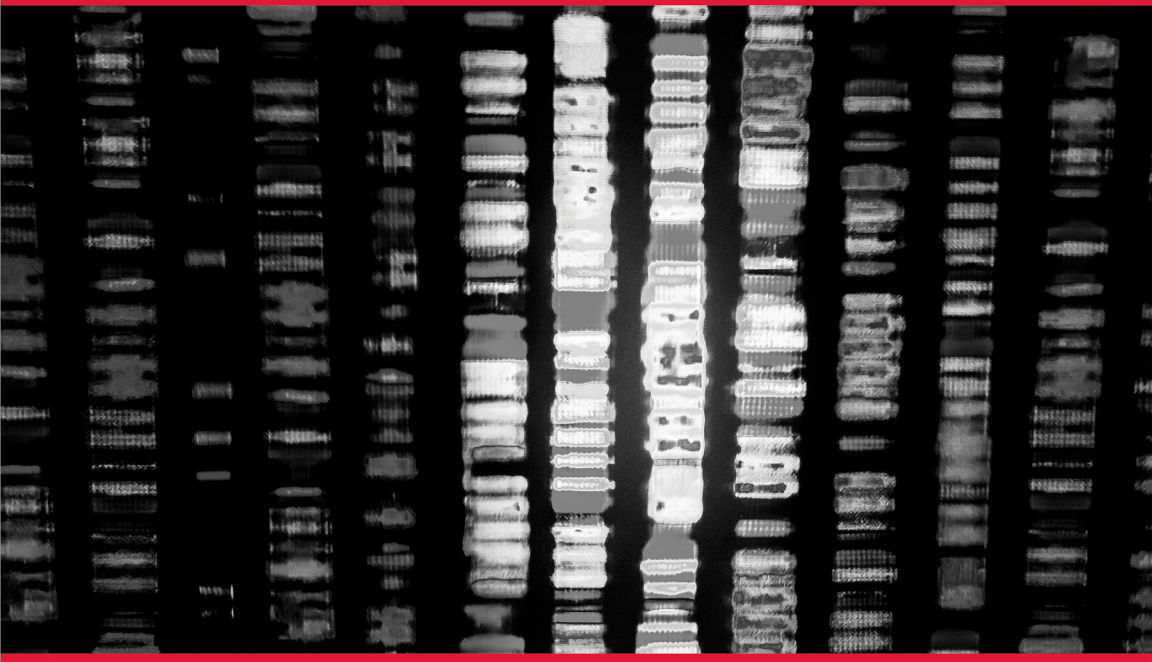
O'REILLY®

The Business of Genomic Data



Brian Orelli

Strata+ Hadoop WORLD

Make Data Work

strataconf.com

Presented by O'Reilly and Cloudera, Strata + Hadoop World is where cutting-edge data science and new business fundamentals intersect—and merge.

- Learn business applications of data technologies
- Develop new skills through trainings and in-depth tutorials
- Connect with an international community of thousands who work with data

Job # 15420

The Business of Genomic Data

Brian Orelli

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

The Business of Genomic Data
by Brian Orelli

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Tim McGovern
Editor: Tim McGovern
Production Editor: Nicholas Adams
Interior Designer: David Futato
Cover Designer: Randy Comer

March 2016: First Edition

Revision History for the First Edition
2016-03-03: First Release

978-1-491-94237-6

[LSI]

Table of Contents

The Business of Genomic Data

Genomic sequencing has come a long way since the international Human Genome Project consortium's first full sequence, which took nearly 20 years and cost about $2.7 billion. Some early pioneers tried to develop new businesses around genomic data—Human Genome Sciences Inc. even named itself after the technology—but it hasn't been until very recently that technological advances have created an opportunity to establish companies with viable business models using genomic data at their forefront.

The price to sequence a genome plummeted to $1,000 last year and might approach $500 this year, which has allowed for a massive increase in the number of genomes sequenced. While the added data makes it easier to identify variations, lower cost of data storage and analysis has been key to identifying which of those variations are important. This report will highlight those big-data issues and how companies are using these swiftly increasing amounts of data to improve diagnostics and treatment.

