring the Relation between Genetic Testing and Racism on Reddit and 4chan

6.1 Introduction

Over the past decade, researchers have made tremendous progress toward understanding the human genome, i.e., the complete set of an individual's DNA, which encodes all of the information needed to build and maintain that organism. With increasingly low costs, millions of people can afford to learn about their genetic make-up, not only in diagnostic settings, but also to satisfy their curiosity about traits, wellness, or discover their ancestry and genealogy.

A number of companies have successfully marketed *direct-to-consumer (DTC)* genetic tests: individuals purchase a kit (typically around $100), mail it back with a saliva sample, and receive online reports after a few days. DTC companies offer a wide range of services, from romantic match-making [71] or identification of athletic skills [201] to reports of health risks (e.g, likelihood of developing Parkinson's), wellness (e.g., lactose intolerance), carrier status (e.g., hereditary hearing loss), traits (e.g., eyes color), etc.

Popular products also include genetic *ancestry* tests, which promise a way to discover one's ancestral roots, building on patterns of genetic variations common in people from similar backgrounds [165]. However, these are subject to limitations, e.g., results differ from provider to provider due to different control groups [193]. AncestryDNA alone has tested more than 16M customers as of June 2020 [7].

Alas, increased popularity of self-administered genetic tests, and in particular ancestry, has also been accompanied by media reports of far-right groups using it to attack minorities or prove their genetic "purity" [32, 184]. This prompts concerns of a new wave of scientific racism [187]. For example, white nationalists have been taped chugging milk at gatherings to demonstrate the ability of people of white color to better digest lactose [104]. Also, statements from USA President Donald Trump led Senator Warren to publicly confirm her Native American ancestry via genetic testing [144].

Interest in DTC genetic testing by right-wing communities comes at a time when racism, hate, and antisemitism on platforms like 4chan, Gab, and certain communities on Reddit is on the rise [108, 81]. Thus, these trends are particularly worrying, also considering how technology has been disrupting society in previously unconsidered ways [97]; the fact that racist, misogynistic, and dangerous behavior festers and spreads on the Web at an unprecedented scale, eventually making its way into the real world, prompts the need for a thorough understanding of how these genetic testing tools are being used and misusedused in online discussions. As genetics-based arguments for discrimination [20], and even genocide (e.g., the Holocaust), have been made in the past, *without* hard, statistical data to back them up, the potential abuse of genetic testing results this should not be overlooked.

While other aspects of genetic testing have been studied (e.g., how they affect one's perception of racial identity [172, 192]), we are interested in the relation between genetic testing and online hate. This is a topic that has not been thoroughly studied by the scientific community, despite, as discussed earlier, increasingly worrisome indications of far-right groups exploiting genetic testing for racist rhetoric. With this motivation in mind, we identify and address the following research questions: (1) What is the overall prevalence of genetic testing discourse on social networks like Reddit and 4chan? (2) In what context do users discuss genetic testing? (3) Is genetic testing associated with far-right views, racist ideologies, hate speech, and/or white supremacy? (4) If yes, in what context? Can we identify specific themes?

We compile and use a set of 280 keywords related to genetic testing to extract all available posts and comments from Reddit and 4chan. We collect 7K threads from the politically incorrect (/pol/) board of 4chan (consisting of 1.3M posts) from June 30, 2016 to March 13, 2018, and 77K comments from Reddit related to genetic testing from January 1, 2016 to March 31, 2018, and analyze them along several axes to understand how genetic testing is being discussed online. We rely on natural language processing, computer vision, and machine learning tools, including (i) Latent Dirichlet Allocation (LDA) [29] to identify topics of discussion, (ii) word embeddings [151] to uncover words used in a similar context across datasets, (iii) Google's Perspective API [175] to measure toxicity in texts, and (iv) Perceptual Hashing to assess the imagery and memes shared in posts.

Overall, the *main* findings of our study include:

1. Genetic testing is often discussed on /pol/ and on subreddits associated with hateful, racist, and sexist content. These communities discuss genetic testing in a highly toxic manner, often suggesting its use to marginalize or even *eliminate* minorities.
2. Our image analysis on /pol/ shows the recurrent presence of popular alt-right personalities and "popular" antisemitic memes along with genetic testing discussions.
3. Word embeddings analysis reveals that certain subreddits use ethnic terms in conjunction with genetic testing keywords in the same way as /pol/, which may be

an indicator of 4chan's fringe ideologies spilling out on more mainstream Web communities.

4. Genetic testing on Reddit is being discussed in a variety of contexts, e.g., dog breeds, debating the validity of evidence in real crimes crime evidence, and issues related to children (e.g., adoption, pregnancy), among others. This indicates how mainstream genetic testing has recently become.
5. Reddit users are not uniformly interested in all aspects of genetic testing, rather, they form groups ranging from enthusiasts (i.e., people who are interested in or have undergone genetic testing), to people who use genetic keywords exclusively in subreddits that discuss fringe political views.

6.2 Datasets

In this section, we present the methodology used to obtain the datasets used in our study.

Genetic Testing Keywords. To extract relevant comments and posts we compile a list of 280 keywords related to genetic testing. First, we use the list of 268 DTC companies offering DNA tests over the Internet between 2011 and 2018 (e.g., 23andme, AncestryDNA, Orig3n) obtained from [177]. We then add 12 more keywords: ancestry testing/test, genetic testing/test, genomic testing/test, genomics, genealogy testing/test, dna testing/test, and GEDMatch (an open data personal genomics database and genealogy website [89]).

Reddit Dataset. We gather all Reddit comments from January 1, 2016 to March 31, 2018 (2B comments in 473K subreddits) via the publicly available monthly releases of `pushshift.io`.[1] We then use the 280 genetic testing keywords as search terms to extract

[1] `https://files.pushshift.io/reddit/`

Reddit	Genetic Testing	Random	*4chan*	Genetic Testing	Random
Comments	77,184	204,713	Threads	6,986	19,530
Subreddits	3,734	12,616	Posts	1,306,671	760,691
Users	48,096	165,127	Posts/T (Mean)	186.5	37.9
			Posts/T (Median)	183	5
			Images	338,540	206,830

TABLE 6.1: Overview of the Reddit and 4chan datasets.

all comments possibly related to genetic testing. This results in a dataset of 77K comments posted in 4.6K subreddits, as summarized in Table 6.1. For comparison, we obtain a 0.0001% subset of all Reddit comments posted between January 1, 2016 to March 31, 2018. To do so, we use a simple Python script that randomly selects 0.0001% of all comments for each month of the same time period (recall that all Reddit comments are publicly available by `pushshift.io`). This results in a set of 204K random comments unrelated to genetic testing.

4chan Dataset. We focus on 4chan's politically incorrect board (/pol/), which has been shown to include a high volume of racist, xenophobic, and hateful content [108]. We choose /pol/ as we study how genetic testing is being discussed in communities that have been associated with alt-right ideologies. We collect 1.9M threads posted on /pol/ from June 30, 2016 to March 13, 2018. Once again, we use the 280 keywords as search terms on each thread: if we find a keyword anywhere in it, we get the *whole thread*. This is slightly different from what we do for Reddit. On 4chan, each discussion is structured as a single-threaded entity where the OP submits an image on which other users respond. There is no official method of responding to a certain comment other than the original one, whereas, on Reddit a user may reply to a specific comment creating a new branch of answers. Also, 4chan threads do not contain titles, thus, it is difficult to understand the context of a discussion without reading the whole thread. In the end, we extract 6.9K threads containing 1.3M posts. For comparison, we also get a random sample of 19K threads, with 760K posts. The 4chan dataset is summarized in Table 6.1, where we report the mean and median number of posts per thread, and the total number of images. Note that, while the threads with genetic testing keywords have 338,540 images, later on we study only images shared in the *posts* containing those keywords (6,375).

Remarks. We look at Reddit and 4chan's politically incorrect board (/pol/) as opposed to mainstream platforms (i.e., Facebook, Twitter) for several reasons. First, Facebook's structure and API do not permit the uniform collection of all available comments containing specific keywords, thus we would have to cherry-pick specific groups possibly leading to a biased dataset, while Twitter's API only allows the collection of 1% of all tweets which would greatly limit the dataset's amplitude. Meanwhile, Reddit's self-organizing content (i.e., subreddits) permits us to easily extract the context in which genetic testing is being discussed, while *all* comments ever posted in the platform are available online (see Section Datasets). Moreover, we are interested in the hateful and racist connotations of genetic testing discourse.

6.3 Genetic Testing Discussions on Reddit

In this section, we study the prevalence of genetic testing comments on Reddit. We start by identifying the subreddits with the highest number of comments related to genetic testing and thematically grouping them. Then, we use Google's Perspective API [175], a publicly available tool geared to identify toxic comments, to measure the toxicity of each group. We also use Latent Dirichlet Allocation (LDA) for basic topic modeling, aiming to extract the most prominent topics of discussion for each group. Finally, we examine comments in which users express privacy concerns.

	Subreddit	Gen Test Comms	Total Comms	Percent.	Tag
1	r/promethease	347	2,580	13%	Genetics
2	r/SNPedia	184	1,774	10%	Genetics
3	r/23andme	4,150	44,225	9%	Genetics
4	r/Ancestry	190	2,793	6%	Ancestry
5	r/Genealogy	3569	95,205	3%	Ancestry
6	r/genetics	347	11,741	2%	Genetics
7	r/Adoption	610	40,667	1%	Children
8	r/IDmydog	175	14,429	1%	Animals
9	r/ehlersdanlos	340	47,303	0.7%	Health
10	r/TheBlackList	288	43,127	0.6%	Entertainment
11	r/Celiac	171	41,444	0.4%	Health
12	r/Testosterone	306	83,997	0.3%	Health
13	r/serialpodcast	745	213,958	0.3%	Entertainment
14	r/EARONS	155	48,613	0.3%	Crime
15	r/StevenAveryIsGuilty	357	126,689	0.2%	Crime
16	r/cancer	172	68,037	0.2%	Health
17	r/dogs	1,627	803,094	0.2%	Animals
18	r/MakingaMurderer	1,198	624,641	0.1%	Crime
19	r/SuperMaM	139	73,997	0.1%	Crime
20	r/Nootropics	613	331,434	0.1%	Drugs
21	r/DebateAltRight	298	169,354	0.1%	Hate
22	r/AugustBumpers2017	120	71,825	0.1%	Children
23	r/ttcafterloss	223	141,992	0.1%	Children
24	r/UnresolvedMysteries	966	667,940	0.1%	Crime
25	r/InfertilityBabies	156	111,862	0.1%	Children
26	r/BeforeNAfterAdoption	107	81,078	0.1%	Animals
27	r/TickTockManitowoc	443	364,725	0.1%	Crime
28	r/pitbulls	108	103,844	0.1%	Animals
29	r/infertility	427	423,863	0.1%	Children
30	r/arabs	128	157,054	0.08%	Race/Countries
31	r/BabyBumps	973	130,1608	0.07%	Children
32	r/altright	108	166,436	0.06%	Hate
33	r/Judaism	178	299,667	0.06%	Race/Countries
34	r/AskDocs	193	385,831	0.05%	Health
35	r/TryingForABaby	192	411,263	0.04%	Children
36	r/slatestarcodex	123	273,357	0.04%	Science
37	r/bipolar	164	396,899	0.04%	Health
38	r/MensRights	399	993,039	0.04%	Sexes
39	r/bestoflegaladvice	144	362,868	0.03%	Legal
40	r/steroids	320	825,647	0.03%	Drugs
41	r/legaladvice	1,081	2,851,210	0.03%	Legal
42	r/hapas	128	368,467	0.03%	Race/Countries
43	r/science	782	2,666,213	0.03%	Science
44	r/ADHD	168	576,203	0.03%	Health
45	r/changemyview	538	1,908,120	0.02%	Other
46	r/TheRedPill	270	1,044,079	0.02%	Hate
47	r/confession	182	710,132	0.02%	Other
48	r/teenmom	203	824,312	0.02%	Entertainment
49	r/TeenMomOGandTeenMom2	133	565,612	0.02%	Entertainment
50	r/Parenting	194	829,177	0.02%	Children
51	r/childfree	350	1,531,152	0.02%	Children
52	r/MGTOW	365	1,625,881	0.02%	Hate
53	r/relationship_advice	309	1,383,111	0.02%	Other
54	r/ShitAmericansSay	122	547,506	0.02%	Comedy
55	r/relationships	1,853	8,538,031	0.02%	Other
56	r/JUSTNOMIL	359	1,790,725	0.02%	Other
57	r/TrueReddit	102	557,598	0.01%	Other
58	r/tifu	390	2,191,142	0.01%	Other
59	r/TwoXChromosomes	488	2,753,369	0.01%	Sexes
60	r/breakingmom	101	609,366	0.01%	Children
61	r/Advice	157	1,021,798	0.01%	Other
62	r/PurplePillDebate	210	1,421,805	0.01%	Hate
63	r/aww	799	5,671,423	0.01%	Other
64	r/history	142	1,054,177	0.01%	Other
65	r/raisedbynarcissists	163	1,214,553	0.01%	Other
66	r/milliondollarextreme	109	895,032	0.01%	Hate
67	r/asktransgender	158	1,307,753	0.01%	Sexes
68	r/exmormon	288	2,444,535	0.01%	Religion
69	r/nottheonion	313	2,898,542	0.01%	News
70	r/MapPorn	114	1,063,518	0.01%	Other
71	r/explainlikeimfive	388	3,741,174	0.01%	Educational
72	r/Futurology	278	2,689,784	0.01%	Science
73	r/NoStupidQuestions	198	1,943,855	0.01%	Educational
74	r/AskWomen	324	3,328,046	<0.01%	Sexes
75	r/UpliftingNews	114	1,214,761	<0.01%	News
76	r/Documentaries	130	1,386,157	<0.01%	Educational
77	r/todayilearned	1,185	13,088,194	<0.01%	Educational
78	r/conspiracy	469	5,281,831	<0.01%	Other
79	r/news	1,717	19,386,087	<0.01%	News
80	r/ireland	138	1,615,105	<0.01%	Race/Countries
81	r/TumblrInAction	216	2,563,058	<0.01%	Hate
82	r/depression	103	1,277,435	<0.01%	Health
83	r/askscience	101	1,289,247	<0.01%	Science
84	r/fatlogic	120	1,543,070	<0.01%	Hate
85	r/IAmA	242	3,521,706	<0.01%	Other
86	r/technology	268	4,072,195	<0.01%	Technology
87	r/AdviceAnimals	372	5,906,232	<0.01%	Other
88	r/Showerthoughts	477	8,034,239	<0.01%	Other
89	r/trashy	110	1,897,268	<0.01%	Funny
90	r/BlackPeopleTwitter	209	3,762,278	<0.01%	Hate
91	r/OldSchoolCool	142	2,593,419	<0.01%	Other
92	r/canada	231	4,341,997	<0.01%	Race/Countries
93	r/CringeAnarchy	217	4,101,269	<0.01%	Politics
94	r/AskMen	195	3,805,036	<0.01%	Sexes
95	r/The_Donald	1,251	28,360,073	<0.01%	Hate
96	r/worldnews	845	20,224,373	<0.01%	News
97	r/europe	219	5,275,810	<0.01%	Race/Countries
98	r/atheism	108	2,626,435	<0.01%	Religion
99	r/AskReddit	5,421	132,899,306	<0.01%	Other
100	r/india	127	3,141,858	<0.01%	Race/Countries
101	r/KotakuInAction	109	2,811,180	<0.01%	Hate
102	r/pics	543	15,528,294	<0.01%	Other
103	r/politics	1,517	46,270,193	<0.01%	Politics
104	r/personalfinance	150	4,671,327	<0.01%	Other
105	r/Philippines	102	3,245,641	<0.01%	Race/Countries
106	r/unitedkingdom	105	3,595,982	<0.01%	Race/Countries
107	r/ukpolitics	125	4,348,955	<0.01%	Race/Countries
108	r/trees	113	4,009,217	<0.01%	Drugs
109	r/WTF	131	5,609,346	<0.01%	Other
110	r/videos	319	13,934,560	<0.01%	Other
111	r/funny	321	15,792,122	<0.01%	Funny
112	r/gifs	111	9,032,723	<0.01%	Other
113	r/movies	111	11,810,334	<0.01%	Other
114	r/nba	183	23,109,676	<0.01%	Other

TABLE 6.2: List of subreddits sorted by normalized number of genetic testing comments.

