Genomics in the AWS® Cloud

Analyzing Genetic Code Using Amazon Web Services

Catherine Vacher • David Wall

WILEY

Genomics in the AWS® Cloud

Genomics in the AWS® Cloud

Analyzing Genetic Code Using Amazon Web Services

Catherine Vacher
David Wall

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada and United Kingdom.

ISBN: 978-1-119-57337-1

ISBN: 978-1-119-57341-8 (ebk.)

ISBN: 978-1-119-57340-1 (ebk.)

Library of Congress Control Number: 2023933145

Cover image: © antoniokhr/Getty Images

Cover design: Wiley

For Roland and Floriane

Acknowledgments

This book represents a lot of work over the course of several years, and we are grateful to the people who helped us complete it.

First among these is Emma Rath, a skilled scientist and good friend who reviewed and contributed to several chapters.

We acknowledge the skill and monumental patience of the team at Wiley, particularly John Sleeva and Kenyon Brown. We appreciate all that they did to bring this book into being.

We are grateful to Carole Jelen, our literary agent, who helped mold this book and also showed extraordinary patience.

We also want to extend special thanks to our family, including Philippe Vacher and Lee and Anne Wall. They were inspirational and always have been.

Thank you, all.

—Catherine Vacher and David Wall

About the Authors

Catherine Vacher, PhD, is a research fellow at the Brain and Mind Centre at the University of Sydney. There, she models complex systems related to healthcare and advises policymakers on the most effective use of limited public health resources.

She previously worked as a scientist at the Garvan Institute of Medical Research. There, she worked in bioinformatics, whole-genome sequencing, transcriptomics, molecular modeling, and metagenomics, mostly with respect to cancers.

She also enjoyed a long career at IBM, in Europe and in the Asia-Pacific region.

She enjoys theater and opera as well as travel.

David Wall is a consultant specializing in the AWS cloud. His projects have included complex database and machine learning solutions as well as telecommunications systems.

He is an avid cyclist and works as a volunteer firefighter.

Contents at a Glance

Contents

Introduction

Welcome to *Genomics in the AWS Cloud*!

From its title, you can conclude that this book is about two things: genomics (the science of sequencing and interpreting genetic data) and Amazon Web Services (one of the three big hosted computing platforms). *Genomics in the AWS Cloud*, therefore, is meant to appeal either to people from a biology background who want to learn how to do genomics work with AWS or to people with a computer background who want to find out how to apply their skills to genomics.

Both of these areas, genomics and cloud computing, are evolving constantly, and practically no one can claim to be completely *au fait* with either. This book, therefore, aims at not one but two separate moving targets. Our goal as authors is not to teach you everything there is to know about AWS and genomics—or even about the intersection of the two fields—but rather to show you the following:

- Enough of the general concepts of cloud computing and genomics that you understand the problems to be solved and the technologies available to work on those problems
- Enough specifics to enable you to work through actual genomics tasks and see results

Who Should Read This Book

This book is intended for people who aren't content to use commercial genome sequencing services and want to do their own analysis. We walk you through the process of getting raw data from a blood sample via a lab and then using the AWS services to analyze it—learning which genes are present in the sample and what they might say about you and your health. This will enable you to

investigate aspects of your genome that commercial services don't explore because they are not allowed to give medical advice.

As well, this book is suited to people who want to learn about the AWS cloud and want to structure their study around a useful field—genomics.

Genomics

At the core of genomics is genome sequencing, which is the process of taking some biological material, such as blood or tissue, and converting it to pure information. This is a complicated process that combines the traditional work of a biologist (which is to say, manipulating actual cells in a "wet lab" environment) with information technology. Cells go into the process; a computer-readable data file comes out.

Genome sequencing took a long time to figure out. Crude, expensive methods were first employed in the 1970s and 1980s. More automated methods became available in the late 1990s, and these enabled the sequencing of relatively simple organisms: yeasts, bacteria, and a nearly microscopic nematode worm (*Caenorhabditis elegans*, long popular in biology labs as an experimental subject). Uncomplicated plants (notably *Arabidopsis thaliana*—a European weed with a particularly small genome) and modest insects (*Drosophila melanogaster*—the fruit fly), both longtime standard experimental subjects, soon followed around the turn of the century.

