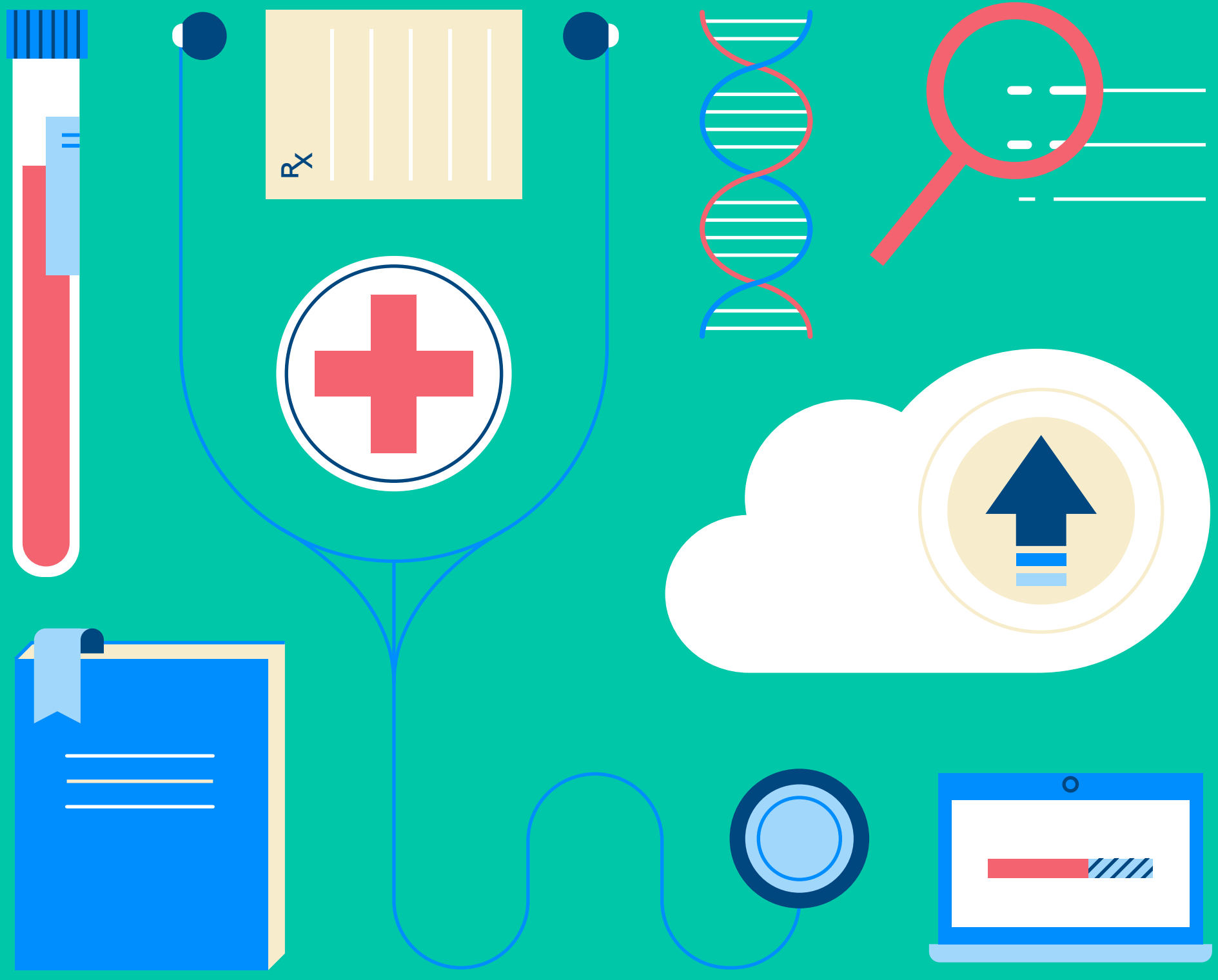
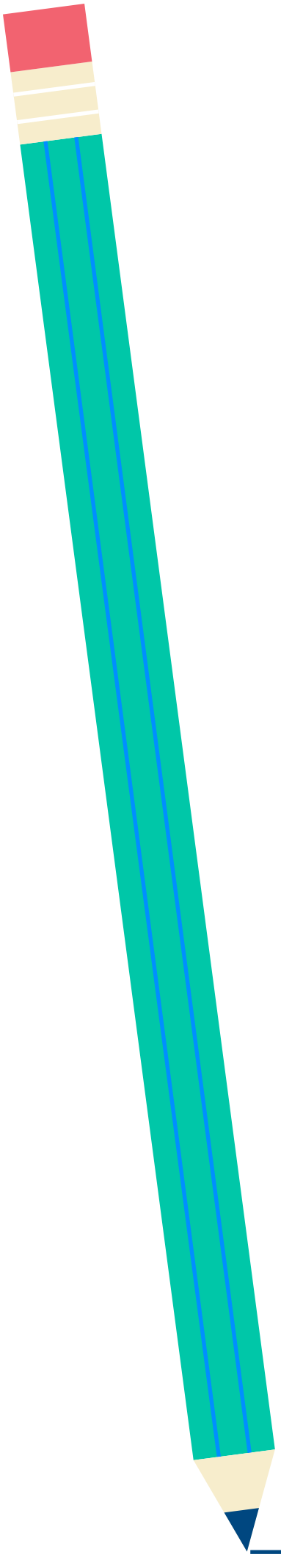
Personalized Medicine and Genomic Research