6.3.1 Methodology

Subreddits selection & grouping. We extract all the subreddits where genetic testing comments have been posted to, but discard subreddits if they either have fewer than 1K comments overall or fewer than 100 comments with one of the keywords. This yields a list of 114 subreddits; see Table 6.2, which reports the normalized number of genetic testing related comments.

We group the subreddits into categories to study them based on (broad) discussion topics. We first turn to `redditlist.com`, a website reporting various subreddits metrics (e.g., number of subscribers, growth, etc.) and thematic tags, however, tags are available only for very popular subreddits, and most of the subreddits in our list do not have them. Thus, we have two annotators browse the subreddits and assign up to five tags based on their thematic content. We then create a dictionary based on all the tags, and pick one tag which represents each subreddit best according to the annotators' judgment (the tag is reported in Table 6.2). Finally, we group them based on this tag, which leads to 18 categories plus a generic one, labeled as "other" (which includes 25 subreddits). We report the subreddits in each category, except "other," in Figure 6.1.

Note that, while the content of most subreddits can be intuitively guessed from the name (e.g., /r/23andMe is about the company 23andMe), that is not always the case. For instance, /r/AdviceAnimals is not about advice on animals, but on humans, and /r/trees is a subreddit about marijuana. Also, we opt to assign a separate 'Ancestry' category rather than 'Genetics', since the former includes subreddits that do not necessarily deal with genetic testing.

Prevalence of genetic testing comments. Unsurprisingly, the top five subreddits with most genetic testing comments are directly related to genetic testing/ancestry. Subreddits like /r/SNPedia or /r/Ancestry have a high fraction of comments with at least one genetic testing keyword; respectively, 10% and 7%. We also find genetic testing to be

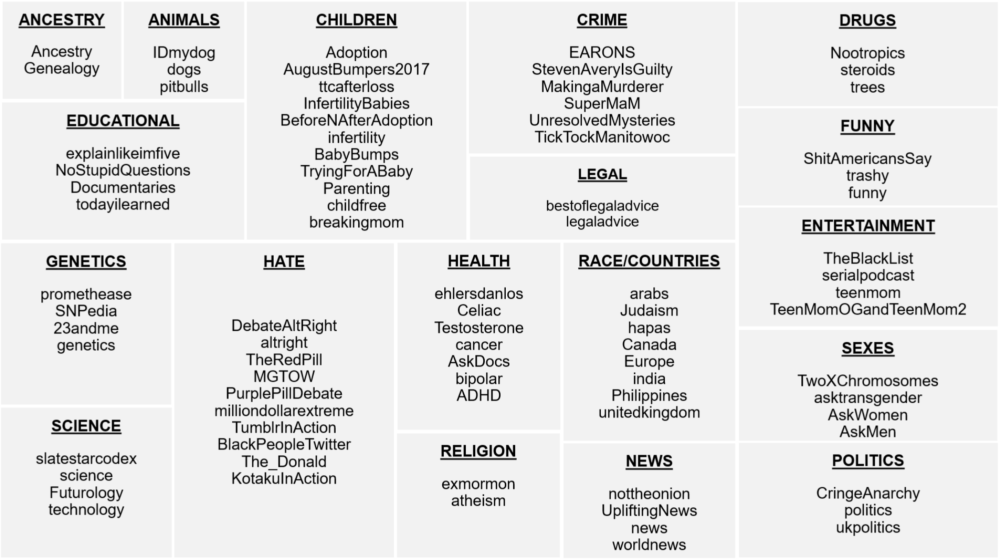
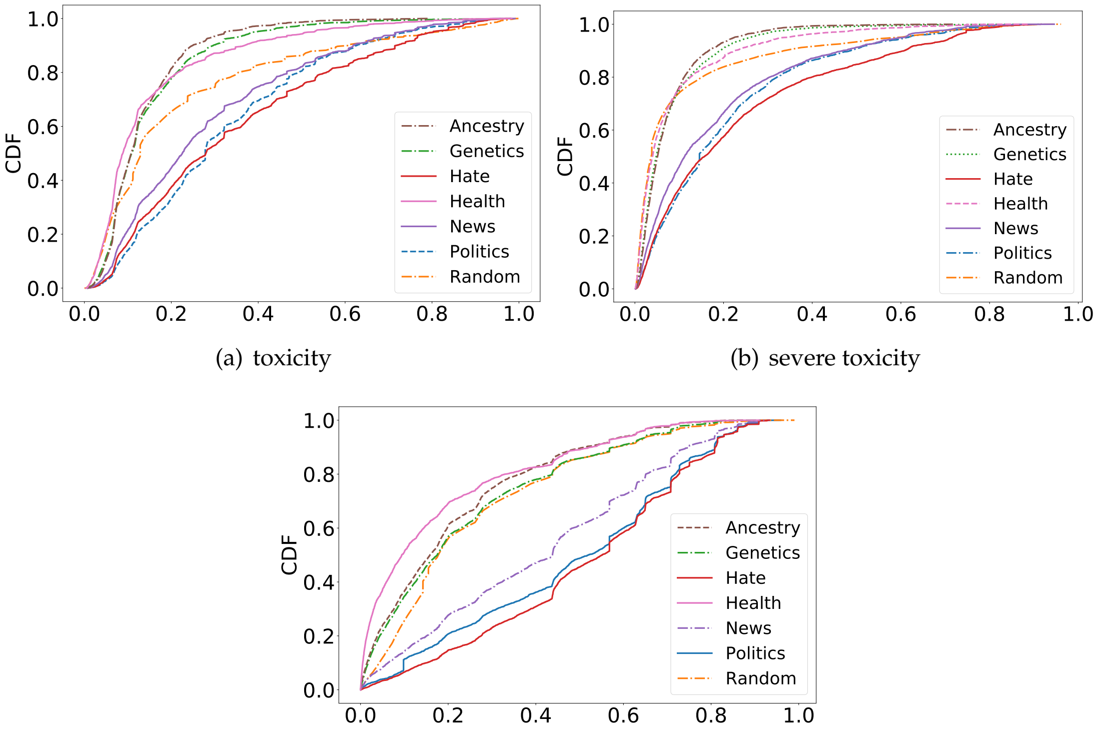


FIGURE 6.1: Subreddits with genetic testing related comments listed in Table 6.2 grouped into categories based on their thematic topics (excluding a generic 'other' category).

relatively popular in subreddits about dog breed identification (/r/IDmydog, 1%), children (/r/Adoption, 1%), entertainment (/r/TheBlackList, 0.6%), health (/r/ehlersdanlos, 0.7%), and crime (e.g., /r/EARONS, 0.3%). By contrast, in the random dataset, only 6 out of 204K comments (0.003%) include a genetic testing keyword. Naturally, these percentages depict conservative lower bounds as: 1) comments can be replied to by other comments, thus creating different branches of discussion, and 2) one can comment on a topic about genetic testing without using a keyword. However, our approach provides ample data points for our analysis.

Topics and toxicity. In the rest of this section, we analyze the 19 categories of subreddits in terms of the topics being discussed as well as the toxicity of the comments therein, using, respectively, LDA and Google's Perspective API [175]. The API returns three values between 0 and 1, pertaining to: 1) Toxicity, i.e., how rude, disrespectful, or unreasonable a comment is likely to be; 2) Severe Toxicity, which is similar to toxicity but only focuses on the "most toxic" comments; and 3) Inflammatory, which focuses on texts intending to provoke or inflame. In Figure 6.2, we plot the CDFs of the toxicity of



(a) toxicity

(b) severe toxicity

(c) inflammatory

FIGURE 6.2: CDFs of Google's Perspective API toxicity on the genetic testing comments for the three most/least toxic subreddit categories.

the comments for the three most and the three least toxic subreddits (we also compare to the random dataset as a baseline). We run two-sample Kolmogorov-Smirnov (KS) tests between the distribution of each category and the random dataset: in all cases, we reject the null hypothesis that they come from a common parent distribution ($p < 0.01$). We note that the two-sample KS test is non-parametric and thus robust in terms of different sample sizes. While we acknowledge this might not be a perfect sampling, it is unlikely that any sampling method would result in perfectly balanced datasets. Also, recall that we are primarily interested in the overall comparison of content related (and unrelated) to genetic testing, thus this is appropriate for our purposes. Overall, the comments originating from subreddits related to genetics, ancestry, and health are less

Topic	Category: Hate
1	dna (0.069), test (0.055), get (0.017), would (0.016), like (0.014), testing (0.013), know (0.012), one (0.011), think (0.009), take (0.008)
2	child (0.037), men (0.023), women (0.022), father (0.019), woman (0.015), support (0.014), man (0.014), paternity (0.014), birth (0.011), get (0.008)
3	white (0.034), people (0.021), african (0.016), black (0.015), european (0.013), race (0.013), ancestry (0.011), like (0.008), american (0.007), genetic (0.006)
4	jewish (0.028), native (0.017), american (0.015), israel (0.015), trump (0.013), clinton (0.010), jews (0.009), cherokee (0.007), citizenship (0.007), indian (0.007)
5	rep (0.027), dem (0.027), act (0.012), gay (0.007), body (0.007), gender (0.006), use (0.004), vote (0.004), proper (0.003), russia (0.003)
6	testing (0.023), genetic (0.022), data (0.008), insurance (0.008), company (0.007), health (0.007), consent (0.007), paternity (0.006), companies (0.005), google (0.005)
7	rape (0.021), women (0.012), lie (0.010), man (0.010), police (0.008), case (0.007), false (0.007), evidence (0.007), sex (0.006), point (0.005)
8	genetic (0.016), human (0.006), even (0.006), testing (0.006), would (0.006), race (0.006), medical (0.006), differences (0.005), social (0.005), could (0.004)
9	youtube (0.010), talk (0.008), islamic (0.007), gedmatch (0.005), watch (0.005), working (0.005), video (0.005), dude (0.004), coast (0.004), saliva (0.004)
10	people (0.009), would (0.008), women (0.008), genetic (0.006), like (0.006), men (0.006), good (0.006), think (0.006), one (0.006), want (0.006)

TABLE 6.3: LDA analysis of the Hate subreddits.

Topic	Category: Genetics
1	dna (0.021), family (0.015), know (0.013), would (0.013), test (0.013), father (0.012), one (0.011), great (0.011), dad (0.009), mother (0.009)
2	european (0.023), ancestry (0.023), dna (0.017), african (0.015), results (0.014), people (0.014), native (0.013), american (0.012), eastern (0.011), german (0.009)
3	chromosome (0.031), haplogroup (0.031), ashkenazi (0.021), jewish (0.019), confidence (0.015), maternal (0.012), paternal (0.011), chromosomes (0.011), also (0.011), line (0.010)
4	genetic (0.021), testing (0.014), test (0.011), would (0.011), information (0.007), like (0.007), people (0.007), results (0.007), get (0.006), know (0.006)
5	data (0.028), snps (0.020), one (0.013), snp (0.013), snpedia (0.011), gene (0.011), genome (0.010), raw (0.009), promethease (0.008), variant (0.008)
6	blood (0.035), hair (0.023), eyes (0.018), type (0.017), cells (0.015), skin (0.015), blue (0.012), dark (0.011), brown (0.010), saliva (0.009)
7	asian (0.055), chinese (0.039), wegene (0.032), south (0.025), results (0.020), east (0.016), korean (0.014), japanese (0.014), southeast (0.013), customers (0.012)
8	sample (0.031), results (0.018), weeks (0.017), received (0.014), time (0.013), kit (0.013), samples (0.012), extraction (0.011), process (0.011), people (0.011)
9	gedmatch (0.054), dna (0.044), data (0.033), ancestry (0.026), results (0.023), raw (0.020), upload (0.016), use (0.015), get (0.013), also (0.012)
10	ancestry (0.025), promethease (0.023), health (0.022), data (0.019), get (0.017), reports (0.017), report (0.014), new (0.011), results (0.011), ancestrydna (0.010)
Topic	**Category: Ancestry**
1	match (0.029), dna (0.026), matches (0.025), one (0.016), cousins (0.014), shared (0.013), share (0.011), cousin (0.011), related (0.011), gedmatch (0.010)
2	dna (0.020), family (0.019), test (0.018), great (0.012), father (0.012), know (0.011), mom (0.011), would (0.011), mother (0.010), side (0.010)
3	native (0.085), american (0.076), cherokee (0.018), ancestry (0.014), indian (0.011), nbsp (0.009), family (0.009), tribe (0.009), claim (0.008)
4	dna (0.026), ancestry (0.018), results (0.011), irish (0.009), people (0.009), european (0.008), like (0.008), african (0.008), ethnicity (0.008), british (0.008)
5	william (0.019), youtube (0.016), watch (0.016), african (0.014), norwegian (0.013), sub (0.011), saharan (0.011), middle (0.009), census (0.008)
6	dna (0.062), test (0.049), testing (0.020), father (0.020), would (0.019), autosomal (0.013), family (0.012), get (0.012), line (0.011), haplogroup (0.010)
7	ancestry (0.049), gedmatch (0.045), ftdna (0.028), dna (0.026), upload (0.024), results (0.024), test (0.023), matches (0.022), get (0.018), data (0.017)
8	jewish (0.031), european (0.023), asian (0.020), europe (0.018), east (0.017), eastern (0.017), italian (0.015), results (0.015), ancestry (0.014), ashkenazi (0.013)
9	dna (0.037), ancestry (0.018), ancestrydna (0.016), test (0.015), testing (0.013), data (0.010), tests (0.009), results (0.008), tree (0.008), information (0.007)
10	tree (0.029), find (0.018), family (0.017), people (0.016), trees (0.013), ancestry (0.012), see (0.012), records (0.012), matches (0.010), search (0.009)

TABLE 6.4: LDA analysis of the Genetics and Ancestry subreddits.

toxic than a random baseline, while comments in news, politics, and "hateful" subreddits are remarkably more toxic.

Remarks. We choose to use Google's Perspective to identify hateful content as other methods, e.g., hate speech detection libraries [61], are primarily trained on short texts with a limited number of training samples. Whereas, our datasets contain lengthy comments; thus, the Perspective API should perform better. In the rest of the section, we report a few representative comments for each category based on our topic analysis.

6.3.2 Genetic Testing & Racism

Remarkably, 10/114 subreddits in our sample are categorized as hateful as they are broadly associated with hateful content. Some are clearly associated with the alt-right [206] (e.g., /r/altright, /r/DebateAltRight, and /r/The_Donald), sexism, or racism. For instance, /r/TheRedPill includes misogyny and toxic behavior towards women [146], while /r/MGTOW, Men Going Their Own Way, is a forum for men who reject romantic relationships with women, and was identified as a supremacist group by the Southern Poverty Law Center [205]. Other subreddits in this group include /r/milliondollarextreme, an American sketch satire show associated with alt-right and antisemitism [199] which was banned in September 2018, as well as /r/KotakuInAction, which is associated with GamerGate-related toxicity [47]. Also, /r/BlackPeopleTwitter makes fun of tweets purporting to originate from African Americans.

With this in mind, we set to study the relation between genetic testing and racism on Reddit. Our Perspective API analysis (see Figure 6.2) shows that the category related to hate is the most toxic, and some of the subreddits (e.g., /r/DebateAltRight, /r/altright) have among the highest number of comments including genetic testing keywords in this category of subreddits. In this context, the LDA modeling gives us insight on how these fringe communities discuss genetic testing; see Table 6.3. Users often discuss their desire to get tested (e.g., dna, test, would, like, know), while others argue on issues related to paternity (e.g., paternity, father, support). Although we find similar topics in genetics/ancestry and parenting subreddits, here they are being expressed in a much more toxic/inflammatory manner; as evidenced by Figure 6.2. For example, a user writes in /r/TheRedPill: “Would get a DNA test on those kids ASAP. I don’t know why all men don’t do them secretly as soon as the kids are born.”