Two biologists described the first draft human genome sequence in an article in the journal *Science* in 2001. Scientists have worked to refine the human genome sequence since then and also have worked to sequence the genomes of thousands of other organisms.

Key to their work has been a continuous drop in the cost of full-genome sequencing. The first human genome sequence in 2001 cost roughly 2.7 billion U.S. dollars to produce—it required funding of the sort only national governments could provide. Within less than a decade, by 2005, the cost had fallen by four orders of magnitude to something like $1 million—still quite a lot. At this writing, in 2022, it is possible to have a human genome sequenced for less than the cost of a high-end smartphone, and plenty of companies are attracting funding for their plans to bring the cost to less than $100. By the time the asteroid 99942 Apophis makes its closest approach to Earth in 2029, sequencing a full genome will almost certainly cost about the same as the simplest routine medical blood test costs now. The cost of knowing everything about your genetic makeup will be trivial (assuming Apophis doesn't render this unimportant, which the latest reports assure us it won't).

The dramatic fall of the price of genome sequencing, from billions of dollars at the turn of the twenty-first century to a few hundred dollars today, makes

it possible for almost all of us to explore our genetic makeup. While we all, as humans, share practically all of our genetic code (upwards of 99.9 percent), the differences make all the difference.

The tiny fraction of our individual genomes that differ from other humans is what accounts for whether we are male or female, all of our physical characteristics, many of our personality traits, and our propensity to health or various kinds of disease.

The availability of low-cost genome sequencing has revolutionized medical and pharmaceutical research and is started to change the practice of medicine. It also enables us to start to understand the building blocks of life and how much, or rather how little, we differ from other life forms.

Genomics in the AWS Cloud is about discovering and studying those differences and learning from them. But there is another part to the equation, which is to say the other set of tools and techniques identified in the title.

Cloud Computing and AWS

Almost in parallel with the advances in genomics that took place between 1995 and the present day, so-called cloud computing evolved enormously during the same time period and today represents a standard way of designing, deploying, and operating information processing systems.

Now, the idea of computing resources that are not local to the people who need them is not new at all. The earliest commercial and scientific computers were, of course, mainframes that were shared across many users—and more than a few of these remain in place today. Servers in organizational or co-location data centers, providing storage and computing resources to privileged users and the general public, have long been part of information technology. In the case of mainframes and client-server systems, users access remote computer systems (often not knowing or caring where they actually are). Functionally, that's *cloud computing*, and it's not a new thing.

What is new is the ease with which modern cloud computing platforms allow rapid construction and cost-effective use of complex and powerful systems. You can quickly set up elaborate workflows, test them with minimal computing power, and then scale them up enormously when it's time for a production run. More or less, you pay only for the computing power you use, and there are ways to schedule the use of processor cycles for times of low demand, when computing is cheaper. With the exception of storage and data transfer—meaning machine images that can be turned into working compute resources, as well as input and output data—systems configured in the cloud can cost practically nothing when they are not doing useful work. Such efficient use of expensive resources isn't possible with on-premises or traditionally hosted solutions.

There are three main players in the cloud-computing industry.

- Amazon Web Services (AWS)
- Google Cloud Platform (GCP)
- Microsoft Azure

Each of them has its points of strength and weakness, and those relative merits are beyond the scope of this publication. Most organizations mix and match parts of each anyway to take advantage of relative technical superiorities and to maintain leverage with vendors. Suffice it to say that we chose to do our genomics work on AWS.

Considering that genome analysis is essentially a workflow to be carried out on storage and computing resources, AWS is well suited to the job. Here are some of the tools we will use:

- Elastic Compute Cloud (EC2) for building and running the Linux servers that actually run the software required for genome analysis
- Elastic Block Storage (EBS) for maintaining updated disk images, ready to attach to a working machine when needed
- Simple Storage Service (S3) for storing input and output files when ready access to them is needed
- Simple Workflow Service (SWF) for automating processes
- Glacier when files need to be archived at low cost
- Identity and Access Management (IAM) for maintaining security and appropriate user privileges

What You'll Learn from This Book

This book is intended to educate its readers in two areas: the science of genomics and the technology of Amazon Web Services. The idea is that you use the latter as a tool to explore the former.

