1 Introduction

Stories of family deceits and deceptions have become commonplace in a media receptive to personal tales of triumph and tragedy. A distinguished geneticist learns in his mature years that his mother, while married to his legal father, had a secret affair that begat him. A best-selling author discovers that her paternal DNA was from a medical student serving as a sperm donor and not her legal father, who traced her ancestry deep into Eastern Europe. A woman who, as a newborn, was left in a bag abandoned in the foyer of a Brooklyn apartment building searches for her biological parents 23 years later. These revelations are the result of the millennial DNA ancestry revolution.

I first learned about the surge of public interest in DNA genealogy when I was researching a book on forensic applications of DNA. After the British scientist Alec Jeffreys created a way to use DNA sequences to provide unique identifiers of individuals in 1983, which later became known as DNA fingerprinting, police agencies throughout the world adopted the DNA identification system. In the United States, the forensic DNA system of personal identification was the basis of a databank known as CODIS (Combined DNA Index System), established in 1990. The US Federal Bureau of Investigation (FBI) created CODIS as a pilot project involving 14 states and local laboratories before it became national. The FBI has a database containing over 12 million DNA profiles and similar databases exist worldwide. The early techniques of gene sequencing used by the FBI were based on markers called short tandem repeats, or STRs (see Chapter 5), which provided no information about a person's appearance or genealogy.

Familial and Historical Genealogy

The tools of genetic sequencing were turned into a commercial opportunity when new genetic markers were developed and the sequencing methods improved. Two strands of interest among people active in exploring genome sequencing were genealogy and health. Those interested in familial genealogy wish to learn about their family ancestry over a few generations. Those interested in historical genealogy want to find genetic connections over hundreds or even thousands of years, utilizing all the tools of historical and genetic research. Someone may, for example, be interested in whether their family tree can be traced to the Roman Empire.

Medical genome sequencing, which predated genealogical sequencing, searches for specific mutations that are linked to diseases. The markers sequenced on the genome are much more specific than genealogical sequencing.

Traditionally, genealogical research – whether familial or historical – has largely been considered a hobby or a family recreational activity. The techniques required for familial genealogy include archival research involving sources such as the federal census, birth and death certificates, probate records, gravestones, obituaries, tax records, church and military documents, immigration records, and all the tangible physical records a person leaves behind when they are gone. With the digitization of documents, these searches have become easier and available to more people.

The Birth of Genome Sequencing

Sequencing the DNA of an organism took off in the 1980s in the fields of medicine and criminal justice after Fred Sanger, Walter Gilbert, and Paul Berg shared the Nobel Prize for their development of DNA sequencing methods. The first bacterial virus (bacteriophage) sequenced in its entirety was reported in 1977. The first animal virus completely sequenced was SV40 in 1978. These viruses had small genomes, were easy to grow and purify, and had been studied as models for many years, which explains why they were sequenced first. The first human virus fully sequenced was the polio virus, in 1981.

DNA sequencing and genetic databases also became tools of anthropologists and population geneticists, who used them to study ancient population movements. By analyzing specific DNA changes (called polymorphisms) in groups of people, scientists were able to track migrations of populations across the globe.

The commercial DNA ancestry sector began in 2000 with the launch of Family Tree DNA, the first of dozens of companies to capture public interest in genealogy. While he was investigating his European ancestry, Bennett Greenspan, the company's founder and CEO, learned that scientists were able to use DNA to track a person's ancestry. He understood there was a void in the marketplace for such applications in genealogical searches. Within a couple of decades, the commercial ancestry sector was occupied by around 70 companies. In order to apply the most current and sophisticated methods of genomic genealogy, companies hired geneticists, anthropologists, statisticians, and experts in data analytics.

Goal for the Book

As I immersed myself in the scientific literature and popular books and articles about DNA ancestry, I came to realize that it took more than a basic understanding of genetics to understand how one's ancestry could be read from one's genome. It was a significant departure from traditional genealogical methods which involve the construction of family trees encompassing several past generations. In contrast, DNA ancestry can take one back many generations to people whose names are not known but who occupied certain regions of the world and share some segments of DNA with the customer. Where traditional genealogical methods involving the search of family documents is straightforward, DNA ancestry is complex and requires investment in equipment and expertise.