Other topics are related to ancestry results (e.g., jewish, american, european) as well as race in general (e.g., white, black, race), which are not as widely discussed in genetics/ancestry subreddits (see Table 6.4). Again, the conversations exhibit clear racist connotations; e.g., a user writes in /r/DebateAltRight: “The Jews know who Jews are [...] It doesn’t require genetic testing [...] We whites know who whites are. Non-whites know who whites are. Anyone with eyes knows who whites are. And we will fight for our race!” Finally, we find topics related to sexual crimes (e.g., ‘rape’, ‘women’, ‘evidence’, ‘sex’), homosexuality and gender (e.g., ‘gay’, ‘gender’), and insurance (e.g., 'insurance', 'company', 'health').

Overall, genetic testing is a relatively popular topic of discussion in subreddits associated with fringe political views. When looking at the comments with the highest toxicity, we find some disturbing content, including instances of xenophobia (e.g., "Can you be Alt-Right and have non-white friends?", receiving the reply "No, as a member of the Alt-Right you have to DNA test all of your friends and if they're not 100% White then you report them to your local Atomwaffen," referring to a neo-nazi terrorist organization [204]). Some users explicitly advocate using genetic testing to eliminate groups of non-white ancestry (e.g., "You know with pre-implantation genetic testing we can breed out non-white ancestry fairly easily [...]").

6.3.3 What Is the Genetic Testing Discourse About?

Next, we select a few categories of subreddits and analyze them via topic modeling and the toxicity metrics, aiming to better understand how users perceive genetic testing in each context. To ease presentation, we only do so on interesting or unexpected categories.

Genetics & Ancestry. As mentioned, the subreddits with the highest ratio of genetic testing keywords (see top five subreddits in Table 6.2) are directly related to genetic testing and ancestry. This is confirmed by LDA (see Table 6.4). In fact, even in the genetics category, the discussion is dominated by ancestry (e.g., european, ashkenazi, african) and family (e.g., family, father, mother). We also observe that the open personal genomics database and genealogy website, GEDmatch [89], is one of the topics with the greatest weights (0.054); see Table 6.4. GEDmatch allows users to upload their genetic data obtained from DTC genetic testing companies to identify potential relatives who have also uploaded their data. Interestingly, in December 2018, US police forces declared that GEDmatch helped them find suspects in 28 cold murder and rape cases [98]. Overall, as shown in Figure 6.2, the subreddits about genetics and ancestry attract far less toxic comments than the random Reddit sample, and are the least toxic categories among the rest in our dataset. In particular, we observe extremely low levels of inflammatory content.

Crime Investigations. Genetic testing appears to be discussed in subreddits falling in the crime category, e.g., /r/EARONS, the East Area Rapist/Original Night Stalker, a.k.a. the Golden State Killer [155]. We also find subreddits covering (often controversial) discussions about Steven Avery, who was wrongly convicted of sexual assault and attempted murder; e.g., /r/StevenAveryIsGuilty seems to firmly believe Avery was justly convicted, while /r/TickTockManitowoc does not. The LDA analysis confirms how discussion in this category revolves around investigation and evidence (e.g., blood, sample, evidence); see Table 6.5. A user writes: "Similarly, why didn't we get more impact out of the DNA test on the key? Specifically, one non-courtroom interviewee makes the point that TH DNA should have been all over the key, because she had owned it for many years. The fact that only SAs DNA was found seems to be evidence that it was in fact wiped/disinfected. Why wasn't this a bombshell to be used in court?", while another says: "I am not convinced the DNA matched Teresa. I think they were probably random bones from a cadaver. Read about the DNA testing. It only matches in 7 of 15 locations." The toxicity and inflammatory levels of the content of this category are similar to the random dataset, which, combined with the LDA results, suggest that genetic testing here is discussed for informational reasons.

Parenting. Users also discuss genetic testing in the context of children, pregnancy, and parenting; e.g., in /r/Parenting, /r/Adoption, /r/TryingForABaby, /r/infertility. From the LDA analysis (see Table 6.5), we find that users often discuss topics related to the identity of the father or child support (e.g., father, support, lawyer), but also health and the characteristics of their child (e.g., ultrasound, gender, embryos). For example, a user is trying to support a woman admitting to having difficulties conceiving a child by saying: "Get a 2nd opinion. And just remember the DNA test was normal. Be calm.

Topic	Category: Crime
1	dna (0.041), would (0.020), testing (0.019), think (0.016), people (0.012), like (0.011), test (0.011), know (0.010), could (0.009), get (0.009)
2	blood (0.060), dna (0.043), testing (0.023), test (0.019), sample (0.013), vial (0.012), samples (0.012), tested (0.010), lab (0.009), tests (0.009)
3	found (0.016), murder (0.014), police (0.013), case (0.010), years (0.009), later (0.009), dna (0.008), man (0.007), went (0.007), convicted (0.006)
4	dna (0.054), test (0.020), evidence (0.019), testing (0.011), would (0.011), bullet (0.010), could (0.009), one (0.008), case (0.007), found (0.007)
5	one (0.011), would (0.007), control (0.007), lab (0.006), test (0.006), like (0.006), case (0.006), evidence (0.005), science (0.005), say (0.005)
6	evidence (0.023), avery (0.020), testing (0.016), dna (0.014), case (0.013), court (0.009), allen (0.008), trial (0.008), would (0.008), state (0.007)
7	father (0.031), family (0.023), mother (0.012), son (0.012), dad (0.011), related (0.011), adam (0.011), cousin (0.010), cousins (0.009), different (0.008)
8	said (0.019), fire (0.016), family (0.012), hobbs (0.008), brendan (0.007), barb (0.006), sketch (0.005), monday (0.005), richard (0.005), death (0.004)
9	avery (0.029), blood (0.017), would (0.017), evidence (0.017), found (0.015), key (0.011), garage (0.010), car (0.008), trailer (0.007), police (0.007)
10	bones (0.049), bone (0.035), remains (0.029), found (0.023), human (0.019), fragments (0.017), burn (0.016), pit (0.015), body (0.014), teresa (0.013)
Topic	**Category: Children**
1	testing (0.023), genetic (0.020), weeks (0.016), back (0.014), pregnancy (0.012), first (0.012), loss (0.010), results (0.010), pregnant (0.009), get (0.009)
2	genetic (0.035), testing (0.017), child (0.017), children (0.014), people (0.012), health (0.012), would (0.011), kids (0.009), medical (0.008), life (0.008)
3	know (0.019), like (0.019), want (0.014), would (0.014), get (0.013), really (0.012), feel (0.011), time (0.010), think (0.009), even (0.009)
4	child (0.050), dna (0.030), test (0.028), father (0.023), support (0.020), kid (0.016), dad (0.011), lawyer (0.011), paternity (0.010), get (0.009)
5	insurance (0.025), testing (0.013), get (0.013), genetic (0.012), doctor (0.012), labcorp (0.011), blood (0.009), pay (0.009), test (0.009), covered (0.008)
6	dna (0.030), test (0.020), family (0.019), parents (0.013), ancestry (0.012), also (0.011), birth (0.011), adoption (0.011), find (0.011), get (0.010)
7	weeks (0.034), genetic (0.029), scan (0.024), girl (0.022), testing (0.022), ultrasound (0.021), boy (0.016), baby (0.014), week (0.013), gender (0.012)
8	test (0.022), dna (0.016), name (0.012), back (0.012), got (0.008), came (0.008), said (0.008), little (0.008), son (0.007), chow (0.006)
9	ivf (0.017), embryos (0.014), testing (0.011), one (0.010), genetic (0.010), pgs (0.010), dog (0.010), sperm (0.010), embryo (0.009), transfer (0.009)
10	genetic (0.032), testing (0.030), test (0.024), would (0.015), risk (0.012), baby (0.011), done (0.011), also (0.009), results (0.008), back (0.008)

TABLE 6.5: LDA analysis of the Crime and Children subreddits.

My uterus is sending yours good vibes.” Once again, the subreddits in this category contain low levels of toxicity.

Animals. Reddit users also use genetic testing keywords in subreddits related to animals, and more specifically those related to dogs .For instance, /r/IDmydog, which is the 8th ranked subreddit in terms of genetic testing comments, focuses on identifying dog breeds from pictures. For example, a user writes: “Turns out, Cherry’s DNA test came out as half purebred Miniature Schnauzer, one-eighth Chihuahua, and the rest possibly unknown terrier mix.” /r/dogs and /r/pitbulls focus on discussion about dogs and pitbulls respectively. This is also confirmed by LDA (e.g., breeds, terrier, mixed); see Table 6.6. An interesting topic of discussion is related to dog breeds banned in certain countries [8], and how one can be identified through DNA testing. For example, a user writes: “Why don’t you get a DNA test done and see what he really is? If just by some chance he’s not a banned breed, you can show that to your vet and get them to change the breed listed on record and then you can use that to show potential landlords if they say he looks like something that he’s not.” Once again, this category has similar levels of toxicity and inflammatory content to the random dataset.

Other categories. Genetic testing is also discussed in educational contexts

Topic	Category: Animals
1	pit, breeds, breed, bull, amp, bulls, terrier, dogs, mixed, jpg
2	dna, test, mix, like, dog, get, know, really, could, would
3	dogs, genetic, still, testing, breed, thing, even, dog, issues, pedigree
4	health, genetic, breed, dogs, breeder, testing, breeding, breeders, dog, puppy
5	dog, breed, dna, banned, type, test, prove, one, court, pit

TABLE 6.6: LDA analysis of the Animals subreddits.

Topic	Category: Privacy
1	dna, privacy, information, data, genetic, testing, ancestry, would, like, people
2	would, child, test, people, one, privacy, father, think, know, right
3	dna, genetic, health, information, employers, bill, wellness, sequencing, would, testing
4	dna, act, table, formatted, view, article, privacy, genetic, gender, testing
5	cancer, breast, congress, house, trump, bill, genetic, republicans, act, laws

TABLE 6.7: LDA analysis of the comments including 'privacy'.

(e.g., /r/explainlikeimfive, /r/ NoStupidQuestions), to learn about science (e.g., /r/science, /r/futurology), discuss their health (e.g., /r/celiac, /r/cancer), or in the context of drugs (/r/Nootropics, /r/steroids). User also use words related to genetic testing in a legal context (/r/legaladvice), to discuss subjects related to their cultural background (e.g., /r/arabs, /r/judaism), as well as religion (e.g., /r/exmormon). Finally, we find genetic testing words in subreddits related to entertainment programs (e.g., /r/TheBlackList), comedy (e.g., /r/funny), and issues related to gender (e.g., /r/AskMen, /r/AskWomen).

6.3.4 Privacy Concerns

We also examine comments where users discuss privacy concerns in the context of genetic testing. We extract comments which include both a genetic testing keyword and the word 'privacy' from our Reddit dataset, getting 560 comments (0.7% of all comments). Obviously, this set is a conservative sample, as it is possible for a user to discuss issues related to privacy without specifically mentioning the word 'privacy'. Then, we use LDA to identify the context in which users discuss issues related to privacy; see Table 6.7.

The most common subreddits in which privacy issues are being discussed are /r/genealogy, /r/news, and /r/23andMe. We find that Reddit users express privacy concerns on the use of genetic testing (e.g., dna, data, information, privacy, gender). Specifically, a topic of discussion is the potential misuse of genetic information by employers, while another topic focuses on paternity tests and whether children have the right to know their biological father. Finally, several users discuss the privacy issues stemming from a bill passed by the Republican Party on March 8, 2017 which allows companies to ask for their employees' genetic test results [21].

6.3.5 Where Do Users Post About Otherwise?

We also examine the overlap in users discussing genetic testing among all 114 subreddits in our sample. We do so to examine whether subreddits that have common interests have also similar user base. For instance, we want to assess if users that post on /r/23andMe, also post on /r/ancestry. To do so, we extract the set of users that posted in each subreddit and calculate the pairwise Jaccard Index scores between the set of users in each subreddit. Next, we create a complete graph where nodes are the subreddits and edges are weighted by the Jaccard Index. We then run the community detection algorithm in [30], which provides a set of communities based on the graph's structure.

Figure 6.3 shows the resulting graph: nodes that have the same color are part of the same community. The main observations are the following: 1) there are high Jaccard Index scores between the nodes in the same community, i.e., there is a substantial overlap of users that posted in all subreddits within the community. 2) Genetic testing subreddits (e.g., /r/genetics, /r/ancestry, /r/23andMe) are part of the same community (pink nodes) as scientific and education ones (e.g., /r/askscience, /r/science), highlighting that "enthusiasts" are also active on scientific subreddits. 3) Subreddits associated with sexist content essentially share the same users (e.g., /r/MGTOW, /r/TheRedPill, lower

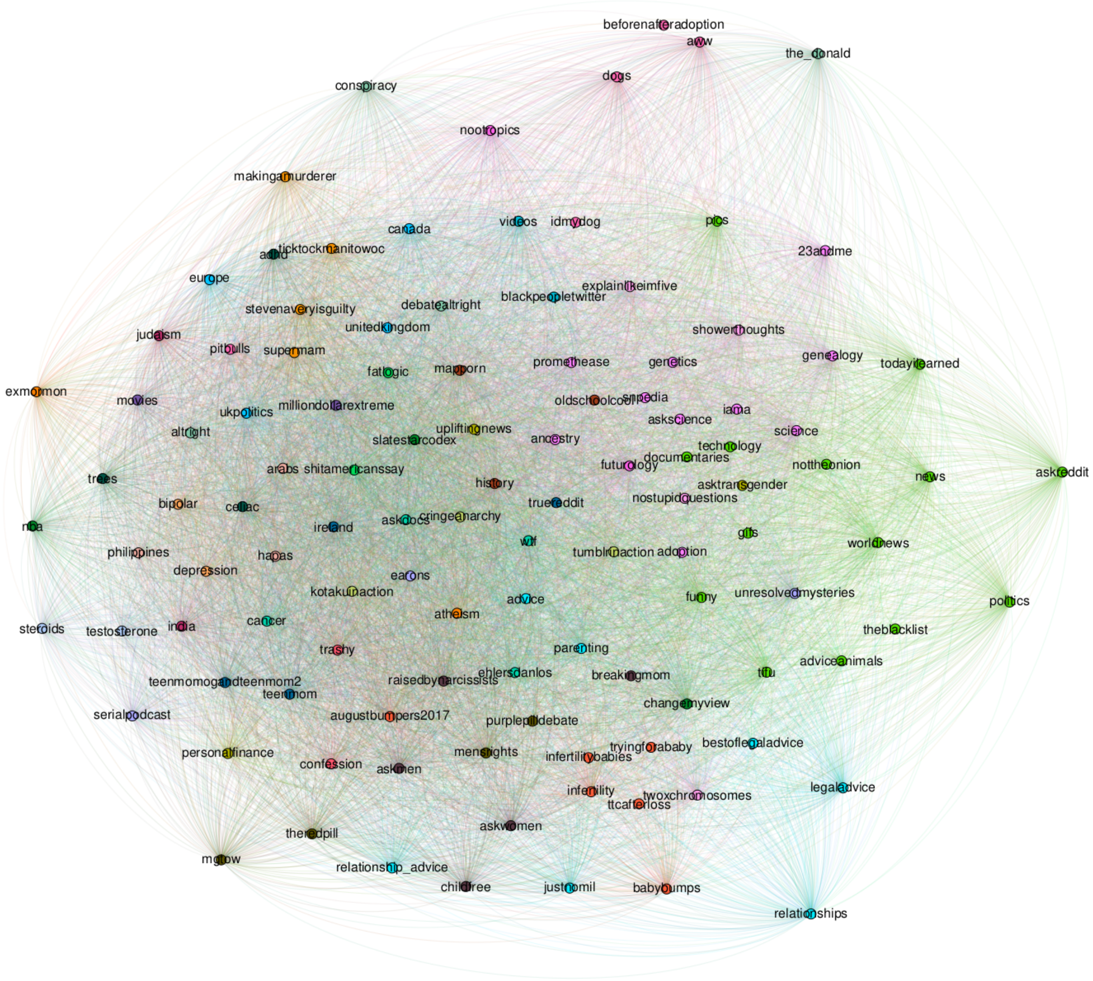


FIGURE 6.3: Graph depicting the Jaccard Index of the users whose comments include genetic testing keywords for each subreddit.

left in olive green); also, users who discuss genetic testing in /r/The_Donald are also active in other alt-right subreddits like /r/AltRight, /r/DebateAltRight (mint green nodes).