Since it's unlikely that many readers are familiar with both genomics and AWS, this book is meant to teach you either subject—or both if you are familiar with neither—and how they work together.

How This Book Is Organized

Here is a quick introduction to each chapter in this book. You can skip directly to the parts that interest you most, or you can read from beginning to end to get a complete picture.

Chapter 1: Why Do Genome Analysis Yourself When Commercial Offerings Exist? This chapter explains what turnkey commercial services (such as 23&Me) exist and what they are good for. It then explains what they do not do and why you might want to do your own genomics work.

Chapter 2: A Crash Course in Molecular Biology This chapter brings you up to speed on the state of biological science as it pertains to genomics. Use this to refresh your knowledge from school or to gain a new understanding.

Chapter 3: Obtaining Your Genome This chapter explains how to get a sample of your blood suitable for sequencing and how to send it off to a lab for conversion into raw genome data.

Chapter 4: The Bioinformatics Workflow This chapter approaches the sequencing process from a biological point of view, explaining how one step of data processing feeds into the next to ultimately produce the results you want.

Chapter 5: AWS Services for Genome Analysis This chapter represents a deep dive into the AWS services you can use for genomics work. If you know biology but aren't familiar with AWS, you might want to begin here.

Chapter 6: Building Your Environment in the AWS Cloud This chapter goes into more detail about how to set up AWS services for genomics work.

Chapter 7: Linux and AWS Command-Line Basics for Genomics All major bioinformatics tools run under Linux, so you'll need to understand that operating environment and how to get things working together within it.

Chapter 8: Processing the Sequencing Data This chapter explains how to get from raw sequencing data to useful information about a genome.

Chapter 9: Visualizing the Genome This chapter introduces you to tools you can use to depict genomic information graphically, enabling you to understand it better.

Chapter 10: Containerizing Your Workflow on the Desktop This chapter explains how to use Docker containers, both locally and in AWS, to process data efficiently and scalably.

Chapter 11: Variants and Applications This chapter explores certain aspects of genome analysis, allowing you to dig deeper into the information you have derived from your sample.

Chapter 12: Cancer Genomics This chapter discusses the analysis of somatic mutations, specifically those found in cancerous tumors, although the workflow could also be applied to any tissue in the body.

How to Use This Book

You can approach this book by reading straight through, from beginning to end. That is probably the best way to approach it if you have knowledge of neither genomics nor AWS. Alternatively, if you feel you have a good handle on one or the other, you can focus on those chapters that cover your weak area and then study the sections on how AWS can be used to create and automate a genomics workflow.

Our Story

This book is not about us, but it's probably fair to explain to you, our reader, who we are and where we are coming from as we write this book.

We are a married couple and the parents of two children. Catherine is a working scientist and bioinformatician, attached to a medical research group in Sydney. David is an information technology and communications consultant who designs and builds AWS solutions, among other things.

A key fact is that one of our two children isn't alive anymore. Our daughter, Floriane, died of a sudden cardiac arrest in April 2015. She was nine years old at the time.

Floriane's death and its consequences are not the subject of this book. However, the way she lived and died inspired us to take our knowledge of genomics and cloud computing and look inward at our family. We wanted to know what happened to Floriane and what bearing that had on the rest of us. One way we honor her memory is by attempting to understand what happened to her. As well, we want to keep the rest of us—particularly her younger brother, Roland—safe.

Floriane was a fit and apparently healthy girl of nine. She had no ongoing medical concerns and was by all appearances developing normally. She was active and happy.

One evening, after dinner, she approached David on the sofa and asked him to chase her around the house—one of her favorite activities that they had indulged in many times before. David chased Floriane around, theatrically waving his arms. Floriane giggled excitedly and ran a few meters. David caught her and flipped her upside-down, which excited her further because she knew it led to tickling. At that point, she went completely limp.