Profiles in Cloud-Enabled Scientific Discovery



amazon web services | intel

Dear Readers,



Hundreds of thousands of customers around the world have joined Amazon Web Services and are using AWS solutions powered by Intel to build their businesses, scale their operations, and harness their technological innovations. We're excited about our work with the hospitals and research institutions using bioinformatics to achieve major healthcare breakthroughs and unlock the mysteries of the human body.

These organizations are revolutionizing our understanding of disease and developing novel approaches to diagnosis and treatment. A human genome contains a complete copy of the genetic material necessary to build and maintain that organism. The sequencing of this code represents one of history's largest scientific endeavors—and greatest accomplishments. When the Human Genome Project began in 1990, researchers had only a rudimentary understanding of DNA and the details of the human genome sequence. It took around 13 years and cost roughly $3 billion to sequence the first genome. But today, even small research groups can complete genomic sequencing in a matter of hours at a fraction of that cost.

The parallel evolution of genomics and cloud computing over the past decade has launched a revolution in discovery-based research that is transforming modern medicine. Doctors and researchers are now able to more accurately identify rare inherited and chromosomal disorders, and develop highly personalized treatment plans that reflect the unique genetic makeup of individual patients.

Unlocking the Code, Unleashing Innovation

This eBook highlights the important work bioinformatics organizations are undertaking and explains how we are helping them achieve their mission. The stories of these four organizations illustrate what is possible with the AWS Cloud:

GenomeNext is a genomic data management and analysis firm created in partnership with Nationwide Children's Hospital. GenomeNext's AWS based platform represents the newest technological benchmark in the history of genomic analysis, and allows even small research groups to complete genomic sequencing in a matter of hours at a fraction of the traditional cost.

The Human Microbiome Project (HMP) was established in 2008 to assist with the comprehensive characterization of the human microbiome—the sum genetic makeup of both a person and the microbes living in and on that person—and analysis of its role in human health and disease. Researchers from all over the globe can now access HMP data through Nephele, an AWS-supported platform, and use that information to identify

possible microbial causes of preterm births, diabetes, inflammatory bowel syndrome, and other disorders.

The Inova Translational Medicine Institute (ITMI) is assembling one of the world's largest whole-genome sequence databases, through which researchers will be able to track 30 billion genetic variants. AWS architecture facilitates the storage and management of this secure data, and enables Inova researchers to develop personalized treatments and predictive care for newborns suffering from congenital disorders and patients of all ages with cancer-causing genetic mutations.

The Center for Computational Biology & Bioinformatics (CCBB) provides expertise in managing and analyzing large genomic data sets. CCBB has seven core AWS-supported analysis pipelines, all optimized to handle next-generation sequencing data. Each pipeline is targeted at identifying small but important molecular differences, whether in a tumor's DNA or in the microbiome, enabling doctors to tailor treatment on an individual level.

Researchers and administrators in the field realized that scalable cloud-computing models are a better fit than large capital purchases that have to be planned for and budgeted over three- to five-year windows. The sheer size of today's genomic data sets—which are measured in terabytes and occasionally petabytes—makes internal network architecture solutions impractical. The research these organizations are conducting is too important to be held up by processing lags and server errors.

Amazon and Genomics

Research institutions and genomics labs can operate faster and more efficiently by using AWS to access powerful computing tools. These tools are accessed with standard, off-the-shelf computers, and include all of the resources necessary to analyze big data genomic pipelines, store petabytes of data, and share results with collaborators around the world.

Since the launch of AWS, our service and sales teams have created and maintained a robust ecosystem for genomics. These efforts include:

Seeding the market with AWS Research Grants to develop open source tools like Galaxy and AMPLab;

Developing and marketing multiple genomics platform-as-a-service partners, including Illumina, DNAnexus, and Seven Bridges Genomics;

Developing features that address this market's business and compliance needs, such as data encryption, high-performance EC2 instances, and robust networking.