Additionally, we find communities with subreddits focused on the geopolitical aspects of genetic testing (see light blue nodes on the top left) like /r/europe, /r/canada, /unitedkingdom, and /r/ukpolitcs, as well as subreddits about personal advice (light blue nodes on the bottom right) like /r/advice, /r/parenting, /r/legaladvice, /r/bestoflegaladvice. Other communities are centered around conceiving children (e.g., /r/infertility, /r/tryingforababy,

/r/babybumps, orange nodes on the bottom right side), crime investigation (e.g., /r/MakingaMurderer, /r/StevenAveryIsGuilty, orange nodes on the top left side), and animals (e.g., /r/dogs, /r/IDmydog, /r/pitbulls, pink nodes on top right side).

Overall, Reddit users are not uniformly interested in every aspect of genetic testing, but rather specific communities focus on specific aspects thereof. For example, we find groups ranging from genetic testing enthusiasts, i.e., those who are interested in or have undergone genetic testing, to people who discuss genetic testing exclusively in subreddits with educational and scientific content, to those who use genetic testing terminology exclusively when discussing fringe political views.

Take-Aways. Our Reddit analysis shows that genetic testing is discussed in a variety of contexts which in itself is an indicator of how mainstream it has become. For instance, users discuss it in the context of issues related to their children, pets, or health, or to debate on their cultural heritage. More interestingly, they are not uniformly interested in every aspect of genetic testing, rather, they form *groups* ranging from genetic testing enthusiasts to individuals with fringe political views. Thus, we observe a dichotomy in the type of users interested in genetic testing: some focus in typical uses of genetic testing, others discuss their use in worrying ways. Specifically, we find evidence of toxic language displaying clear racist connotations, and of groups of users using genetic testing to push racist agendas, e.g., to eliminate or marginalize minorities. This is worrying since Reddit is a mainstream platform (5th most visited site in the US [181]).

6.4 Genetic Testing Discussions on 4chan's /pol/

In this section, we study the prevalence of genetic testing comments on 4chan's politically incorrect board (/pol/). We first conduct a general characterization of the threads containing genetic testing keywords where we, similarly to the previous section, use Google's Perspective API to measure the toxicity of the contents, and LDA modeling to

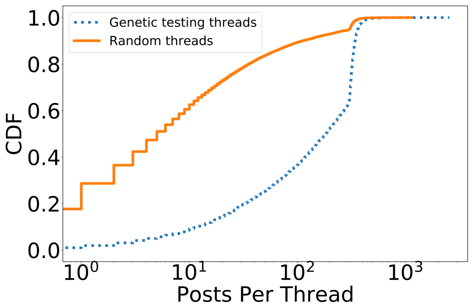


FIGURE 6.4: CDF comparing 4chan threads with genetic testing keywords and random threads in terms of number of posts.

extract the most prominent topics of discussion. Then, we use Perceptual Hashing [156] and clustering to study the imagery and memes of the dataset.

6.4.1 General Characterization

Thread Activity. We begin by measuring the number of posts in threads where genetic testing keywords appear, aiming to examine whether these threads attract more or less activity than "usual." On /pol/, there is a limit on how many threads can simultaneously be active: whenever a new one is created, the one with the oldest last post is purged. There is also a "bump" limit that prevents a thread from never being purged. As per [108], the majority of threads attract only a few posts before being archived, while some—often covering controversial or popular topics—get many posts and possibly hit the bump limit.

In Figure 6.4, we plot the CDF of the number of posts per thread, for both the genetic testing threads and our random sample. The former have an order of magnitude more posts than the latter (the median is 183 and 5 posts, respectively), which indicates that genetic testing is often discussed in long-lasting/interesting threads and may attract more attention by users. We also run a two-sample KS test on the distributions and we reject the null hypothesis that they come from a common parent distribution ($p < 0.01$).

Topic	4chan
1	ancestry (0.048), african (0.046), european (0.023), white (0.015), american (0.012), north (0.011), americans (0.010), population (0.008), south (0.008), europeans (0.008)
2	youtube (0.030), watch (0.028), jewish (0.020), king (0.013), company (0.010), lauren (0.010), tut (0.009), monkey (0.008), igenea (0.007), haplogroup (0.006)
3	ancient (0.023), modern (0.020), egyptians (0.015), egypt (0.012), years (0.009), national (0.008), egyptian (0.008), greeks (0.008), roman (0.007), saharan (0.007)
4	women (0.015), children (0.015), woman (0.011), men (0.010), man (0.009), genes (0.009), kids (0.009), child (0.008), two (0.008), birth (0.008)
5	genetic (0.030), data (0.022), ancestrydna (0.014), information (0.014), health (0.013), company (0.012), testing (0.011), research (0.011), use (0.008), send (0.007)
6	back (0.022), got (0.021), european (0.020), family (0.020), german (0.013), took (0.012), irish (0.011), hair (0.011), came (0.011), eyes (0.010)
7	dna (0.063), test (0.042), white (0.024), like (0.017), people (0.015), would (0.012), genetic (0.012), one (0.011), get (0.011), even (0.010)
8	gedmatch (0.024), raw (0.014), creation (0.008), human (0.007), far (0.007), data (0.007), got (0.007), son (0.006), run (0.006), forum (0.006)
9	screw (0.016), tweet (0.010), bill (0.010), tea (0.010), news (0.010), reddit (0.009), look (0.007), fda (0.005), search (0.005), guy (0.005)
10	companies (0.018), pay (0.016), child (0.015), order (0.015), racists (0.014), support (0.012), testing (0.011), adding (0.011), admit (0.011), law (0.011)

TABLE 6.8: LDA analysis of 4chan's /pol/.

Toxicity & Hate. We then measure hate and toxicity in /pol/ threads by computing: 1) the percentage of hate words, and 2) the toxicity/inflammatory levels. For the former, we use a dictionary of hate words compiled by and available from hatebase.org, as used in [108]; for the latter, we again rely on the Perspective API. However, we find no major differences between the genetic testing threads and the random sample—which is not surprising as /pol/ is known for its high level of hate speech [108]—thus, we omit related plots to ease presentation.

Topic Modeling. We also use LDA modeling to identify the most prominent topics of discussion; see Table 6.8. Similar to Reddit, 4chan users use keywords suggesting their intention to get tested (e.g., would, get, dna, test). Several topics are related to ancestry, which is also among the words with the highest weights (0.048); for instance, users often discuss the ancestral background of the American population (e.g., american, african, european, white), others debate the cultural connection of modern humans to ancient civilizations (e.g., egyptians, greeks, roman), and the facial traits of modern europeans (e.g., german, irish, eyes, hair). Interestingly, another prominent topic of discussion is related to Lauren Southern (e.g., lauren, jewish, youtube), an Internet personality associated with the alt-right, whose popularity rose after being detained in Italy for trying to block a ship rescuing refugees [54]. Other conversations likely relate to how genetic testing companies use their data (e.g., genetic, data, use, research), as well as legal issues related to child support (e.g., child, birth, support, law).

Entity	Clusters (%)	Entity	Clusters(%)
/pol/	15 (6.9%)	Video	3 (1.4%)
Lauren Southern	15 (6.9%)	Jewish people	3 (1.4%)
23andMe	13 (6.0%)	Logo	3 (1.4%)
Pepe the Frog	9 (4.1%)	White	3 (1.4%)
United States of America	8 (3.7%)	Shaun King	2 (0.9%)
Richard Spencer	5 (2.3%)	Screenshot	2 (0.9%)
Genetic	4 (1.8%)	4chan	2 (0.9%)
Meme	4 (1.8%)	The Holocaust	2 (0.9%)
Europe	3 (1.4%)	Race	2 (0.9%)
Greece	3 (1.4%)	Adolf Hilter	2 (0.9%)

TABLE 6.9: Top 20 entities with the most clusters.

6.4.2 Image Analysis

Next, we look at the images and memes that are shared in /pol/ posts including genetic testing keywords. We use the image analysis pipeline introduced in [235] which uses Perceptual Hashing [156] and clustering techniques to group together images that are visually similar. We run the pipeline on the 6,375 images included in *posts* where at least one genetic testing keyword appears; as discussed earlier, this is in contrast to the textual analysis where we look at whole threads. (Recall from Table 6.1 that the total number of images in threads containing genetic testing keywords is 338,540.) We obtain 215 clusters including 543 total images; the other 5,832 images are labeled as noise by the clustering algorithm and thus we discard them. This high noise ratio mirrors findings in [235] and is likely due to 4chan users creating a lot of original content [108]. Also, our dataset only includes a few thousand images, thus not a lot of images are visually similar.

We annotate each cluster using Google's Cloud Vision API[2], specifically, we calculate the medoid of each cluster (i.e., its "representative" image) following the methodology by [235], and use that image to query the API. This returns a set of meaningful entities, which are obtained by searching labeled images across the Web, along with their confidence scores. The exact methodology for extracting the entities is not known,

[2] https://cloud.google.com/vision/

Topic	Entity: 23andMe
1	dna (0.050), ancestry (0.035), tests (0.024), results (0.018), one (0.018), percent (0.016), african (0.016), got (0.014), would (0.014), could (0.014)
2	could (0.030), jewish (0.030), even (0.023), pol (0.023), people (0.023), also (0.023), company (0.016), test (0.016), results (0.016), markers (0.016)
3	white (0.039), genetic (0.034), test (0.034), heritage (0.022), european (0.022), dna (0.018), jew (0.018), like (0.018), nigger (0.014), still (0.014)
Topic	**Entity: United Stated of America**
1	white (0.044), ancestry (0.038), americans (0.031), self (0.028), african (0.021), european (0.018), even (0.018), whites (0.018), race (0.018), american (0.018)
2	white (0.039), roman (0.024), people (0.021), whites (0.018), full (0.018), empire (0.016), citizenship (0.016), held (0.016), admixture (0.016), like (0.016)
3	sargon (0.042), get (0.037), spencer (0.032), enoch (0.032), like (0.027), anyone (0.027), think (0.022), say (0.017), would (0.017), even (0.017)

TABLE 6.10: LDA analysis of the texts in the /pol/ posts with imagery annotated as '23andMe' or 'United Stated of America'.

however, upon manual examination, we can confirm that the API is indeed able to extract fine-grained entities. For instance, given an image with Donald Trump, the API returns an entity called "Donald Trump" and not generic labels like "man" or "politician."

For each cluster, we extract the entity with the highest confidence score and analyze the top 20 entities, as reported in Table 6.9. The most popular entries are /pol/ itself and Lauren Southern with 6.9% of all clusters. The latter is interesting as it adds to the evidence that discussions about genetic testing frequently involve alt-right celebrities. In fact, pictures of American white-supremacist Richard Spencer [227] (6th most popular with 2.3% of all clusters), and Carl Benjamin, a YouTuber known for his misogynistic involvement in the GamerGate controversy [27], are also popular. We also find clusters related to: 1) 23andMe (6.0%), e.g., screenshots of genetic testing results from 23andMe or images with the 23andMe logo, 2) memes including Pepe the Frog (4.1%), a 4chan-popularized hate symbol [4], and 3) geographic images related to, e.g., the US (3.7%), Europe (1.4%), or Greece (1.4%). The latter is likely mirroring discussions about the connection of modern humans to ancient civilizations; see topic 6 in Table 6.8. We also find imagery related to the Jewish community (1.4%), as well as the Holocaust (0.9%) and Hitler (0.9%), suggesting that, on 4chan, genetic testing terms and Nazi-related imagery are used together for the dissemination of hateful and antisemitic content.

We also examine the entities in Table 6.9 more closely to shed light on the context

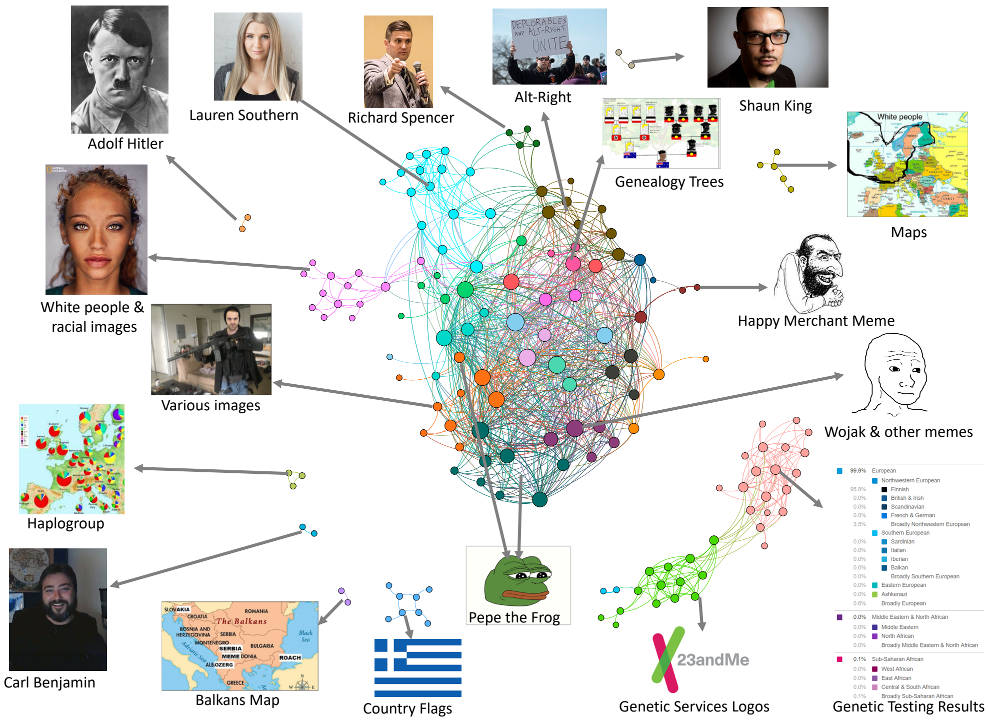


FIGURE 6.5: Visualization of the image clusters with manual annotation.

in which images are being discussed. Specifically, we extract text from the posts appearing alongside the images and use LDA modeling on the posts of each entity separately. We set LDA to produce only three topics per entity given the limited number of posts per entity. Among other things, we find that posts containing images related to 23andMe (see Table 6.10) actually include discussions with racial connotations; for instance, whether test results show signs of African ancestry (e.g., ancestry, percent, african), or whether people with Jewish heritage are behind the company (e.g., jewish, company, results). For example, a user writes: "Can a genetics company founded by a Jew be trusted?" Similarly, posts with images annotated as United States of America (see Table 6.10) reveal discussions on the ancestral background of the American population (e.g., americans, ancestry, african, whites). A user writes: "Less than 5% of White Americans have even negligible amounts of African DNA".

Cluster visualization. Finally, we provide a visualization of the clusters in Figure 6.5.

Nodes in the graph represent clusters, while edges represent the Jaccard Index between clusters (as per the entities returned by the Cloud Vision API). To ease presentation, we only consider edges where the Jaccard Index is greater than 0.2, a threshold we select after inspecting the distribution of all the Jaccard Index scores. This corresponds to selecting 4.1% of the edges with the highest Jaccard Index, allowing us to understand the *main* connections between clusters.

Then, we perform community detection, using the approach presented in [30]. This considers the structure of the graph and decomposes it into a set of communities, where each community includes a set of highly inter-connected nodes. The resulting graph is presented in Figure 6.5, with each color representing a different community. For each community, we have manually inspected the images in the clusters and added a high-level description as well as a representative image.

The figure highlights the presence of two tightly-knit communities (bottom right): the green community includes images with logos of genetic testing companies, while the light red community covers images with screenshots of genetic testing results. We also find communities with images related to Haplogroups and Genealogy Trees, as well as others related to the alt-right (top of the graph). In fact, a few communities exhibit clear racial connotations (pink), e.g., a cluster including an image from National Geographic predicting how the average American woman will look like in 2050 [86], which, unsurprisingly, attracted numerous posts on 4chan. Finally, a few communities are related to hateful memes like Pepe the Frog and the Happy Merchant, a caricature of a manipulative Jew used on 4chan in racist contexts [81].