David was confused. He thought he'd accidentally hit Floriane's head on something. Catherine, nearby, thought it was a seizure, even though Floriane had never had one previously. It quickly became clear that Floriane was in cardiac arrest.

Not having a defibrillator on hand, we called for an ambulance and did our best in the seconds before she died with manual cardiopulmonary resuscitation

(CPR) that proved ineffective. Paramedics showed up with a defibrillator and adrenaline, and used both, but it was all over by then. Floriane died at home, a few minutes after the onset of her cardiac arrest.

(Again, it's not the subject of this book, but please take this opportunity to contemplate the fleetingness of life and to appreciate the people you love. We'll wait.)

And so we were left with questions: What just happened? Why did it happen? Is it going to happen to any of the rest of us? There were many more questions, far more existential in nature, and those questions remain, but those questions are subjects for other venues. Looking at the situation in a limited way, our daughter had just died, completely unexpectedly, and we wanted to know why.

Speculation on the cause of Floriane's death began almost immediately. She hadn't been visibly sick at all. She hadn't exhibited the symptoms—fever and so on—of an infection, such as meningitis. She was far too mature—nine years old, tall, and apparently fit—to have succumbed to the strange and almost random things that claim the lives of babies and are grouped under the catchall descriptor Sudden Infant Death Syndrome (SIDS). She didn't have any known allergies. She wasn't taking any medicines. She hadn't suffered an injury, hadn't eaten anything unusual, hadn't traveled anywhere dangerous—she hadn't done anything outside the realm of what is normal for a fifth-grade girl living in Australian suburbia. And yet she was no longer alive.

The emergency-room doctor offered a broad guess that turned out to be right: cardiomyopathy, or a disease of the heart muscle. Further investigation by the medical examiner showed some characteristics of hypertrophic cardiomyopathy (HCM), which is one of several kinds of cardiomyopathy, but also revealed some characteristics of arrhythmogenic right ventricular cardiomyopathy (ARVC), another disease of the heart. Genetic testing showed two interesting genes: MYH7 (associated with HCM) and RYR2 (associated with ARVC).

The medical examiner officially attributed Floriane's death to HCM but noted that her case had "unusual features"—a reference to the traces of ARVC that had been found. The stated cause of death could just as well have been ARVC, we were told.

We began to research and to think. We discovered that RYR2 is associated with disturbances to the heart rhythm due to adrenaline. Floriane had gone into cardiac arrest when she was wound up with excitement, playing with David. She had received an injection of adrenaline from the paramedics. Could the undeniable structural defects caused by HCM, which usually do prove fatal, but not until age 30 or so, have been amplified by some form of ARVC?

The question gained urgency when we learned that while Floriane's MYH7 was a *de novo* mutation, she hadn't inherited it from us, and her brother didn't have it.

So began our research into our family's genomes.

Getting Under Way

Our goal in *Genomics in the AWS Cloud* is to make the study of your genome, or whatever genome you choose (some are available on the Internet for practice; see Chapter 4), as simple and straightforward as possible. Chapter 2 describes things you should consider before starting. As genome analysis requires computer resources that exceed the average desktop, we set up the analysis environment on the AWS Cloud—that's the subject of Chapters 5 through 7. This book does not presuppose any particular knowledge in molecular biology (which we introduce in Chapter 3) or computers (we talk about Linux in Chapter 8). We paid particular care in explaining how to make sense of our genetic variants in Chapters 11 and 12.

Now, let's start the journey.

How to Contact Wiley and the Authors

Sybex strives to keep you supplied with the latest tools and information you need for your work. If you believe you have found a mistake in this book, please bring it to our attention. At John Wiley & Sons, we understand how important it is to provide our customers with accurate content, but even with our best efforts an error may occur.

In order to submit your possible errata, please email it to our Customer Service Team at `wileysupport@wiley.com` with the subject line "Possible Book Errata Submission."

You can email David Wall with your comments or questions at `david@davidwall.com`.

CHAPTER

1

Why Do Genome Analysis Yourself When Commercial Offerings Exist?