When I began looking into the stages involved in developing an ancestry DNA profile of an individual, I realized I was looking into black boxes that I could barely decipher. Forensic DNA methods have a universally accepted and validated methodology because the technology is simpler and was funded by the government, but the same cannot be said for company-developed ancestry DNA profiles. The literature on DNA ancestry tests was

bifurcated between technical scientific studies and popular magazine and newspaper articles. The scientific studies are out of reach for most readers and the popular stories are deficient in explaining the science.

My goal in writing this book is to decipher the process of DNA ancestry testing and to demystify the elusive technical components while exploring the applications of genealogical ancestry beyond that of creating family networks. To begin, I recognized that there are a number of terms that appear in the scientific, commercial, and popular writings about the role of DNA in understanding biogeographical populations (groups of people with distinct genetic markers inhabiting regions of the world) and their descendants. Terms such as ancestry, genealogy, pedigree, descent, lineage, and genetic inheritance are among the most frequently used.

The Varieties of Ancestry

In an essay titled "What is ancestry?" geneticists Iain Mathieson and Aylwyn Scally made some important distinctions such as that between genealogical ancestry and genetic ancestry. Genealogical ancestry refers to identifiable ancestors in your family tree, which is constructed from historical documents such as those in public records, as well as family lore. If you search back n generations, you will have 2^n ancestors. For example, your great-grandparents are three generations away from you, and your family tree exhibits $2^3 = 8$ great-grandparents. The term "pedigree" refers to how all your genealogical ancestors are related to one another: great-grandfather, second cousin, third cousin once removed, etc. Genealogical ancestry has its limitations because few people have comprehensive knowledge of their ancestors beyond a small number of generations, for which they may not even have records. Assuming a generation of 25 years, on average, 250 years or 10 generations ago there existed, in theory, 2^{10} or 1,024 ancestors for each one of us.

Genetic ancestry refers to the people in your past who contributed to the composition of your genome beyond the 50 percent from your parents and 25 percent from your grandparents. For many generations in the past, the genealogical family will have many people who do not share genetic sequences because a descendant inherited the DNA sequences of some ancestors but not those of others. Therefore, a person's genetic ancestry

consists of a small part of one's genealogical ancestry. One way to understand the differences between genealogical and genetic ancestry is that full siblings have identical genealogical ancestry but differ in their genetic ancestry, because they inherit different chromosomal segments from their parents. Two siblings have the same parents but have not inherited exactly the same DNA from them. Rather, two siblings have, on average, 50 percent of the same DNA as each other.

Traditional genealogical research seeks to construct family trees of individuals, exhibiting the pedigrees (relations between your genealogical ancestors) of descent. Genetic ancestry allows scientists to compare individual genomes with the average genome of a reference sample of some population, which may not be a random or representative sample, but a convenience sample that the ancestry company was able to obtain. While you might have no ancestral link with most members of the population, you might share certain population-specific markers indicating the region of your descent. Thus, genetic ancestry has led to population ancestry. What geneticists and ancestry DNA companies mean by ancestry is the genetic similarity between individuals and populations.

Cultural ancestry is another category of how people relate to their genealogy. Native American ancestry is based on whether someone was embedded in the cultural traditions of a tribe, not on their DNA or on the construction of a family tree. Native American cultural studies scholar Kim TallBear argues in her book *Native American DNA* that DNA connections are no substitute for kinship relations. Much of kinship and tribal citizenship is biological, but not in ways captured by genetic lineages.

Consider a person with a single Native American great-great-great-grandparent. They might not have inherited any Native American chromosome segments, so their genetic ancestry would be 0 percent Native American. Yet, if they were brought up in a Native American tribe, that is their cultural ancestry.

The practice of genealogy has been popularized by television series such as Britain's *Who Do You Think You Are?* and the US Public Broadcasting System series *Finding Your Roots*. These programs have contributed to the commercial market for genealogy. What are these programs tapping into? Cultural

studies scholar Julie Rak's answer to the question of why so many people around the world have become interested in genealogy is that "doing genealogy is about 'doing kinship,' a way to facilitate connection between the living and the dead, to construct identity not just from one's own experience but from knowledge of one's ancestors, to work through grief and loss."