Broadly speaking, companies can be sorted into two classes: those that create the sequence—either by selling DNA sequencers or by using those sequencers to create the sequence—and companies that use the genomic data to create new products: drugs, biomarkers to facilitate precision medicine, or genomic tests to determine which drugs will work best.

Creating Genomic Data

The first sequencing technology, Sanger sequencing, has given way to next-generation sequencing technology that can produce data

faster and cheaper. Next-generation sequencing comes in two general categories: short-read sequencing, in which DNA is hybridized to a chip, amplified, and then read through synthesis of the complementary strand; and long-read sequencing, in which a single DNA molecule is lead through nanopores and the individual bases are read.

Short-read sequencing, pioneered by Illumina and later produced by Thermo Fisher Scientific's Ion Torrent using a different readout for the synthesis step, has the advantage of low cost and high accuracy. Short reads—50 to 300 base pairs—are generally matched to a known sequence, gaining coverage of most of the genome through overlapping the individual short reads. Unfortunately, the short reads make it difficult to match-up sequences in repetitive areas, often leaving holes in the genome.

Nanopore technology from Pacific Biosciences of California, Oxford Nanopore, and others can produce long sequences averaging 10,000 to 15,000 base pairs, allowing the sequencing through repetitive regions and matching the sequences at the ends of the reads.

"We know 75 percent of the human genome really well. For the remaining 25 percent, it's going to give you fantastically better results," Frank Nothaft, a graduate student at UC Berkley's AMPLab said of long-read sequencing.

The longer reads create more overlap for each fragment, facilitating *de novo* construction of the genome without the use of a template. The lack of a template makes it easier to identify genomic rearrangements that might be missed with short reads.

How important finding rearrangements will end up being remains to be seen, Nothaft noted, "It's a chicken and egg thing. We don't understand structural variation because we don't have enough structural variation data."

The high cost of long-read sequences has limited its use to projects where the organism's genome hasn't been sequenced, where knowing the repetitive sequence is important, or when studying genomic rearrangements. Last year, Pacific Biosciences of California released a new machine, the Sequel System, aimed at lowering the cost of nanopore sequencing. The list price for the Sequel System in US dollars is $350,000, less than half that of its predecessor, PacBio RS II.

Pacific Biosciences of California has a deal with F. Hoffman-La Roche to develop diagnostics tests on the Sequel System. Roche initially plans to develop the machine for clinical research, with a launch planned for the second half of 2016, followed later by a launch of the sequencer for *in vitro* diagnostics to be used in diagnostic labs.

It's possible for long-read sequences to use a reference genome for quicker assembly, but currently most of the long-read sequencing is using *de novo* assembly. "If you're going to pay for the cost, you might as well pay to do the *de novo* assembly," Nothaft said.

But he hypothesized that as the cost of long-read sequencing comes down and the amount of data created with the technique increases, there will be a push to make *de novo* assembly more efficient by decreasing the computing power required. It may also be possible to develop assembly techniques that use better algorithms to blend the best of both *de novo* and reference-assembly techniques.

There are some outlets catering to the retail market—Illumina's TruGenome Predisposition Screen for example—and 23andMe offers a $199 kit that isn't a full genomic sequence, but offers carrier status, ancestry, wellness, and trait reports. But most individual human genome sequencing is being carried out directly for diagnosis of patients.

Rare Genomics Institute started as a way to help patients with rare diseases get connected with research studies to get their genomes sequenced or alternatively to find a way to fund their sequencing on their own, including crowdfunding from friends and family. But as the cost of DNA sequencing has fallen dramatically, the institute has shifted focus.

"The problem is downstream now. Patients don't know what to do once they get their data," said Jimmy Lin, founder and president of Rare Genomics Institute. The institute offers a pro bono consulting team of physicians and researchers in rare diseases to offer support and link patients with specialists that can help with their case.

There are several large genomic sequencing projects being run to create databases that can be analyzed to find connections between genetic differences and phenotypes, the clinical manifestations of the genetic changes.