Take-aways. Overall, we find that genetic testing is a rather popular topic of discussion in 4chan's /pol/, often appearing in long/active threads. Also, genetic testing topics are often accompanied by images and memes with clear racial or hateful connotations. While the presence of highly toxic content in /pol/ is unsurprising, the specific content which accompanies threads related to genetic testing is very worrying. We find imagery with prominent figures of the alt-right movement (e.g., Lauren Southern, Richard Spencer), antisemitic memes (e.g., Happy Merchant), and topics of discussion using words with racial/hateful meaning (e.g., jewish, nigger), which may be an indicator that groups adjacent to the alt-right are using genetic testing to bolster their ideology.

6.5 A Language Comparison of Reddit & 4chan

Although they both provide discussion platforms, Reddit and 4chan operate in different ways: e.g., the former requires registration, while the predominant mode of operation on the latter is via anonymous and ephemeral posting. Naturally, they also attract different sets of users and content, e.g., 4chan is typically identified as a fringe community, while, Reddit, though also hosting fringe communities, is overall a mainstream site (5th most visited in the US).

Our analysis of genetic testing on the two platforms thus far has highlighted that genetic testing is a subject which is discussed frequently; on Reddit, in subreddits ranging many aspects of the every day life of the users, on 4chan, in threads that attract an order of magnitude more posts. At the same time, on both platforms, fringe political groups express their wish to marginalize minorities using genetic testing. Next, we provide a comparison of the *language* used in the context of conversations that are likely to include genetic testing. To do so, we turn to word embeddings, specifically, word2vec [151]. Word2vec models are trained on large corpora of text, and generate a high-dimensional vector for each word that appears in the corpus; words that are used in similar context also have a closer mapping to the high-dimensional vector space. This allows us to study which words are used in similar contexts.

Methodology. We train a separate word2vec model, as per the implementation provided by [186], for each of the 19 groups of subreddits (see Figure 6.1) and 4chan's /pol/, using all of the posts made between January 1, 2016 and March 31, 2018, and

Group	# of Words in Vocabulary	Group	# of Words in Vocabulary
4chan's /pol/	31,337	Hate	40,223
Ancestry	122	Health	11,101
Animals	8,065	Legal	4,655
Children	15,858	News	32,097
Crime	11,649	Politics	41,057
Drugs	7,858	Race/Countries	46,978
Educational	23,151	Religion	12,431
Entertainment	7,743	Science	18,341
Funny	5,641	Sexes	20,743
Genetics	1,178	Other	24,767

TABLE 6.11: Words in the vocabulary of the word2vec models trained for each group of subreddits and /pol/.

June 30, 2016 and March 13, 2018, respectively. We pre-process each corpus as follows: 1) we remove special symbols, punctuation, URLs, and numbers; 2) we tokenize each word that appears on each post; and 3) we perform stemming on the words using the Porter algorithm. Next, we train word2vec models for each community on all the pre-processed posts and all words that appear at least 100 times in each corpus. We use a *context window* equal to 7, i.e., the model considers a context of up to 7 words ahead and behind the current word.

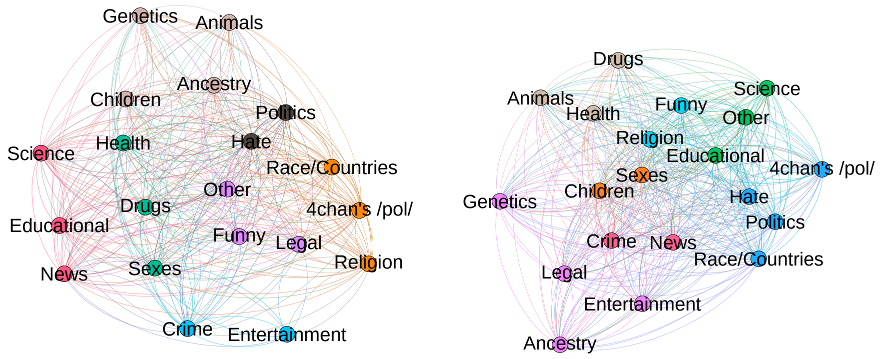
Vocabulary. Table 6.11 reports the number of words that are considered in each word2vec model. Vocabulary sizes vary greatly, e.g., from 122 in the Ancestry subreddits to 46K in Race/Culture subreddits. This is due to the fact that we only consider words that appear at least 100 times.

Training. To assess how each community discusses topics related to ethnicity and genetic testing words, we use the methodology described above and for each word2vec model, we get the 10 most similar words (based on the cosine similarities obtained from the word2vec model) for two groups of seed words: 1) 91 genetic testing keywords obtained from the list of 280 keywords (the other 189 including multiple words so we do not consider them) 2) a hand-picked set of words, namely, "white," "black," "jew," "kike," "ancestry," "dna," and "test." The latter are added aiming to assess whether ethnic terms (e.g., “white”) and genetic testing keywords (e.g., “dna”) are used in different contexts than the set of genetic keywords (e.g., “23andMe”).

Visualization. We calculate the similarity of all the possible combinations of word2vec models using the Jaccard Index scores of all the similar words for all the seed words. Then, we create two complete graphs (see Figure 6.6), one for each set of seed keywords, where nodes are the trained word2vec models and edges are weighted by the Jaccard Index score between the similar words for all the seed words. Once again, we use the community detection algorithm by [30].

When using the genetic testing keywords as seeds (Figure 6.6(a)), we find that communities about genetics, ancestry, animals, and children discuss genetic testing in very similar contexts (light brown nodes). Similarly, we find a cluster with subreddits with scientific, educational, and news content (red nodes on the left), and another related to health, drugs, and sexes (green nodes). Interestingly, the subreddits in the hate category discuss genetic testing in a similar manner as the political ones (brown nodes); this is not entirely surprising also considering that these categories have the two highest toxicity levels (cf. Figure 6.2). Also, /pol/ users seem to discuss genetic testing in a context similar to subreddits related to race/countries and religion (orange nodes). This may be because /pol/ frequently discusses Judaism (with references to Israel and the Jewish community), as well as other religions [81].

When using the set of hand-picked seed words (Figure 6.6(b)), /pol/ is similar to the hateful subreddits, as well as the subreddits about politics and race/countries (blue nodes). In other words, Hate, Politics, Race/Countries subreddits, and /pol/, use ethnic terms in conjunction with genetic testing keywords in similar contexts. Overall, the fact that that certain subreddits share language characteristics with /pol/ is particularly worrying as it may be an indicator of 4chan’s fringe ideologies propagating into more mainstream media.



(a) Genetic Testing Keywords (b) Selected Keywords

FIGURE 6.6: Graph representation of the word2vec models, using as seeds: (a) all the genetic testing keywords, (b) the terms “white,” “black,” “jew,” “kike,” “ancestry,” “dna,” and “test.”

6.6 Discussion

Direct-to-consumer (DTC) genetic testing is one of the first revolutionary technologies with the potential to transform society by improving people’s lives. Nowadays, citizens of most developed countries have easy and affordable access to a wealth of informative reports, which allow them to better understand themselves, learn about their health and their cultural heritage, and find lost relatives [33]. However, this new technology also harbors societal dangers as it is used by fringe groups as “evidence” on which to build discrimination and prejudice, and potentially increase ethnic sectarianism. Considering how information has become increasingly misused on the Web, the potential abuse of genetic testing on online platforms is not be underestimated.

Nevertheless, prior work on this topic has mostly been limited to relatively small (qualitative) studies [172, 192], which discuss how DTC genetic testing may have a negative societal impact due to their results possibly reinforcing the concept of racial privilege. In that respect, our analysis furthers this line of research by taking a large-scale, data-driven approach, which provides new insight into both the breadth and depth of the issue (of which hate speech is an important aspect). We believe that our findings broaden the discussion around DTC genetic testing and its potential misuse in furthering hateful rhetoric and ideology as we provide quantitative evidence for the prior qualitative work.

More specifically, we shed light on online discussions about genetic testing on two social networking sites, Reddit and 4chan's politically incorrect board (/pol/), which are known to provide a platform to fringe and alt-right communities. We analyzed 1.3M comments spanning 27 months using a set of 280 keywords related to genetic testing as search terms, relying on a mix of tools including Latent Dirichlet Allocation, Google's Perspective API, Perceptual Hashing, and word embeddings to identify trends, themes, and topics of discussion.

Our analysis showed that genetic testing is frequently discussed on both platforms. For instance, on /pol/, we find an order of magnitude increase in activity on threads related to genetic testing when compared to a random sample. Interestingly, images appearing along genetic testing conversations often include alt-right personalities and antisemitic memes. On Reddit, genetic testing is discussed in a wider variety of contexts, however, while there are communities building around the more positive aspects (e.g., health, cultural heritage, etc.), we also found others where conversations include racist, hateful, and misogynistic content.

Overall, we uncovered evidence of genetic testing being misused in online discussions, further ingraining and empowering genetics-based prejudice, discrimination, and even calls for genocide. For instance, comments on both /pol/ and a set of "hateful" subreddits often contain highly toxic language, with users even suggesting leveraging genetic testing tools to further marginalize or even eliminate minorities. In fact, word embeddings showed that /pol/ and certain subreddits share worrying language characteristics, which may be an indicator of 4chan's fringe ideologies spilling out to more mainstream platforms.

Chapter 7

Conclusion

Direct-to-consumer genetic testing has the potential to transform society by improving people's lives, however, it also harbors dangers as it prompts important privacy and societal concerns. With this motivation in mind, in this thesis, we presented a multidisciplinary evaluation of the privacy and societal challenges faced by the adoption of DTC genetic testing. Our research focused on whether PETs that are used for testing, storing, and sharing genomic data offer adequate short- and long-term privacy mechanisms. We also explored how the popularity of genetic testing is reflected on three social platforms focusing on the content of the posts as well as the nature of the users that discuss them. Finally, we shed light on a disturbing new phenomenon, the weaponization of DTC genetic testing by fringe political groups against minorities. Overall, the contributions of this thesis can be summarized as follows:

- We identified and discussed ten open research problems faced by the genome privacy community. Specifically, we found that several of them are unlikely to be addressed organically as they are inherently tied to the unique properties of the human genome.
- We demonstrated that DTC genetic testing is a popular topic of discussion on Twitter, Reddit, and 4chan. We found that on Twitter, it is mostly viewed under a positive light and the discourse about it is dominated by users that might have a

vested interest in its success. On Reddit, it is discussed under a variety of contexts but users are not uniformly interested in all aspects of it. On 4chan, genetic testing is discussed mostly in a toxic manner.

- We uncovered evidence of genetic testing being misused on 4chan's /pol/ and in specific subreddits by groups adjacent to fringe political agendas ingraining and empowering genetics-based prejudice and discrimination, and even calling for genocide.

7.1 High-Level Cross-Platform Examination

We also provide a comparison across the three social networks we have analyzed. This is not without challenges as 1) the three platforms differ in several functionality aspects, and 2) we have focused on different aspects of the discourse and through a slightly different set of keywords. Regarding the former, tweets may or may not be responses to other tweets, while on Reddit, a comment follows an original post and each comment can be replied to. On 4chan, all comments are responses to an original post but there is only one thread of discussion. Also, a tweet contains a maximum of 280 characters, while, on Reddit and 4chan, there is no restriction. Furthermore, Twitter and Reddit require the creation of a profile, while, 4chan comments can be and in most cases are anonymous.

Regarding the latter, we have actually attempted to conduct the same analysis on Twitter, in terms of keywords and dates, as the one done on Reddit/4chan, using a dataset containing 1% of all tweets (using Twitter's streaming API) from 2015, which only yields a set of only 35K total tweets. A preliminary study of this dataset, unfortunately, did not yield conclusive findings due to its relative limited size, and thus we do not include it.

For these reasons, it would arguably be impossible to conduct a one-to-one comparison of our three datasets. In fact, our aim is not to do so, nor to compare how the users of each platform differ with respect to their views on genetic testing, but rather to reason around some of the aspects of DTC genetic testing discourse we extract via our quantitative analysis. As a result, we set to perform a high-level cross-platform examination, focusing on three axes: topics, user engagement, and hateful content.

Topics. On Twitter, our hashtag and URL analyses show that a large part of the genetic testing discourse is generated from news and technology websites. We also find several social media marketing strategies at play, with some DTC companies employing traditional giveaways, and others promoting third-party articles about their brands. On the other hand, we find that the genetic testing discourse on Reddit is centered around 18 "mega-topics," ranging from educational and scientific content to politics, religion, crimes, and a set of subreddits strongly associated with hateful content. Interestingly, our /pol/ analysis shows that genetic testing is a popular topic of discussion, often accompanied by images and memes with clear racial or hateful connotations.

Users. On Twitter, the conversation around genetic testing is often dominated by users with a vested interest in its success, such as journalists, medical professionals, and entrepreneurs. Conversely, on Reddit, users who discuss genetic testing tend to form distinct groups ranging from enthusiasts (e.g., those who are interested in or have undergone genetic testing), to people who use genetic keywords exclusively in subreddits that discuss fringe political views. Due to the fact that most 4chan posts are anonymous, we obviously are unable to go down to the user-level. Overall, we observe a dichotomy in the type of users interested in genetic testing among all three datasets: some focus in typical uses of genetic testing (Twitter and most subreddits), while others discuss their use in worrying ways (subreddits associated with hateful content and 4chan's /pol/).

Hateful Content. Toxic content surrounding genetic testing conversations is rather sparse on Twitter, but not inexistent, despite Twitter's conduct policy which should lead to account suspension [217]. Whereas, on Reddit and /pol/, our analysis shows that genetics-based racism is rather systematic. Specifically, we uncover evidence of genetic testing being misused in online discussions, further ingraining and empowering genetics-based prejudice, discrimination, and even calls for genocide. For instance, comments on both /pol/ and a set of subreddits associated with hateful content often contain highly toxic language, with users even suggesting leveraging genetic testing tools to further marginalize or even eliminate minorities. We also find that images appearing along genetic testing conversations often include alt-right personalities and anti-semitic memes. Word embeddings reveal that certain subreddits use ethnic terms in conjunction with genetic testing keywords in the same way as /pol/, which may be an indicator of 4chan's fringe ideologies spilling out on more mainstream Web communities.

7.2 Discussion

The findings presented in this thesis have several real-world implications for the future of personal genomic testing.

Genome Privacy. First, we demonstrated that the use of PETs to protect genome privacy faces several obstacles, some of which are unlikely to be addressed organically due to the unique properties of the human genome. According to a survey we conducted with 21 genome privacy experts (see Chapter 4.5), the most important problems are the lack of long-term security, the inherent utility loss in terms of functionality, and the lack of solutions that can be applied to current genomics initiatives.

Our research indicates that for these problems to be solved alternative solutions or significant technological breakthroughs may be needed. One example of such solutions is the work by Wan et al. [222] who address the privacy-utility trade-off in emerging genomic data sharing initiatives. To do so, they rely on a game-theoretic approach which accounts for the capabilities and the behavior of the adversary, so that the defender can choose the best strategy satisfying their privacy requirements without crippling utility.

Nevertheless, as the number and size of both publicly- and privately-owned biorepositories increases [26] it is evident that, at the moment, we lack the means of adequately protecting the genomic privacy of the donors. In that respect, it is still unclear whether such an undertaking can be completely addressed by solely technological means (e.g., by inventing new cryptographic primitives) or whether a hybrid approach is necessary (e.g., a combination of cryptographic tools, access control, and policy making).

Genetic Testing Discourse. Then, we showed that genetic testing has become a popular topic of discussion on both mainstream and fringe social platforms, namely, Twitter, Reddit, and 4chan. The popularity of genetic testing as well as how it is perceived by the public is becoming increasingly important, because, besides the benefits in terms of health and wellness (both actual and anticipated), genetic testing is also often associated with concerns about privacy, ethics, the legal system, and its societal impacts [87]. Furthermore, genetic testing has even become "politicized." For example, USA President Trump has repeatedly accused Senator Elizabeth Warren of making up her Native American ancestry, leading her to publicly confirm it via genetic testing [144] and, in a 2018 interview, Senator Lindsey Graham stated it would be "terrible" if a DNA test showed he had Iranian ancestry [211].