As you begin to explore sequencing a genome and analyzing the results, you will no doubt become aware of a number of commercial operations that offer to do the job for you, neat and tidy, in exchange for a modest amount of money. Why would you want to go through the time and hassle involved in doing the work yourself when such convenient offerings are so easy to use? Why not engage a service to do your genetic analysis and enjoy the benefits of something that "just works"?

The answer has, essentially, two parts.

The first is that the commercial services may not provide all the information you want, not least because they are hamstrung by regulations that govern the provision of medical advice. They tend to provide "novelty" information—Larry Lightbulb kinds of things about hair color and finger length, as well as about racial, national, and tribal heritage. They are great for educating children about the sorts of information that DNA can carry and for talking about heritability of characteristics good, bad, and neutral. When you do the sequencing and analysis yourself, you can extract whatever information you want.

The second is that you are the kind of person who likes to do things yourself, either just for the satisfaction of it or because you want to understand how everything works and fits together. Working on your personal genome in Amazon Web Services (AWS) is an excellent way to learn about those services, and that knowledge can then be put to use for fun and profit.

Commercial Sequencing Services

As you survey the Web, you will find that there are several popular consumer-grade genome sequencing services, including 23andMe, Ancestry, and MyHeritage. Others include the following:

- **African Ancestry**: This service is marketed to Africans and people of African heritage. Some users have regretted that the service does not show DNA broken down by origin in the various regions of the African continent.
- **Athletigen**: This service focuses on markers related to physical fitness, such as those affecting endurance and speed of recovery after exertion.
- **DNAFit**: This service provides diet plans designed around certain nutrition-related markers.
- **Fitness Genes**: This service offers training regimes that fit genomic markers related to physical exertion.
- **GEDmatch**: This service focuses on genealogy and links with others who have submitted their sequences.
- **Genome Link**: This service provides information on a series of characteristics, such as physical endurance and skin color.
- **Genopalate**: Focused on nutrition, this service aims to help its customers optimize their diets.
- **Living DNA**: This is an ancestry-focused service with a user community primarily from the United Kingdom and Ireland.
- **MyHeritageDNA**: Ancestry focused, this service connects its users with possible relatives and suggests possible genetic risks to health.
- **Nebula Genomics**: Offering a monthly subscription that entitles its users to monthly updates as new information becomes available, this service includes data on the oral microbiome (i.e., the bacteria found in your saliva).
- **Promethease**: This is a modestly priced service that detects a number of single nucleotide polymorphisms (SNPs, or "snips").
- **Sano Genetics**: This free service concentrates on SNPs related to autism and mathematical reasoning.
- **SelfDecode**: This is a general-purpose detector of several thousand genetic markers.
- **Vitagene**: Focused on fitness and athletic performance, this service includes ancestry information as well.
- **Xcode Life**: This service offers several low-cost specialty tests including one on skin care and another on metabolic diseases.

Typical Results

Of the aforementioned services, perhaps the best-known of the direct-to-consumer genetic testing services is 23andMe, a California company that pioneered the industry in 2007. When someone places an order with 23andMe, the company sends out a kit containing materials needed for the collection of saliva, which is then sent back for analysis. (The idea is that everyone's saliva contains cells that have been shed from the interior of the mouth.) The company presents its report to the customer via its website.

The company ran afoul of the U.S. Food and Drug Administration (FDA) in 2013, when the regulator objected to 23andMe (and other genetics services providers) advertising that its service provided its customers with information on their susceptibility to various genetically linked conditions, such as male-pattern baldness and certain kinds of cancer. This, the FDA said, constituted medical advice of the sort that should be formulated and delivered by a qualified doctor. The 23andMe tests were medical devices and should be regulated as such.

After going quiet for several years, 23andMe applied to the FDA for permission to include information in its reports about a number of mutations and alleles that are well-understood to be associated with pathogenic conditions, including Alzheimer's disease, Parkinson's disease, celiac disease, and a number of BRCA1 and BRCA2 mutations associated with breast cancer. The company argued—and the FDA ultimately agreed—that the test methods used by 23andMe were sufficiently reliable and understood as to not require the involvement of a medical professional. As well, the entities agreed that the relationships between the tested sequences and the various diseases were adequately proven, and that if an individual was found to have a sequence known to be pathogenic, there was no need to hide the truth behind the medical establishment.