Structure of the Book

The book is organized around 11 thematic chapters. Chapter 2 discusses the business behind DNA ancestry and how gathering genomic data has allowed companies to create a bifurcated business model of genealogy and health, where the same company collects both types of information. Chapter 3 explores the origin of early populations and discusses the "Out of Africa" theory and its significance in understanding genetic diversity in human populations. I begin to explore the science behind DNA ancestry testing in Chapter 4, which introduces the reader to the concept of genetic markers. Some of the early patents submitted for DNA ancestry tests provide unusual clarity about the science used in the process. While not all the patent applications were awarded, they give us an insight into how early inventors conceptualized the methods of using DNA for disease risk diagnosis, for reading phenotypes, and for ancestry inference.

Chapter 5 examines different markers used in ancestry testing from mono-parental DNA markers (markers that appear exclusively on either the male or female genome), such as on the Y chromosome of the male and the female mitochondrial DNA, to bi-parental markers (which appear on both male and female genomes) situated on the autosomes (all the chromosomes except the sex chromosomes). In Chapter 6, I discuss the reference panels used by DNA ancestry companies. Markers on the consumer DNA sample are compared to markers on reference panels, which are supposed to represent populations in different regions of the world. These reference panels tend to be proprietary for each company, although they may in part utilize public data. Chapter 7 examines how a person's DNA is compared to population reference panels.

A critical component in DNA ancestry testing is reading the thousands of markers on the donor's DNA sample. Chapter 8 discusses the development of the microarray for analyzing the DNA of an ancestry customer and identifying

the DNA markers in the sample. With the development of the microarray, the cost of analyzing a DNA sample was dramatically reduced, making it possible to charge approximately $100 for an estimate of ancestry.

The growth of the DNA ancestry sector has also had repercussions in criminal justice and provided a new method of finding criminals from crime-scene evidence when DNA matches on national databases did not reveal suspects. Chapter 9 discusses new police methods for finding felons through open-access DNA databases.

Chapter 10 explores some of the ethical and privacy issues in commercial ancestry tests, including whether they reinscribe race into the scientific vernacular and reinforce genetic essentialism, and how one person's DNA ancestry test affects other members of one's family. It also discusses why, for some people, the results of their DNA ancestry tests provide validation and assurance of their personal identity. This chapter explores how people use the DNA ancestry tests to establish their ethnicity and when this genealogical information is advantageous to them on social or civil rights grounds.

Many people are now known to have discovered unknown relatives or mistaken relatives through their DNA ancestry results. Chapter 11 selects some iconic stories to illustrate the elation and heartbreak of discovering the truth about family. Chapter 12 examines the accuracy, consistency, and validation of DNA ancestry testing. It explores why test results among different companies vary and why twins tested by the same company have given different results. Chapter 13 concludes the book.

2 The Business of DNA Ancestry

Over a period of 20 years, family genetic genealogy, through the purchase of consumer ancestry testing kits, has been one of the fastest growing family activities of this generation. Citing data from the International Society of Genetic Genealogy, the *Washington Post* reported in 2017 that 8 million people worldwide were involved with recreational genomics. It is estimated that by 2019 about 25 million people had signed up for a DNA ancestry test offered by one of the dozens of companies that have entered this marketplace. The kits are sent to a person's home with return packaging that includes a reservoir for depositing saliva or swabs for sampling cheek cells. The *MIT Technology Review* predicted that by 2021 there would be 100 million consumers of ancestry DNA services.

In 2010, researchers reported 38 companies worldwide had entered the home DNA ancestry marketplace. Six years later, 74 such companies were competing, and by the end of 2019, 61 ancestry testing companies were identified by the International Society of Genetic Genealogy. Some of the earlier companies were either absorbed by competitors or disappeared.

The First DNA Ancestry Company

The first company offering direct to consumer (DTC) genetic ancestry DNA tests was Family Tree DNA (FTD), which was incorporated in the year 2000. Initially, it used mitochondrial DNA and Y chromosome DNA for ancestry testing. DNA Print Genomics offered the first genomic (autosomal) ancestry

test in 2002, which soon became the standard for other companies. Autosomal tests utilize the entire human genome.