Thus, it is important to have a good understanding not only of *what* the discourse about genetic testing is about, but also of *who* posts about genetic testing, and *how* they use the various online platforms for that. In that respect, we found that genetic testing is often discussed on Twitter by tech-savvy users who are overall interested in tech and digital health and by users and entities that might have a vested interest in its success, such as, specialist journalists, medical professionals, and entrepreneurs. We also found a number of enthusiastic users who broadcast their test results through screenshots notwithstanding possible privacy implications. On Reddit, genetic testing is being discussed in a variety of contexts, e.g., dog breeds, crime evidence, and issues related to children (e.g., adoption, pregnancy) which indicates how mainstream genetic testing has become. Furthermore, we discovered that Reddit users are not uniformly interested in all aspects of genetic testing, rather, they form groups ranging from enthusiasts (e.g., those who are interested in or have undergone genetic testing), to people who use genetic keywords exclusively in subreddits that discuss fringe political views.

Genetic Testing & Racism. This latter finding is, perhaps, the most worrying of this thesis. By focusing on specific subreddits and on 4chan's /pol/ we uncovered quantitative evidence of genetic testing being misused in online discussions by fringe groups. These groups use the technology of genetic testing as a means for further ingraining and empowering genetics-based prejudice, discrimination, and even calling for genocide. Our findings are particularly timely as recent events indicate that those interested in societal disruption have successfully seized upon technological innovations and used them in ways that were not intended by their creators. Specifically, information has been increasingly weaponized, including by state actors, to sew racial discontent [208] and even instigate public health crises [37].

Impact. This thesis revolved around the privacy and societal impacts of personal genomic testing. It investigated whether the privacy of those who contribute their data to genetic testing services and public initiatives can be adequately protected both in the short- and the long-term, how personal genomic testing is reflected online, and whether it is being misused. Thus, this thesis impacts various entities, such as researchers, policy makers, companies, and individuals.

First, we demonstrated that current genome privacy tools developed for testing, storing, and sharing genomic data are unable to adequately protect the privacy of the users, both in the short- and the long-term. Our evaluation showed that the overwhelming majority of the proposed techniques aiming to scale up to large genomic datasets need to opt for weaker security guarantees or weaker models. Furthermore, one serious challenge stems from lack of long-term security protection, which is difficult to address as available cryptographic tools are not suitable for this goal. We believe that our work will inspire researchers to find alternative and/or creative ways of addressing some of the non-trivial open problems we identify in this thesis.

The results of this thesis have also real-world implications on the DTC genetic testing industry and policy makers. Specifically, we showed that DTC genetic testing is being used by fringe groups online to ingrain and empower genetics-based prejudice and discrimination, and even call for genocide. Considering that platforms like Facebook and Twitter have begun to be held accountable when their services enable harmful behavior [229], we believe that our work will motivate policy makers to address the DTC genetic testing industry and legislate so that these companies consider the potential abuse of their services and attempt to find ways of minimizing this behavior. Finally, it is our hope that this thesis will raise awareness about the potential privacy and societal implications of DTC genetic testing to the general public.

7.3 Limitations

Like any research study, the work presented in this thesis is not without limitations. In Chapter 4, we conducted a systematic analysis of genome privacy research by reviewing a total of 25 *representative* papers. When deciding whether to include one paper over another, we preferred papers published in venues that are more visible to the privacy-enhancing technologies community or that have been cited significantly more, as they arguably have a stronger influence on the community over time. However, since the reviewed papers were handpicked, the final selection is bound to be biased. Nevertheless, we stress that our methodology constitutes a best-effort approach and that reviewing *all* genome privacy papers would be infeasible. Overall, if we added or replaced one paper with another, the main takeaways would not be considerably altered.

Then, in Chapters 5 and 6 we conducted three large-scale quantitative studies to better understand *what* the discourse about genetic testing is about, *who* posts about genetic testing, and *how* they use the various online platforms for that. Analyzing the content of social platforms can help us answer these questions, however, regardless of how popular a platform is, its content is not necessarily representative of the world population. Thus, each dataset carries a demographical bias that depends on the nature of the platform. More specifically, the geographic analysis of our Twitter dataset showed that the top 5 countries in our dataset are mostly English-speaking ones (see Figure 5.5). Also, both datasets are biased as the keywords used to create them are in English and thus are bound to better reflect English-speaking countries than the rest of the world, and because certain platforms are more popular in certain countries than others (e.g., at the time of the analysis Reddit was the 5th most visited site in USA [181]).

Furthermore, even though our datasets span relatively long time periods of time (approximately 2 years each), they should be treated as "time snapshots," i.e., repeating the same analysis on the same platform a few years later may yield different results, especially when considering the rapid progress in genomics. Also, our datasets do not span the same time periods. Namely, our Twitter dataset starts on January 1, 2015 and ends on July 31, 2017, our Reddit dataset spans from January 1, 2016 to March 31, 2018, and our 4chan dataset spans from June 30, 2016 to March 13, 2018.

Also, in Chapter 6 we leverage on Google's Perspective API [175] to measure the toxicity of the comments we study. However, effective and timely detection of abusive speech is a difficult task. APIs like Perspective can be evaded with simple text or word modifications [111, 125], and might be biased, e.g., against African-American English [60].

Finally, our Twitter dataset (see Chapter 5) was collected using our own custom Python crawler instead of using Twitter's "Sample Stream" API which returns a random sample of all public tweets and then filtering the results using our list of genetic testing keywords. This is intentional as during the early stages of this research the latter method yielded an underwhelming number of tweets (i.e., less than 5,000). Nevertheless, we note that our crawler uses a library that is a wrapper of Twitter's official API and that, when we applied for an API key, we described in our form our intended methodology and our application was accepted.

7.4 Future Work

We believe that the methodological approach used in all three of our research studies is generalizable and can be used in the future to compare and contrast how the field evolves. Specifically, in regards to genome privacy, one can repeat the selection process and use the same systematization criteria used in Chapter 4 to assess the progress in the field. Furthermore, considering that datasets containing comments from social platforms are "time snapshots" we anticipate that our methodology will be used in a similar manner in a few years time to compare how the perception of DTC genetic testing has evolved.

Furthermore, there are several interesting research directions that could be explored in the future. First, we believe that new genome manipulation techniques, such as CRISPR [141], should be extensively studied regarding their potential impact on both security (e.g., causing harm) and privacy (e.g., editing genomes to recover from data exposure or to hinder re-identification). Then, we believe that the methodology we introduced in Chapter 4 should be also used in a broader biomedical context, such as considering users' health-related as well as genomic data.

Also, we believe that our large-scale studies (Chapters 5 and 6) will be useful to researchers in Social Sciences and Genetics that are interested in better understanding the impact of this revolutionary technology to society. In that respect, we believe that more similar studies should be conducted on other social platforms, both mainstream and non-mainstream.

Finally, considering that previous qualitative studies [172, 192] demonstrate how the commercialization of genetic testing may have a negative societal impact, and since our study provides quantitative data on the matter, we believe that the next natural step is to examine whether genetic *ancestry* testing has an (indirect) effect on the levels of racism and discrimination online. Naturally, such correlation is not easy to identify and it may require a mixed-methods methodological approach (e.g., interviews with people adjacent to the far-right), but our work provides a stepping stone toward this.

Bibliography

[1] 23andMe. *About Us*. https://mediacenter.23andme.com/company/about-us/. 2020.

[2] *95+ Social Networking Sites You Need To Know About*. https://makeawebsitehub.com/social-media-sites/. 2020.

[3] Sofiane Abbar, Yelena Mejova, and Ingmar Weber. "You Tweet What You Eat: Studying Food Consumption Through Twitter". In: *CHI*. 2015.

[4] ADL. *Pepe the Frog*. https://www.adl.org/education/references/hate-symbols/pepe-the-frog. 2019.

[5] Rakesh Agrawal et al. "Order Preserving Encryption for Numeric Data". In: *ACM SIGMOD*. 2004, pp. 563–574.

[6] Mete Akgün et al. "Privacy Preserving Processing of Genomic Data: A Survey". In: *Journal of Biomedical Informatics* 56 (2015), pp. 103–111.

[7] AncestryDNA. *Company Facts*. https://www.ancestry.com/corporate/about-ancestry/company-facts. 2020.

[8] Anna Jones. *Brief Summary of Breed Specific Legislation*. https://www.animallaw.info/intro/breed-specific-legislation-bsl. 2017.

[9] Euan A Ashley. "Towards Precision Medicine". In: *Nature Reviews Genetics* 17.9 (2016), p. 507.

[10] Erman Ayday et al. "Privacy-Preserving Processing of Raw Genomic Data". In: *DPM*. 2014, pp. 133–147.

[11] Erman Ayday et al. “Protecting and Evaluating Genomic Privacy in Medical Tests and Personalized Medicine”. In: *ACM WPES*. 2013, pp. 95–106.

[12] Erman Ayday et al. “The Chills and Thrills of Whole Genome Sequencing”. In: *Computer* (2013).

[13] Erman Ayday et al. “Whole Genome Sequencing: Revolutionary Medicine or Privacy Nightmare?” In: *IEEE Computer* (2015).

[14] Al Aziz et al. “Secure and Efficient Multiparty Computation on Genomic Data”. In: *IDEAS*. 2016, pp. 278–283.

[15] Md Momin Al Aziz et al. “Privacy-Preserving Techniques of Genomic Data – A Survey”. In: *Briefings in Bioinformatics* September (2017), pp. 1–9.

[16] Michael Backes et al. “Membership Privacy in MicroRNA-Based Studies”. In: *ACM CCS*. 2016, pp. 319–330.

[17] Michael Backes et al. “Privacy in Epigenetics: Temporal Linkability of microRNA Expression Profiles”. In: *USENIX Security Symposium*. 2016, pp. 1223–1240.

[18] Pierre Baldi et al. “Countering GATTACA: Efficient and Secure Testing of Fully-Sequenced Human Genomes”. In: *ACM CCS*. 2011, pp. 691–702.

[19] Ludovic Barman et al. “Privacy threats and practical solutions for genetic risk tests”. In: *IEEE Security and Privacy Workshops*. 2015, pp. 27–31.

[20] BBC. *James Watson: Scientist loses titles after claims over race*. `https://www.bbc.co.uk/news/world-us-canada-46856779`. 2019.

[21] Sharon Begley. *House Republicans Would Let Employers Demand Workers' Genetic Test Results*. `https://www.statnews.com/2017/03/10/workplace-wellness-genetic-testing/`. 2017.

[22] Doron M. Behar, Mait Metspalu, Yael Baran, et al. “No Evidence from Genome-Wide Data of a Khazar Origin for the Ashkenazi Jews”. In: *Human Biology* 85.6 (2013).

[23] Anat Ben-David and Ariadna Matamoros-Fernandez. “Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain”. In: *IJOC* 10 (2016), pp. 1167–1193.

[24] Michael Bernstein et al. “4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community”. In: *ICWSM* (2011).

[25] Raghav Bhaskar et al. “Discovering Frequent Patterns in Sensitive Data”. In: *KDD*. 2010, pp. 503–512.

[26] Biobanking. *10 Largest Biobanks in the World*. https://www.biobanking.com/10-largest-biobanks-in-the-world/. 2018.

[27] Joe Bish. *Vice News. Examining the Right Wing British Blowhards Using YouTube to Prove Everybody Wrong*. https://bit.ly/2qN4SMG. 2016.

[28] Marina Blanton and Mehrdad Aliasgari. “Secure Outsourcing of DNA Searching via Finite Automata”. In: *DBSec*. 2010, pp. 49–64.

[29] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent Dirichlet Allocation”. In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[30] Vincent D Blondel et al. “Fast Unfolding of Communities in Large Networks”. In: *JSTAT* 2008.10 (2008), P10008.

[31] Joseph Bonneau et al. “The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes”. In: *IEEE Security & Privacy*. 2012.

[32] Eric Boodman. *White Nationalists Are Flocking To Genetic Ancestry Tests – But Many Don't Like Their Results*. https://read.bi/2DEaQYY. 2016.

[33] Katie Sullivan Borrelli. *PressConnects. DNA Tales: These People Found Long-Lost or Never-Known Relatives*. https://bit.ly/2FxDye2. 2018.

[34] Joppe W Bos, Kristin Lauter, and Michael Naehrig. “Private Predictive Analysis on Encrypted Medical Data”. In: *Journal of Biomedical Informatics* 50 (2014), pp. 234–243.

[35] Russell Brandom. *New Documents Reveal Which Encryption Tools the NSA Couldn't Crack*. https://www.theverge.com/2014/12/28/7458159/encryption-standards-the-nsa-cant-crack-pgp-tor-otr-snowden. 2014.

[36] Ferdinand Brasser et al. “Software Grand Exposure: SGX Cache Attacks are Practical”. In: (2017).

[37] David A Broniatowski et al. “Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate”. In: *American journal of public health* 108.10 (2018).

[38] Pete Burnap et al. “Tweeting the Terror: Modelling the Social Media Reaction to the Woolwich Terrorist Attack”. In: *Social Network Analysis and Mining* 4.1 (2014).

[39] Paul R Burton et al. “Size Matters: Just How Big is BIG? Quantifying Realistic Sample Size Requirements For Human Genome Epidemiology”. In: *International Journal of Epidemiology* 38.1 (2008), pp. 263–273.

[40] John M Butler. “Short Tandem Repeat Typing Technologies Used In Human Identity Testing”. In: *Biotechniques* 43.4 (2007), pp. 2–5.

[41] Mustafa Canim, Murat Kantarcioglu, and Bradley Malin. “Secure Management of Biomedical Data With Cryptographic Hardware”. In: *IEEE Transactions on Information Technology in Biomedicine* 16.1 (2012), pp. 166–175.

[42] Timothy Caulfield and Amy L McGuire. “Direct-To-Consumer Genetic Testing: Perceptions, Problems, and Policy Responses”. In: *Annual Review of Medicine* 63 (2012), pp. 23–33.

[43] Patricia A Cavazos-Rehg et al. “Hey Everyone, I'm Drunk. An Evaluation Of Drinking-Related Twitter Chatter”. In: *JSAD* 76.4 (2015).

[44] Eshwar Chandrasekharan et al. “The Bag of Communities”. In: *CHI*. 2017, pp. 3175–3187.

[45] Eshwar Chandrasekharan et al. “You Can’t Stay Here: The Efficacy of Reddit’s 2015 Ban Examined Through Hate Speech”. In: *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW (2017), p. 31.

[46] Despoina Chatzakou et al. “Mean Birds: Detecting Aggression and Bullying on Twitter”. In: *Proceedings of the 2017 ACM on web science conference*. ACM. 2017, pp. 13–22.

[47] Despoina Chatzakou et al. “Measuring #GamerGate: A Tale of Hate, Sexism, and Bullying”. In: *WWW 2017*. 2017.

[48] Feng Chen et al. “PRINCESS: Privacy-Protecting Rare Disease International Network Collaboration via Encryption through Software Guard extensionS”. In: *Bioinformatics* 33.6 (2017), pp. 871–878.

[49] Yangyi Chen et al. “Large-Scale Privacy-Preserving Mapping of Human Genomic Sequences on Hybrid Clouds.” In: *NDSS*. 2012.

[50] Jung Hee Cheon, Miran Kim, and Kristin Lauter. “Homomorphic Computation of Edit Distance”. In: *FC*. 2015, pp. 194–212.

[51] Benny Chor et al. “Private Information Retrieval”. In: *FOCS*. 1995, pp. 41–50.

[52] Peter Chow-White et al. “‘Warren Buffet Is My Cousin’: Shaping Public Understanding of Big Data Biotechnology, Direct-To-Consumer Genomics, and 23andMe on Twitter”. In: *Information, Communication & Society* 21.3 (2018), pp. 448–464.

[53] Emily Christofides and Kieran O’Doherty. “Company Disclosure and Consumer Perceptions of the Privacy Implications of Direct-To-Consumer Genetic Testing”. In: *New Genetics and Society* 35.2 (2016), pp. 101–123.