Human Longevity, the newest project from J. Craig Venter, the man behind the company that competed with the NIH to develop the first draft human genome sequence, plans to sequence up to 40,000 human genomes per year, with plans to rapidly scale to 100,000 human genomes per year.

The company made a deal with South African insurer Discovery Health last year to offer exome sequencing—the exome is the portion of the genome that covers the genes, about 2 percent of a person's genetic data—to Discovery's customers. Discovery Health will cover half of the $250 cost while the patient covers the rest. Human Longevity gives the DNA sequence to the patients' doctors, but will retain a copy and also have access to the patients' medical records to study in large-scale projects.

Human Longevity was spun out of the J. Craig Venter Institute (JCVI), which is a non-profit focused on sequencing a variety of organisms, including viruses and bacteria, to understand human diseases. "Sequencing is the basic assay there," Venter said.

JCVI also spun out another company, Synthetic Genomics, focused on writing genetic code. For example, the company is working on a project to rewrite the pig genome to develop organs for transplants. It also has partnerships with Monsanto to sequence microbes found in the soil and with Novartis to develop next-generation vaccines using JCVI's genomic sequencing and synthetic genomic expertise.

The Million Veteran Program, run by the Department of Veterans Affairs Office of Research & Development, seeks to collect blood samples for DNA sequencing and health information from one million veterans receiving care in the VA Healthcare System. The database of DNA sequences and medical records has 4 petabytes of memory dedicated to storing the information and is already starting to run out of space.

Similarly, Genomics England plans to sequence 100,000 genomes from around 75,000 people and combine it with the health information for patients in the England's National Health Service, the publicly funded nationalized healthcare system. The project, which started in late 2012, is slated for completion in 2017.

Genomics England is split evenly between patients with a rare disease and their families and patients with cancer. The patients with rare diseases will have two blood relatives also sequenced to help

find the underlying genetic changes that cause the disease. The cancer patients will have both normal and tumor tissue sequenced.

Seven Bridges Genomics is working with Genomics England to develop a better way to align short-read sequences. Rather than using a static linear reference to align the sequences, Seven Bridges has designed a Graph Genome based on graph theory that takes into account the observed variations—and their frequencies—at each point in the genome.

"By doing it this way, we allow the alignment to be more accurate," said James Sietstra, president and cofounder of Seven Bridges Genomics.

As new genomes are sequenced, they are added to the Graph Genome, which makes it more useful for aligning future sequences. And by incorporating an individual's variations into the Graph Genome, their data is essentially anonymized but remains part of the population genetics data that can be used to determine the significance of other observed variations.

While the initial DNA sequencing projects were just focused on obtaining the sequence, the latest round is clearly centered on linking genomic changes to clinical outcomes. "We try not to do any sequencing if we don't have phenotype or clinical data," Venter said.

Big Data

The sequencing projects are creating a plethora of data that can be analyzed, but it creates new challenges of how to handle the large amount of data.

The National Cancer Institute (NCI) has funded projects that have generated genomic data on nearly two dozen tumor types from more than 10,000 patients, but the data is stored in different locations and in different formats, making it very difficult to analyze the data in aggregate. To bring data into one place, NCI has partnered with the University of Chicago to develop the Genomic Data Commons (GDC).

In addition to getting the data into one place, GDC analyzed the data and found that there were a lot of batch effects with the way that different researchers handled their respective data. "Just bringing the data into a harmonized, common format so that we could do

a common analysis was a significant amount of effort over almost a year," says Robert Grossman, director of the Center for Data Intensive Science and Chief Research Informatics Officer of the University of Chicago's Biological Sciences Division.

GDC was developed with open source code based on the University of Chicago's Bionimbus Protected Data Cloud that was designed to allow researchers authorized by the National Institutes of Health to access and analyze data in The Cancer Genome Atlas.

But the size of GDC created technical problems for the development that needed to be solved. "A lot of the open source software doesn't scale to the sizes we need, Grossman said. "We're breaking some of these pieces of open source software into what are sometimes called availability zones that we separately manage. And then we bring together separate availability zones to get the scale we need that's required by the project."