Most recently, AWS launched two controlled-access cancer genomics public data sets, The Cancer Genome Atlas on AWS and International Cancer Genome Consortium on AWS, that qualified researchers can access at no cost as part of the AWS Public Data Sets program. Providing access to these petabyte-scale genomic data sets as shared resources on AWS lowers the barrier to entry, expanding the research community and accelerating the pace of research and discovery in the development of new treatments for cancer patients.

Looking Ahead

Medical and scientific communities around the world are just starting to take advantage of the transformative opportunities that personalized genomic medicine offers patients. The four organizations highlighted in this eBook are at the forefront of that medical revolution.

1 GenomeNext

Bringing Chromosome Analysis to the Cloud

By the Numbers

4 million
babies born every year in the U.S.

1 in 10,000
babies born every year who will be afflicted with SMA

2 years
average lifespan of a child diagnosed with SMA Type 1

A Race Against Time

The baby seemed fine at eight weeks.

She made adorable cooing noises in between lopsided smiles and her eyes were little saucers of wonder.

But by six months her parents were concerned by her lack of movement. By eight months, the doctors were alarmed as well. She responded to her mom and dad with her eyes, but the rest of her was motionless in the crib. By ten months she was back in the hospital, unable to move her arms or sit up.

Week by week, her condition became worse. The medical team treating her tried their best to make her comfortable, but couldn't come to a conclusive diagnosis. By the time the specialists realized she was afflicted with spinal muscular atrophy, or SMA, there was very little that could be done to halt the disease and save her life.

By eighteen months she was gone.

This hypothetical scenario is all too real for many parents. SMA is the most common genetic cause of infant mortality, and children suffering from the most severe type of SMA are not likely to live past their second birthday.

SMA is caused by a gene mutation which leads to the loss of motor neurons in the spinal cord, and eventually, a loss of movement in the head, arms, and legs. Of the roughly four million babies that will be born in the U.S. this year, about 400 will be afflicted with SMA.

The disease is passed genetically from both parents, who are usually unaware that they are carrying the abnormality.

There are three types of SMA in children, with Type 1 being the most severe—and the easiest for trained specialists to recognize, as symptoms are evident by the time a child is six months old. In Type 2 and Type 3 cases, symptoms may not show up until the child is ten years old or even a teenager. In all cases, however, symptoms can include weakness of the voluntary muscles, diminished limb movements, difficulty swallowing and feeding, and impaired breathing.

"Spinal muscular atrophy is the most common genetic cause of infant mortality."

—Dr. Darryl De Vivo, Sidney Carter Professor of Neurology, Pediatric Neurology Service at Columbia University Medical Center

Simple genetic testing can identify most standard cases of SMA. But some cases—like the hypothetical one explored above—are unusual, and difficult to detect through normal testing procedures, which focus on one isolated gene at a time.

Genome

An organism's complete set of DNA instructions, including all of its genes.

If there was a reliable and fast way to perform *whole genome* testing, however, doctors could diagnose all manifestations of the disorder at the earliest possible point, treatment could begin sooner, and the prognosis for improvement would increase across cases. Genomic analysis could also be the key to unlocking a cure for this devastating disease.

From Mutant Genes to Complete Genomes

The advent of genetic testing and concurrent advances in human genetics transformed our understanding of many inherited conditions, and led to improved diagnostic techniques, classification systems, and treatment methods.

For years, genetic testing was the least invasive and most accurate way to identify chromosome mutations. The procedure involves a close examination of genetic material to detect changes (mutations) from the usual sequences of chemicals, and can be performed with only a blood sample from the patient.[1]

Unfortunately, a negative test result decreases the likelihood but does not exclude the diagnosis of a genetic disorder such as spinal muscular atrophy. While most SMA cases are the result of a mutation at a specific location on one particular gene, a small but growing percentage of cases involve mutations in different genes.

Additional genetic testing is time and labor intensive—and expensive, with bills quickly spiraling above ten thousand dollars. If a diagnosis hasn't been made after several rounds of testing, many doctors will recommend forgoing further genetic testing entirely.

As costs for genetic sequencing continue to plummet, scientists, doctors, and researchers across disciplines are embracing the new technology.

Whole genome testing removes this ambiguity. A genome is an organism's complete set of genetic material, which includes around 20,000 genes in humans.[2] While traditional genetic testing focuses on a single gene or a panel of genes, whole genome testing examines the entire genome one nucleotide at a time.[3] The benefits are extraordinary:

Patients with unexplained genetic disorders have a good chance to find the cause of their condition.