[54] Matthew Claxton. *Abbotsford News. Former Langley Libertarian candidate detained in Italy*. https://bit.ly/2PUIQWC. 2017.

[55] Ellen Wright Clayton et al. “The Law of Genetic Privacy: Applications, Implications, and Limitations”. In: *Journal of Law and the Biosciences* 6.1 (2019), pp. 1–36.

[56] EW Clayton et al. “A Systematic Literature Review of Individuals’ Perspectives on Privacy and Genetic Information in the United States”. In: *PLoS ONE* 13.10 (2018).

[57] Glen Coppersmith, Mark Dredze, and Craig Harman. “Quantifying Mental Health Signals In Twitter”. In: *CLPsych*. 2014.

[58] Nick Couldry and Jun Yu. “Deconstructing Datafication’s Brave New World”. In: *New Media & Society* 20.12 (2018), pp. 4473–4491.

[59] BF Darst et al. “Perceptions of Genetic Counseling Services in Direct-To-Consumer Personal Genomic Testing”. In: *Clinical genetics* 84.4 (2013), pp. 335–339.

[60] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. “Racial Bias in Hate Speech and Abusive Language Detection Datasets”. In: *Proceedings of the Third Workshop on Abusive Language Online*. 2019.

[61] Thomas Davidson et al. “Automated Hate Speech Detection and the Problem of Offensive Language”. In: *ICWSM*. 2017.

[62] Munmun De Choudhury and Sushovan De. “Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity”. In: *ICWSM*. 2014.

[63] Munmun De Choudhury et al. “Predicting Depression via Social Media”. In: *ICWSM*. 2013.

[64] Emiliano De Cristofaro. “Genomic Privacy and the Rise of a New Research Community”. In: *IEEE Security & Privacy* 12.2 (2014), pp. 80–83.

[65] Emiliano De Cristofaro, Sky Faber, and Gene Tsudik. “Secure Genomic Testing With Size-and Position-Hiding Private Substring Matching”. In: *ACM WPES*. 2013, pp. 107–118.

[66] Emiliano De Cristofaro, Paolo Gasti, and Gene Tsudik. “Fast and Private Computation of Cardinality of Set Intersection and Union”. In: *CANS*. 2012, pp. 218–231.

[67] Emiliano De Cristofaro, Kaitai Liang, and Yuruo Zhang. “Privacy-Preserving Genetic Relatedness Test”. In: *GenoPri*. 2016.

[68] Emiliano De Cristofaro and Gene Tsudik. “Practical Private Set Intersection Protocols With Linear Complexity”. In: *FCDS*. 2010, pp. 143–159.

[69] Fabio Del Vigna et al. “Hate Me, Hate Me Not: Hate Speech Detection on Facebook”. In: *CEUR Workshop*. 2017, pp. 86–95.

[70] Mentari Djatmiko et al. “Secure Evaluation Protocol for Personalized Medicine”. In: *ACM WPES*. 2014, pp. 159–162.

[71] DNARomance. *Online Dating Based On Science*. `https://www.dnaromance.com/`. 2018.

[72] DNAstack. *Beacon Network*. `https://beacon-network.org/`. 2020.

[73] Yevgeniy Dodis, Leonid Reyzin, and Adam Smith. “Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data”. In: *Eurocrypt*. 2004, pp. 523–540.

[74] Cynthia Dwork et al. “Calibrating Noise to Sensitivity in Private Data Analysis”. In: *TCC*. Vol. 3876. 2006, pp. 265–284.

[75] Cynthia Dwork et al. “Robust Traceability from Trace Amounts”. In: *FOCS*. 2015, pp. 650–669.

[76] Yaniv Erlich and Arvind Narayanan. “Routes for Breaching and Protecting Genetic Privacy”. In: *Nature Reviews Genetics* 15.6 (2014), pp. 409–421.

[77] Evangelos Evangelou and John Ioannidis. “Meta-Analysis Methods for Genome-Wide Association Studies and Beyond”. In: *Nature Reviews Genetics* 14.6 (2013), pp. 379–389.

[78] FDA. *FDA allows marketing of first direct-to-consumer tests that provide genetic risk information for certain conditions*. https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm551185.htm. 2017.

[79] Ari Feldman. *23andMe Backpedals On Khazar Theory But The 'Alt-Right' Eats It Up, Anyway*. http://forward.com/news/national/381500/23andme-backpedals-on-khazar-theory-but-the-alt-right-eats-it-up-anyway/. 2017.

[80] Stephen E Fienberg, Aleksandra Slavkovic, and Caroline Uhler. "Privacy Preserving GWAS Data Sharing". In: *ICDM Workshops*. 2011, pp. 628–635.

[81] Joel Finkelstein et al. "A Quantitative Approach to Understanding Online Antisemitism". In: *CoRR* abs/1809.01644 (2018).

[82] Claudia Flores-Saviaga, Brian C. Keegan, and Saiph Savage. "Mobilizing the Trump Train: Understanding Collective Action in a Political Trolling Community". In: *ICWSM*. 2018.

[83] Antigoni Maria Founta et al. "A Unified Deep Learning Architecture for Abuse Detection". In: *Proceedings of the 10th ACM Conference on Web Science*. ACM. 2019, pp. 105–114.

[84] James H Fowler, Jaime E Settle, and Nicholas A Christakis. "Correlated Genotypes in Friendship Networks". In: *Proceedings of the National Academy of Sciences* 108.5 (2011), pp. 1993–1997.

[85] Matthew Fredrikson et al. "Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing". In: *USENIX Security*. 2014, pp. 17–32.

[86] Amanda Froelich. *True Activist. This is What Americans Will Look like by 2050*. https://bit.ly/2vpAIEH. 2014.

[87] FTC. *DNA test kits: Consider the privacy implications*. https://www.consumer.ftc.gov/blog/2017/12/dna-test-kits-consider-privacy-implications. 2017.

[88] Borko Furht. “Cloud Computing Fundamentals”. In: *Handbook of Cloud Computing*. Springer, 2010, pp. 3–19.

[89] GEDmatch. `https://en.wikipedia.org/wiki/GEDmatch`. 2019.

[90] Genetics Home Reference. *What is the Precision Medicine Initiative?* `https://ghr.nlm.nih.gov/primer/precisionmedicine/initiative`. 2019.

[91] Genomics England. `https://www.genomicsengland.co.uk`. 2017.

[92] GenoPri. *International Workshop on Genome Privacy and Security*. `http://www.genopri.org`. 2018.

[93] Reza Ghasemi et al. “Private and Efficient Query Processing on Outsourced Genomic Databases”. In: *IEEE Journal of Biomedical and Health Informatics* 21.5 (2016), pp. 1466–1472.

[94] Global Alliance for Genomics and Health. `https://www.ga4gh.org/`. 2017.

[95] Oded Goldreich and Rafail Ostrovsky. “Software protection and simulation on oblivious RAMs”. In: *Journal of the ACM* 43.3 (1996), pp. 431–473.

[96] Sara Goodwin, John D McPherson, and W Richard McCombie. “Coming of Age: Ten Years of Next-Generation Sequencing Technologies”. In: *Nature Reviews Genetics* 17.6 (2016), pp. 333–351.

[97] Yuriy Gorodnichenko et al. *Social Media, Sentiment and Public Opinions: Evidence from #Brexit and #USElection*. National Bureau of Economic Research. 2018.

[98] Ellen M Greytak, CeCe Moore, and Steven L Armentrout. “Genetic Genealogy for Cold Case and Active Investigations”. In: *Forensic Science International* (2019).

[99] Melissa Gymrek et al. “Identifying Personal Genomes by Surname Inference”. In: *Science* 339.6117 (2013), pp. 321–324.

[100] Marcus Hähnel, Weidong Cui, and Marcus Peinado. “High-Resolution Side Channels for Untrusted Operating Systems”. In: *USENIX*. 2017.

[101] Kay Hamacher, Jean Pierre Hubaux, and Gene Tsudik. "Genomic Privacy (Dagstuhl Seminar 13412)". In: *Dagstuhl Reports*. Vol. 3. 2014.

[102] Katie EJ Hann et al. "Awareness, Knowledge, Perceptions, and Attitudes Towards Genetic Testing for Cancer Risk Among Ethnic Minority Groups: A Systematic Review". In: *BMC public health* 17.1 (2017), p. 503.

[103] Liz Harley. *White House hosts Precision Medicine Initiative Summit*. http://www.frontlinegenomics.com/white-house-hosts-precision-medicine-initiative-summit/. 2016.

[104] Amy Harmon. *New York Times. Why White Supremacists Are Chugging Milk (and Why Geneticists Are Alarmed)*. https://nyti.ms/2Afg4Ho. 2018.

[105] Carmit Hazay and Yehuda Lindell. *Efficient Secure Two-Party Protocols: Techniques and Constructions*. Springer Science & Business Media, 2010.

[106] Dan He et al. "Identifying Genetic Relatives Without Compromising Privacy". In: *Genome Research* 24.4 (2014), pp. 664–672.

[107] Helix. *Genotyping vs. Sequencing, and What They Mean For You*. https://blog.helix.com/dna-technologies-genotyping-vs-sequencing/. 2020.

[108] Gabriel Emile Hine et al. "Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web". In: *ICWSM*. 2017.

[109] Nils Homer et al. "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays". In: *PLoS Genet* 4.8 (2008), e1000167.

[110] Farhad Hormozdiari et al. "Privacy Preserving Protocol for Detecting Genetic Relatives Using Rare Variants". In: *Bioinformatics* 30.12 (2014), pp. 204–211.

[111] Hossein Hosseini et al. "Deceiving Google's Perspective API Built for Detecting Toxic Comments". In: *arXiv:1702.08138* (2017).

[112] Homa Hosseinmardi et al. “Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network”. In: *SocInfo*. 2015.

[113] Zhicong Huang et al. “A Privacy-Preserving Solution for Compressed Storage and Selective Retrieval of Genomic Data”. In: *Genome Research* 26.12 (2016), pp. 1687–1696.

[114] Zhicong Huang et al. “GenoGuard: Protecting Genomic Data Against Brute-Force Attacks”. In: *IEEE Security & Privacy*. 2015, pp. 447–462.

[115] Jean Pierre Hubaux et al. “Genomic Privacy (Dagstuhl Seminar 15431)”. In: *Dagstuhl Reports*. Vol. 5. 2016.

[116] Mathias Humbert et al. “Reconciling Utility With Privacy in Genomics”. In: *ACM WPES*. 2014, pp. 11–20.

[117] *iDash Privacy & Security Workshop 2019 – Secure Genome Analysis Competition*. http://www.humangenomeprivacy.org/2019/. 2019.

[118] Hae Kyung Im et al. “On Sharing Quantitative Trait GWAS Results in an Era of Multiple-Omics Data and the Limits of Genomic Privacy”. In: *The American Journal of Human Genetics* 90.4 (2012), pp. 591–598.

[119] Business Insider. *White nationalists are flocking to genetic ancestry tests — but many don't like their results*. https://www.businessinsider.com/white-nationalists-genetic-ancestry-tests-dont-like-results-2017-8?r=UK. 2017.

[120] National Human Genome Research Institute. *The Cost of Sequencing a Human Genome*. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost. 2020.

[121] *International Workshop on Genome Privacy and Security*. https://genopri.org. 2019.

[122] Internet Live Stats. *Internet Users by Country (2016)*. http://www.internetlivestats.com/internet-users-by-country/. 2017.

[123] ISOGG. *List of DNA Testing Companies*. https://isogg.org/wiki/List_of_DNA_testing_companies. 2017.

[124] Karthik A Jagadeesh et al. "Deriving Genomic Diagnoses Without Revealing Patient Genomes". In: *Science* 357.6352 (2017), pp. 692–695.

[125] Edwin Jain et al. "Adversarial Text Generation for Google's Perspective API". In: *International Conference on Computational Science and Computational Intelligence (CSCI)*. 2018, pp. 1136–1141.

[126] Xiaoqian Jiang et al. "A Community Assessment of Privacy Preserving Techniques for Human Genomes". In: *BMC Medical Informatics and Decision Making* 14.Suppl 1 (2014), S1.

[127] Aaron Johnson and Vitaly Shmatikov. "Privacy-Preserving Data Exploration in Genome-Wide Association Studies". In: *ACM KDD*. 2013, pp. 1079–1087.

[128] Ari Juels and Thomas Ristenpart. "Honey Encryption: Security Beyond the Brute-Force Bound". In: *Eurocrypt*. 2014, pp. 293–310.

[129] Liina Kamm et al. "A New Way To Protect Privacy in Large-Scale Genome-Wide Association Studies". In: *Bioinformatics* 29.7 (2013), pp. 886–893.

[130] Murat Kantarcioglu et al. "A Cryptographic Approach to Securely Share and Query Genomic Sequences". In: *IEEE Transactions on Information Technology in Biomedicine* 12.5 (2008), pp. 606–617.

[131] Nikolaos Karvelas et al. "Privacy-Preserving Whole Genome Sequence Processing Through Proxy-Aided ORAM". In: *ACM WPES*. 2014, pp. 1–10.

[132] Anna Kasunic and Geoff Kaufman. ""At Least the Pizzas You Make Are Hot": Norms, Values, and Abrasive Humor on the Subreddit r/RoastMe". In: *ICWSM*. 2018.

[133] Brandon Keim. *10 years on, the genome revolution is only just beginning*. https://www.wired.com/2010/03/genome-at-10/. 2010.

[134] Sheharbano Khattak et al. "SoK: Making Sense of Censorship Resistance Systems". In: *Proceedings on Privacy Enhancing Technologies* 2016.4 (2016), pp. 37–61.

[135] Chiea Chuen Khor et al. "Genome-Wide Association Study Identifies FCGR2A as a Susceptibility Locus for Kawasaki Disease". In: *Nature Genetics* 43.12 (2011), pp. 1241–1246.

[136] Miran Kim and Kristin Lauter. "Private Genome Analysis Through Homomorphic Encryption". In: *BMC Medical Informatics and Decision Making* 15.5 (2015), S3.

[137] Evan Klinger and David Starkweather. `http://www.phash.org/`. 2018.

[138] Haewoon Kwak et al. "What Is Twitter, A Social Network Or A News Media?" In: *WWW*. 2010.

[139] Eric S Lander et al. "Initial Sequencing and Analysis of the Human Genome". In: *Nature* 409.6822 (2001), pp. 860–921.

[140] Kristin Lauter, Adriana López-Alt, and Michael Naehrig. "Private Computation on Encrypted Genomic Data". In: *Latincrypt*. 2014, pp. 3–27.

[141] Heidi Ledford. "CRISPR, The Disruptor". In: *Nature* (2015).

[142] Kristina Lerman et al. "Emotions, Demographics and Sociability in Twitter Interactions". In: *ICWSM*. 2016.

[143] Ninghui Li et al. "Membership Privacy: A Unifying Framework for Privacy Definitions". In: *ACM CCS*. 2013, pp. 889–900.

[144] Annie Linskey. *The Boston Globe. Warren Releases Results of DNA Test*. `https://bit.ly/2Chey99`. 2018.

[145] Christoph Lippert et al. "Identification of Individuals by Trait Prediction Using Whole-Genome Sequencing Data". In: *Proceedings of the National Academy of Sciences* 114.38 (2017), pp. 10166–10171.

[146] Stephen Marche. *The Guardian. Swallowing the Red Pill: A Journey to the Heart of Modern Misogyny*. https://bit.ly/2Chey99. 2016.

[147] Adam Marcus et al. “Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration”. In: *CHI*. 2011.

[148] Paul Marks. *Submarine Internet Cables Are a Gift for Spooks*. https://www.newscientist.com/article/dn23752-submarine-internet-cables-are-a-gift-for-spooks/. 2013.

[149] Deborah Mascalzoni et al. “Informed consent in the genomics era”. In: *PLoS Medicine* 5.9 (2008).

[150] Paul J McLaren et al. “Privacy-Preserving Genomic Testing in the Clinic: A Model Using HIV Treatment”. In: *Genetics In Medicine* 18.8 (2016), pp. 814–822.