With the shift toward presentation of information about ancestry rather than medical conditions, direct-to-consumer online genetic analysis services have, perhaps predictably, begun to appeal to those whose ancestry is more than passing interest.

So, what's in a set of 23andMe reports? If you undergo the saliva test and log into the 23andMe website today, you will get a lot of novelty information about probable hair color, the shapes of certain body parts, and the aspects of aging.

Sometimes, 23andMe gets things right. For example, 23andMe predicted the following for me:

- A 67 percent probability of little to no back hair. (Correct!)
- A 32 percent probability of a bald spot on the top rear of the head. (Correct, pretty much. It's kind of thinning there, but certainly not bald. Certainly not.)
- A 74 percent chance that the earlobes are separate from the sides of the face. (Hear, hear!)

- A 71 percent chance that the ring fingers are longer than the index fingers. (Yep.)
- A 1 percent chance of red hair. (It's brown.)
- A 62 percent chance that dandruff is sometimes a problem. (It is.)
- A 25 percent chance of being afraid of heights. (I am a qualified pilot.)

Sometimes, it gets things wrong. In my case, it forecast the following:

- A 51 percent chance of blue eye color. (They're green.)
- A 66 percent chance of wavy hair. (It's straight.)
- A 33 percent chance of a widow's peak across the front of the scalp. (There is definitely a prominent one.)
- A 4 percent chance of bunions in the feet. (A substantial one on the right foot has been causing trouble since high school.)

From there, the 23andMe report can get a little comical. For example, my report states that I am likely to wake up at precisely 7:34 a.m. Apparently, the company uses statistical analysis of surveys conducted on people with certain similar genetic markers (those associated with early or late rising) to arrive at the precise time. The suggestion is that genetics determine your wake-up time to the minute, which is just not correct. This is a manufactured novelty "fact," and it's not correct: I almost always gets out of bed before 6 a.m. Hilariously, my wife Catherine's 23andMe report has her waking up significantly earlier than I do, which has happened exactly zero times.

The company provides an ancestry report, which attempts to describe which part of the world your forebears came from.

Some background here: my wife and I live in Australia, to which we both immigrated (Catherine from France and me from the United States) around the turn of the 21st century. We are both people of the New World, largely characterized by its relatively recent immigrants, now.

In my case, the ancestry timeline (shown in Figure 1.1) appears to agree with much of what I know about my family history. It shows ancestors from Scandinavia as recently as 1940. Family lore states that my maternal grandmother was born in Sweden and was brought to Chicago as a baby in the 1920s, so that makes sense. Similarly, my family records have my maternal great-great-grandfather emigrating from Hesse, in the western part of what had recently become a unified German Empire, in 1879. That fits with 23andMe's report of French and German ancestry in the late 19th century.

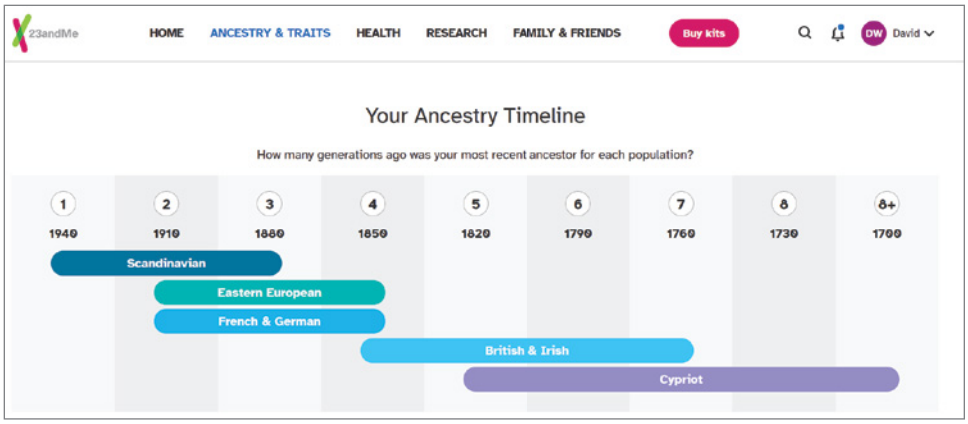

Figure 1.1: My ancestry timeline

My paternal side is less well documented, probably because those ancestors have been in the United States for several generations. The surname Wall fits with heritage from the British Isles, though, and there was always vague talk about my paternal grandmother having ancestors from the eternally contested regions of Central and Eastern Europe.