GDC is in beta testing as of this writing, with plans of going live in the "June timeframe," Grossman said. The storage includes 2.2 petabytes of legacy data, with plans to add another petabyte or more of additional storage each year to accommodate new projects.

Like Bionimbus, Berkeley's AMPLab is developing tools that help researchers process large-scale data, including a general-purpose API for working with genomic data at scale. "We're getting people to speak the same format for how they're saving data," AMPLab's Nothaft said.

Through the use of on-premises machines, cloud-based computing, and improved algorithmic methods, AMPLab can achieve a four times cost improvement compared to similar tools. Much of the savings comes from avoiding expensive high-performance computing style architecture that isn't as good of a match to the data access pattern that genomic analysis entails.

"While the cost of doing the sequencing has gone down, the cost to do the analysis hasn't gone down much," Nothaft said. "It's not greater than sequencing cost, but it's something people have to think about" as computing becomes a larger percentage of the overall cost of a project, he added.

Human Longevity is also working on developing in-house tools to handle and analyze the large amounts of data that the company is

generating. The company recently hired a new chief data scientist, Franz Och, who was previously head of Google Translate.

"The computer world needs to step up and keep up with the sequencing world," Venter said.

Computer analysis may be the hard part of deriving an answer from big data, but the answer you get may not always be the right one; the most you can truly tell from a database is a correlation between a genomic change and clinical manifestation. The correlation has to be validated by scientists studying the underlying biology.

"Our approach is not to just take a statistical angle at what the data is telling you," said Renée Deehan Kenney, SVP of research and development at Selventa, a big data consulting company. "We're very mindful that correlation doesn't equal causation."

The correlation issue can be further complicated by the quality of the database, which may have data normalization errors. "It's essentially a garbage in, garbage out issue," Nothaft said. "If you don't solve them, any conclusions can be statistically bogus."

Kenney acknowledged that people can publish erroneous data that can't be repeated, but Selventa gets around that by trying to collect enough data that the flawed data is drowned out. "It's getting better, but we have a ways to go in terms of quality," Kenney said.

At some point, we'll reach a critical mass where adding additional data won't be as beneficial, but neither Kenney nor Grossman thinks we're close to reaching that point.

"I don't think we've gotten close to diminishing returns yet," Kenney said, pointing out that rare and pediatric diseases are suffering the most from a lack of data due to the lack of patients and unwillingness to add to the test burden for children.

"Because cancer is often times about combinations of relatively rare mutations, you need enough data so that you have statistical power to understand what's going on," Grossman said. "I don't think we're anywhere near having enough data to do what we need to do."

Data Silos

While there are plenty of projects creating genomic data, they're often isolated in silos that make them unavailable to other researchers.

Part of the isolation stems from a lack of a standard framework for sharing data that UC Berkeley's AMPLab and University of Chicago and NCI's GDC are seeking to break down.

The Global Alliance for Genomics and Health, of which Berkeley, the University of Chicago, the NCI, and 238 other institutions are members, seeks to "create a common framework of harmonized approaches to enable the responsible, voluntary, and secure sharing of genomic and clinical data."

But the key point there is "voluntary." Many investigators will hold back some of the patient-level data even while releasing the key points of the study that are required to get it published. "It's the juiciest and most important information that they want to keep proprietary," Selventa's Kenney said, using the example of scientists publishing expression data, but holding back the information about the patients the tested sample were taken from.

The data commons format that GDC uses is designed to support so-called *strength-of-evidence* databases that can inform treatment and seeks to overcome the issue of separate data silos. "We're trying to open data up through commons, in contrast to a lot of companies that are buying data, siloing it, and sending small amounts back at a proprietary price to those that contributed data," Grossman said. "We're trying to create a critical mass of data that's open so that we can make discoveries in cancer and other difficult diseases."

Developing Products

While big data is helpful for finding correlations and eventually causations, it's the clinical utility of that information that will eventually benefit patients through tests and therapeutics.