After several years, FTD partnered with National Geographic, which had founded the Genographic Project in 2005, the goal of which was to collect DNA samples throughout the globe to understand the patterns of human migration. By 2019, the Genographic Project claimed over 1 million participants in 140 countries. FTD needed the worldwide DNA samples to advance its commercial venture for paid consumer ancestry testing.

FTD was founded by Bennett Greenspan, who had a childhood interest in genealogy. In seeking to trace his ancestors from Poland, he learned that some had emigrated to Buenos Aires, Argentina. After reading that Thomas Jefferson's descendants could be traced by DNA, he contacted geneticist Michael Hammer, a professor at the University of Arizona whose name had been on a DNA ancestry publication known to Greenspan. That's when he realized there was a business opportunity in using DNA to trace ancestry. After selling his photographic supply business he turned his avocation interest in genealogy into a commercial venture.

Greenspan collected DNA samples from people in various regions of the world and used existing publicly funded, open-access DNA databases of different ethnicities. He signed up a group of Sephardic Jews from Seattle, Washington, for DNA tests that allowed him to build a Sephardic DNA database. This was a database of opportunity built through Greenspan's contacts. After purchasing some smaller genetic testing companies, in 2011 FTD renamed itself Gene By Gene, and within eight years had a staff of 150 people.

Growth of the Digital Ancestry Sector

Of the 246 DTC companies listed in 2018, 74 offered ancestry services. Others focused on legal paternity, maternity, grandparent identification, and sibling identification testing.

In 2010, the DNA ancestry industry was valued at $15 million; and six years later, in 2016, that value had grown to $173 million. The two giant ancestry DNA companies, by virtue of the number of tests kits sold, are AncestryDNA,

founded in 2007 as an offshoot of National Geographic, located in Lehi, Utah, and 23andMe, founded in 2006 by Linda Avey and Anne Wojcicki (funded by Google), located at Mountain View, California. Three years after incorporation, AncestryDNA had acquired 1 million customers. It started a health division in 2015 and began collecting customer health information. By November 2019, AncestryDNA reported that it had distributed 14 million kits. Meanwhile, 23andMe established a partnership with GlaxoSmithKline to leverage genetic data for drug development. The business model for a number of companies, including 23andMe, involved selling the DNA data to pharmaceutical companies. Drug development benefited from the large genome databases that could provide information on which mutations were antagonistic to specific drug therapies and how rare these mutations were. Figure 2.1 shows the rapid growth of autosomal tests administered by five ancestry companies between 2012 and 2020.

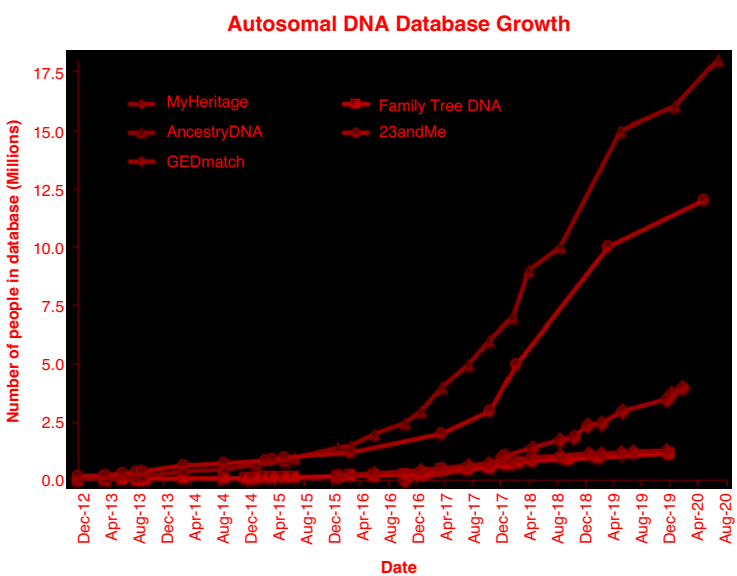

Figure 2.1 Growth in autosomal DNA ancestry testing. Source: Leah Larkin, 2020. Autosomal DNA testing growth. https://thednageek.com/dna-tests.