[151] Tomas Mikolov et al. “Distributed Representations of Words and Phrases and Their Compositionality”. In: *NIPS*. 2013.

[152] Richard A. Mills. “Pop-up Political Advocacy Communities on Reddit.com: SandersForPresident and The Donald”. In: *AI and Society* 33.1 (2018), pp. 39–54.

[153] Alexandros Mittos, Bradley Malin, and Emiliano De Cristofaro. “Systematizing Genome Privacy Research: A Privacy-Enhancing Technologies Perspective”. In: *Proceedings on Privacy Enhancing Technologies* 2019.1 (2019), pp. 87–107.

[154] Alexandros Mittos et al. “"And We Will Fight For Our Race!" A Measurement Study of Genetic Testing Conversations on Reddit and 4chan”. In: *Thirteenth International AAAI Conference on Web and Social Media* (2020).

[155] Megan Molteni. *Wired. The Creepy Genetics Behind the Golden State Killer Case*. https://bit.ly/2HYECJE. 2018.

[156] Vishal Monga and Brian L. Evans. “Perceptual Image Hashing Via Feature Points: Performance Evaluation and Tradeoffs”. In: *IEEE Transactions on Image Processing* (2006).

[157] National Human Genome Research Institute. *The Cost of Sequencing a Human Genome*. https://www.genome.gov/sequencingcosts/. 2017.

[158] Nature. *The Ethics of Catching Criminals Using Their Family's DNA*. https://www.nature.com/articles/d41586-018-05029-9. 2018.

[159] Gonzalo Navarro. "A Guided Tour To Approximate String Matching". In: *ACM Computing Surveys* 33.1 (2001), pp. 31–88.

[160] Muhammad Naveed et al. "Controlled Functional Encryption". In: *ACM CCS*. 2014, pp. 1280–1291.

[161] Muhammad Naveed et al. "Privacy In The Genomic Era". In: *ACM Computing Surveys* 48.1 (2015), pp. 1–43.

[162] Science News. *What I actually learned about my family after trying 5 DNA ancestry tests*. https://www.sciencenews.org/article/family-dna-ancestry-tests-review-comparison. 2020.

[163] Daiva E Nielsen, Sarah Shih, and Ahmed El-Sohemy. "Perceptions of Genetic Testing for Personalized Nutrition: A Randomized Trial of DNA-based Dietary Advice". In: *Lifestyle Genomics* 7.2 (2014), pp. 94–104.

[164] NIH. *The All of Us Research Program*. https://allofus.nih.gov/. 2017.

[165] NIH. *What Is Genetic Ancestry Testing?* https://ghr.nlm.nih.gov/primer/dtcgenetictesting/ancestrytesting. 2019.

[166] Niu, Xia-mu and Jiao, Yu-hua. "An Overview of Perceptual Hashing". In: *Acta Electronica Sinica* (2008).

[167] Alicia L Nobles et al. ""Is This an STD? Please Help!" Online Information Seeking for Sexually Transmitted Diseases on Reddit". In: *ICWSM*. 2018.

[168] Alexandra Olteanu et al. "The Effect of Extremist Violence on Hateful Speech Online". In: *ICWSM*. 2018.

[169] Raphael Ottoni et al. “Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination”. In: *WebSci*. 2018.

[170] Katie Palmer. *Another Personal Genetics Company Is Sharing Client Data*. `https://www.wired.com/2015/07/another-personal-genetics-company-selling-client-data/`. 2015.

[171] Aaron Panofsky and Joan Donovan. “Genetic ancestry testing among white nationalists: From identity repair to citizen science”. In: *Social studies of science* (2019).

[172] Aaron Panofsky and Joan Donovan. *When Genetics Challenges a Racist's Identity: Genetic Ancestry Testing among White Nationalists*. `https://osf.io/preprints/socarxiv/7f9bc/`. 2017.

[173] European Parliament. “General Data Protection Regulation (EU) 2016/679”. In: *Official Journal of the European Union* L119 (2016), pp. 1–88.

[174] Michael J Paul and Mark Dredze. “You Are What You Tweet: Analyzing Twitter for Public Health”. In: *ICWSM*. 2011.

[175] Perspective. `https://www.perspectiveapi.com/`. 2019.

[176] Anthony A Philippakis et al. “The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery”. In: *Human Mutation* 36.10 (2015), pp. 915–921.

[177] Andelka M Phillips. *Data on Direct-to-Consumer Genetic Testing and DNA Testing Companies*. 10.5281/zenodo.1175800. 2018.

[178] Andelka M Phillips. “DTC Genetics for Ancestry, Health, Love and More: A View of the Business and Regulatory Landscape”. In: *Applied & Translational Genomics* 8 (2016).

[179] Nugroho Dwi Prasetyo et al. “On the Impact of Twitter-Based Health Campaigns: A Cross-Country Analysis of Movember”. In: *EMNPL*. 2015.

[180] Jean Louis Raisaro et al. "Addressing Beacon Re-Identification Attacks: Quantification and Mitigation of Privacy Risks". In: *Journal of the American Medical Informatics Association* (2017), ocw167.

[181] Reddit. https://www.redditinc.com/press. 2020.

[182] Elspeth Reeve. *Alt-Right Trolls Are Getting 23andMe Genetic Tests To 'Prove' Their Whiteness*. https://www.vice.com/en_us/article/vbygqm/alt-right-trolls-are-getting-23andme-genetic-tests-to-prove-their-whiteness. 2016.

[183] Elspeth Reeve. *Vice News – White Nonsense: Alt-right trolls are arguing over genetic tests they think prove their whiteness*. http://bit.ly/2DhP90h. 2016.

[184] Elspeth Reeve. *Vice News. White Nonsense*. https://bit.ly/2DhP90h. 2016.

[185] Genetics Home Reference. *What is genetic ancestry testing?* https://ghr.nlm.nih.gov/primer/dtcgenetictesting/ancestrytesting. 2020.

[186] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". In: *NLPFrameworks*. 2010.

[187] David Reich. *New York Times. How Genetics Is Changing Our Understanding of 'Race'*. https://nyti.ms/2pUxF0w. 2018.

[188] MIT Technology Review. *More Than 26 Million People Have Taken an At-Home Ancestry Test*. https://www.technologyreview.com/2019/02/11/103446/more-than-26-million-people-have-taken-an-at-home-ancestry-test/. 2019.

[189] Manoel Horta Ribeiro et al. "Auditing Radicalization Pathways on YouTube". In: (2020), pp. 131–141.

[190] Manoel Horta Ribeiro et al. "Characterizing and Detecting Hateful Users on Twitter". In: *ICWSM*. 2018.

[191] Aja Romano. *DNA profiles from ancestry websites helped identify the Golden State Killer suspect*. https://www.vox.com/2018/4/27/17290288/golden-state-killer-joseph-james-deangelo-dna-profile-match. 2018.

[192] Wendy D Roth and Biorn Ivemark. “Genetic Options : The Impact of Genetic Ancestry Testing on Consumers’ Racial”. In: *American Journal of Sociology* 124.1 (2018), pp. 150–184.

[193] Tina Hesman Saey. *What I Actually Learned About My Family After Trying 5 DNA Ancestry Tests*. https://bit.ly/2zaUIKy. 2018.

[194] Xinghua Shi and Xintao Wu. “An Overview of Human Genetic Privacy”. In: *Annals of the New York Academy of Sciences* 1387.1 (2017), pp. 61–72.

[195] Kana Shimizu, Koji Nuida, and Gunnar Rätsch. “Efficient Privacy-Preserving String Search and an Application in Genomics”. In: *Bioinformatics* 32.11 (2016), pp. 1652–1661.

[196] Suyash S Shringarpure and Carlos D Bustamante. “Privacy Risks from Genomic Data-Sharing Beacons”. In: *The American Journal of Human Genetics* 97.5 (2015), pp. 631–646.

[197] Leandro Silva et al. “Analyzing the Targets of Hate in Online Social Media”. In: *ICWSM*. 2016.

[198] Sean Simmons and Bonnie Berger. “Realizing Privacy Preserving Genome-Wide Association Studies”. In: *Bioinformatics* 32.9 (2016), pp. 1293–1300.

[199] David Sims. *The Battle Over Adult Swim’s Alt-Right TV Show*. https://bit.ly/2g06PPK. 2016.

[200] Nigel P. Smart et al. *Algorithms, Key Size and Parameters Report*. https://www.enisa.europa.eu/publications/algorithms-key-size-and-parameters-report-2014/at_download/fullReport. 2014.

[201] SoccerGenomics. *Unlock The Player Within You*. https://www.soccergenomics.com/. 2018.

[202] Ebrahim M Songhori et al. "Tinygarble: Highly Compressed and Scalable Sequential Garbled Circuits". In: *IEEE Security and Privacy*. 2015, pp. 411–428.

[203] João Sá Sousa et al. "Efficient and Secure Outsourcing of Genomic Data Storage". In: *BMC Medical Genomics* 10.2 (2017), p. 46.

[204] SPLC. *Atomwaffen Division*. https://www.splcenter.org/fighting-hate/extremist-files/group/atomwaffen-division. 2019.

[205] SPLC. *Male Supremacy*. https://www.splcenter.org/fighting-hate/extremist-files/ideology/male-supremacy. 2017.

[206] Liam Stack. *New York Times. Alt-Right, Alt-Left, Antifa: A Glossary of Extremist Language*. https://nyti.ms/2uGOTV5. 2017.

[207] Björn Stade et al. "GrabBlur: A Framework to Facilitate the Secure Exchange of Whole-Exome and-Genome SNV Data Using VCF Files". In: *BMC Genomics* 15.4 (2014), S8.

[208] Leo G Stewart, Ahmer Arif, and Kate Starbird. "Examining Trolls and Polarization with a Retweet Network". In: *Proceedings of ACM WSDM: Workshop on Misinformation and Misbehavior Mining on the Web*. 2018.

[209] Barbara E Stranger et al. "Relative Impact Of Nucleotide And Copy Number Variation On Gene Expression Phenotypes". In: *Science* 315.5813 (2007), pp. 848–853.

[210] Latanya Sweeney, Akua Abu, and Julia Winn. "Identifying Participants in the Personal Genome Project by Name". In: *arXiv:1304.7605* (2013).

[211] The Washington Post. *Lindsey Graham says it would be, 'like, terrible' if a DNA test found that he had Iranian heritage*. https://www.washingtonpost.com/politics/lindsey-graham-says-it-would-be-like-terrible-if-a-dna-test-found-that-he-had-iranian-heritage/2018/10/16/0de362e6-d169-11e8-83d6-291fcead2ab1_story.html. 2018.

[212] Mike Thelwall et al. "Sentiment Strength Detection In Short Informal Text". In: *Journal of the American Society for Information Science and Technology* 61.12 (2010), pp. 2544–2558.

[213] Ian Thomson. *Microsoft Researchers Smash Homomorphic Encryption Speed Barrier*. https://www.theregister.co.uk/2016/02/09/researchers_break_homomorphic_encryption/. 2016.

[214] Florian Tramèr et al. "Differential Privacy with Bounded Priors: Reconciling Utility and Privacy in Genome-Wide Association Studies". In: *ACM CCS*. 2015, pp. 1286–1297.

[215] Juan Ramón Troncoso-Pastoriza, Stefan Katzenbeisser, and Mehmet Celik. "Privacy Preserving Error Resilient DNA Searching Through Oblivious Automata". In: *ACM CCS*. 2007, pp. 519–528.

[216] *Tweepy*. https://www.tweepy.org/. 2020.

[217] Twitter. *Hateful Conduct Policy*. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy. 2019.

[218] Caroline Uhlerop, Aleksandra Slavković, and Stephen E Fienberg. "Privacy-Preserving Data Sharing for Genome-Wide Association Studies". In: *The Journal of Privacy and Confidentiality* 5.1 (2013), pp. 137–166.

[219] Onur Varol et al. "Online Human-Bot Interactions: Detection, Estimation, and Characterization". In: *ICWSM*. 2017.

[220] Isabel Wagner. "Genomic Privacy Metrics: A Systematic Comparison". In: *IEEE Security & Privacy Workshops*. 2015, pp. 50–59.

[221] Zhiyu Wan et al. "Controlling the Signal: Practical Privacy Protection of Genomic Data Sharing Through Beacon Services". In: *BMC Medical Genomics* 10.2 (2017), p. 39.

[222] Zhiyu Wan et al. "Expanding Access to Large-Scale Genomic Data While Promoting Privacy: A Game Theoretic Approach". In: *The American Journal of Human Genetics* 100.2 (2017), pp. 316–322.

[223] Rui Wang et al. "Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study". In: *ACM CCS*. 2009, pp. 534–544.

[224] Shuang Wang et al. "Genome Privacy: Challenges, Technical Approaches to Mitigate Risk, and Ethical Considerations in the United States". In: *Annals of the New York Academy of Sciences* 1387.1 (2017), pp. 73–83.

[225] Shuang Wang et al. "HEALER: Homomorphic computation of ExAct Logistic rEgRession for secure rare disease variants analysis in GWAS". In: *Bioinformatics* 32.2 (2016), pp. 211–218.

[226] Xiao Shaun Wang et al. "Efficient Genome-Wide, Privacy-Preserving Similar Patient Query Based on Private Edit Distance". In: *ACM CCS*. 2015, pp. 492–503.

[227] Chris Welch and Sara Ganim. *CNN. White Supremacist Richard Spencer: 'We reached tens of millions of people' with video*. https://cnn.it/2T7z5D8. 2016.

[228] Danielle Welter et al. "The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations". In: *Nucleic Acids Research* 42.D1 (2013), pp. D1001–D1006.

[229] Queenie Wong. *Facebook's Privacy Mishaps: Zuckerberg Could Be Held Accountable, Report Says*. https://cnet.co/2VDJUlu. 2019.

[230] Wei Xie et al. “SecureMA: Protecting Participant Privacy in Genetic Association Meta-Analysis”. In: *Bioinformatics* 30.23 (2014), pp. 3334–3341.

[231] Lei Xu et al. “Privacy Preserving Large Scale DNA Read-Mapping in MapReduce Framework Using FGPAs”. In: *2014 24th International Conference on Field Programmable Logic and Applications (FPL)*. IEEE. 2014, pp. 1–4.

[232] Andrew Yao. “How to Generate and Exchange Secrets”. In: *FOCS*. 1986, pp. 162–167.

[233] Masaya Yasuda et al. “Secure Pattern Matching Using Somewhat Homomorphic Encryption”. In: *ACM CCSW*. 2013, pp. 65–76.

[234] Fei Yu et al. “Scalable Privacy-Preserving Data Sharing Methodology for Genome-Wide Association Studies”. In: *Journal of Biomedical Informatics* 50 (2014), pp. 133–141.

[235] Savvas Zannettou et al. “On the Origins of Memes by Means of Fringe Web Communities”. In: *IMC*. 2018.

[236] Savvas Zannettou et al. “The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources”. In: *IMC*. 2017.

[237] Christop Zauner. `http://www.phash.org/docs/pubs/thesis_zauner.pdf`. 2010.

[238] ZDNet. *IBM Warns Of Instant Breaking of Encryption by Quantum Computers: 'Move Your Data Today'*. `https://www.zdnet.com/article/ibm-warns-of-instant-breaking-of-encryption-by-quantum-computers-move-your-data-today/`. 2018.

[239] Zephoria. *Twitter Statistics*. `https://zephoria.com/twitter-statistics-top-ten/`. 2020.

[240] Yihua Zhang, Marina Blanton, and Ghada Almashaqbeh. "Secure Distributed Genome Analysis for GWAS and Sequence Comparison Computation". In: *BMC Medical Informatics and Decision Making* 15.5 (2015), S4.

[241] Yuchen Zhang et al. "FORESEE: Fully Outsourced secuRe gEnome Study basEd on homomorphic Encryption". In: *BMC Medical Informatics and Decision Making* 15.5 (2015), S5.

[242] Yongan Zhao et al. "Choosing Blindly but Wisely: Differentially Private Solicitation of DNA Datasets for Disease Marker Discovery". In: *Journal of the American Medical Informatics Association* 22.1 (2014), pp. 100–108.