Pathway Genomics has used genomic data to develop a series of genetic tests to answer specific questions. Rather than sequencing the entire genome, Pathway Genomics' tests look at specific genes depending on what the doctor is interested in. For example, a series of hereditary cancer tests look for mutations associated with breast

cancer or colon cancer. The company also offers a liquid biopsy blood test that looks for circulating tumor DNA to either try to detect cancer or monitor the disease progression, including examining genetic changes in the tumor that might make certain treatments more effective.

Pathway Genomics has tests for general health and wellness too, including a test that helps patients lose weight by using genetic results to estimate the likelihood of overeating and developing diabetes, and recommended nutritional needs.

The company also has three pharmacogenomics tests that help doctors optimize the use of prescription medications. One test focuses on mental health treatments, another on pain medications, and a third for heart drugs.

While Pathway Genomics is a genetic testing company at heart, the company has spent a lot of effort to simplify the outputted report so it's easy for doctors and patients to understand. The company even has a mobile app that allows patients to share data with multiple doctors without having to keep a copy of the paper report. "We find that to be very powerful for patients," said Ardy Arianpour, chief commercial officer of Pathway Genomics.

Pathway Genomics is even developing a health and wellness mobile app called OME that will dynamically collect data and use machine-based deep learning powered by IBM Watson to offer actionable advice.

While Pathway Genomics spends a lot of effort curating the public databases to determine if genes should be included in its tests, Arianpour noted that "the biggest challenge is developing tests that everyone actually wants or needs."

Pathway Genomics isn't the only company that has developed tests to help doctors make better decisions about treatments. Genomic Health and Myriad Genetics, for example, both have tests to help doctors understand the genetic changes in tumor DNA. Myriad's Prolaris prostate cancer test, for instance, predicts the 10-year survival rate and whether active surveillance versus treatment is a better option for the slow-growing tumor.

In January, Genomic Health announced plans to launch a liquid biopsy cancer test, Oncotype SEQ. "This test is a blood-based mutation panel that uses next-generation sequencing to identify select

actionable genomic alterations for the treatment of patients with late-stage lung, breast, colon, melanoma, ovarian, and gastrointestinal cancers," Phillip Febbo, Genomic Health's chief medical officer told investors on a recent conference call.

While the current cancer tests look at specific genes, Human Longevity announced in January that it plans to offer a comprehensive sequencing of both normal tissue and tumor genome analysis, as well as tumor and germline exome analysis products.

Human Longevity also offers a product called Health Nucleus designed to understand individual health and disease risk. The $25,000 health workup includes whole genome sequencing, sequencing of the patient's microbiome—the bacteria that live inside humans' bodies—and other laboratory tests, including a comprehensive body MRI.

In addition to helping develop tests that can diagnose patients, genomic databases can also help scientists discover new proteins that drugs can target.

Selventa helps drug companies that don't have bioinformatics capabilities discover those new targets. "We reduce the complexity to pathways and elements that a human can wrap their brain around," Selventa's Kenney said.

Human Longevity signed a multi-year agreement with Genentech, a member of the Roche Group, in 2015 to conduct whole-genome sequencing of tens of thousands of patients to identify new therapeutic targets and diagnostic biomarkers.

Even after a drug has been developed, the genomic databases can be helpful in stratifying the patients that would benefit most from the drug based on their genetic makeup—so-called personalized medicines.

While single protein changes have made good diagnostic biomarkers for drugs that target a specific protein, researchers are discovering that it may take measuring the expression of 100 different genes to know the best drug for a cancer therapy. These sorts of correlations can only be discovered by sequencing the DNA of matched pairs of tumors and normal tissue from a large number of patients and require specialized machine learning to indentify the significance of the changes.

Unfortunately, Kenney warned that it's "very hard to develop a diagnostic like that and get it approved by the FDA right now." Regulators will only become more comfortable with the complex algorithmic diagnostics with increasing validation of the complex correlations involved. This will necessarily involve greater sophistication in understanding the results and processes of machine learning, as well as deeper understanding of the biological causal mechanisms.

She also noted that getting insurers to pay for so-called companion diagnostics that tell doctors whether a drug will help a patient remains challenging without a clear intellectual-property element ensuring a period of exclusivity. "Until things change, we're not going to see investments," Kenney said. "That's putting a damper on personalized medicines."