The report includes Cypriot heritage in the 18th century. I have no idea what that is about, having never heard anything about ancestors on an island of the Ottoman Empire in the Eastern Mediterranean. The 1700s are the distant past for us, as far removed in memory as the original ancestors that first distinguished themselves from the other apes.

Conclusion: no one really knows, but it kind of fits.

As for Catherine, she understands her heritage to be French for many generations back, as far as anyone has been able to trace genealogy. I get scolded when I suggest that the family tree may be particularly branchless. The report from 23andMe (shown in Figure 1.2) concurs, sort of. It shows French heritage back into the 19th century, as well as British and Irish ancestors during the same time period. This latter characteristic may be explained by Catherine's maternal antecedents being from Brittany, a maritime region in the extreme west of France. The language and culture there are Celtic and distinct from those of mainstream France—a distinction that was even more pronounced in the past. The Bretons have a lot in common with other Celtic people, including the Irish, Welsh, Manx, and Cornish. A number of her ancestors were sailors, which probably contributed to the genetic intermixing.

It's all interesting in a bourgeois sort of way, in which people living comfortably enough to spend a hundred bucks on a saliva test can look back at their

ancestors, whose lives were foreign (in time, if not always location) and therefore conventionally romantic and interesting. In the case of people who find that their ancestors migrated from place to place, it becomes possible to conclude that the lives of the ancestors were unstable or otherwise crappy enough to motivate them to relocate, which allows the patrons of 23andMe to burnish their claims to modest origins. It's all good fun and harmless enough, facilitating European stuff on North American Christmas trees and a bottomless market for generic Irish music among Boomers in the former colonies.

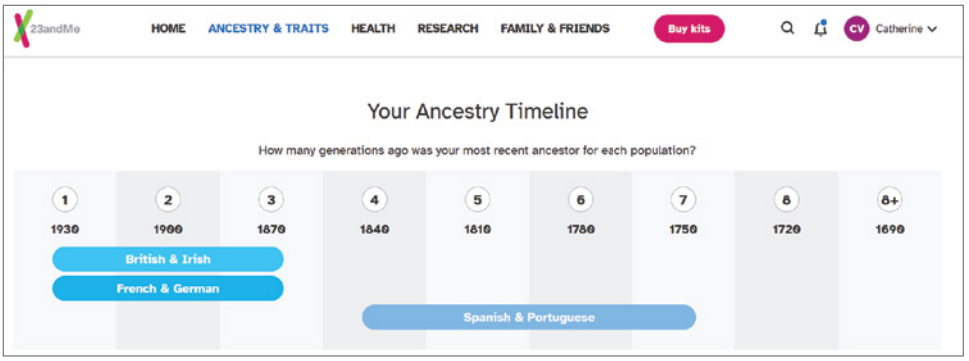

Figure 1.2: My wife's ancestry timeline

There's a dark side to genetic ancestry services like that of 23andMe, though. Some people assign extreme importance to their personal ancestry and believe that certain types of genetic heritage are inherently better than others. A scan of white supremacist websites, for example, will reveal screen shots of ancestry reports from 23andMe and similar services, posted by people attempting to fit in with the group. It's simultaneously risible and sad—or would be so, if it weren't for the undercurrent of violence—to see genetic profiling used in this way. Genetic testing is like any other power tool, though, capable of being used for ill as well as for good.

Perhaps the best part of 23andMe is the fact that it makes available what it calls *raw data*, which is a text file of 7–10 MB containing information about base pair sequences (see Figure 1.3).

You can take the 23andMe raw data file and use it as input to many of the other services listed earlier in this chapter.