Physicians can see how previously unknown genes may be contributing to the disease state. Traditional genetic testing examines only the known "troublemaker" genes.

Physicians can better understand how specific treatments for a disease will be affected by the patient's unique genetics.

Because of its scale and expense, whole genome testing used to be a rarely employed diagnostic technique. But as costs for genetic sequencing have plummeted, scientists, doctors, and researchers across disciplines are embracing the technology.

GenomeNext is leading this new wave of genome-driven innovation, enabled by Amazon Web Services.

GenomeNext and AWS

GenomeNext is a genomic data management and analysis firm established in 2014. The company was created in partnership with Nationwide Children's Hospital, one of the largest and most comprehensive pediatric hospitals and research institutes in the United States. GenomeNext provides the Columbus, Ohio-based institution with genome sequencing and molecular diagnostics—among other services—on an AWS cloud-based platform that facilitates faster analysis.

As the name suggests, GenomeNext specializes in advancing our understanding of the human genome and harnessing it as a research tool. While genome sequencing—the act of mapping out and analyzing the entire genetic makeup of a person—has been at the forefront of medical

Under the Hood: Intel Solutions

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases.

For GenomeNext: C3 instances are based on high frequency Intel Xeon E5-2680 v2 ("Ivy Bridge") processors, and are designed for running compute-intensive applications.

research for some time, the sheer size of the data has presented a number of technical obstacles for researchers and practitioners.

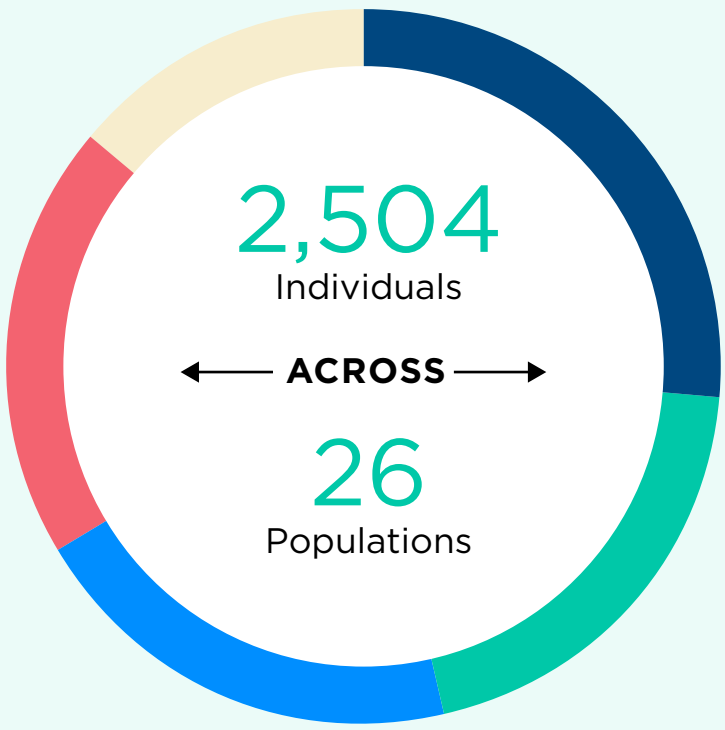
To overcome this Big Data challenge, the GenomeNext team, led by Dr. Peter White, developed a computational pipeline called Churchill that allows efficient analysis of a whole genome sample in as little as 90 minutes with results that are completely reproducible.

AWS provides critical support for this process in two key ways:

GenomeNext's entire software-as-a-service platform is hosted on the AWS cloud. Moving data to the cloud freed up a tremendous amount of local server capacity and allowed for easy remote access. Clinical groups and researchers around the world are now able to upload and compute large amounts of data on the platform through a completely automated system.

The speed of GenomeNext's platform is enabled by AWS parallel processing architecture, which is able to handle simultaneous analyses that each require significant amounts of computational resources. Because of this, GenomeNext has been able to reduce the timeframe for genome analysis from two weeks to two hours or less—a monumental achievement.

GenomeNext's Churchill platform was able to upload and analyze the entire 1000 Genomes Phase III dataset in only seven days. The 1000 Genomes dataset represents:



- **661 African**
- **489 South Asian**
- **504 East Asian**
- **347 Admixed American**
- **503 European**

2,504 Exomes

The exome makes up a very small portion (1.5%) of the genome, but contains all of the protein coding genes. Most genetic disorders are correlated with mutations in protein coding genes.

85.4 Trillion Base Pairs

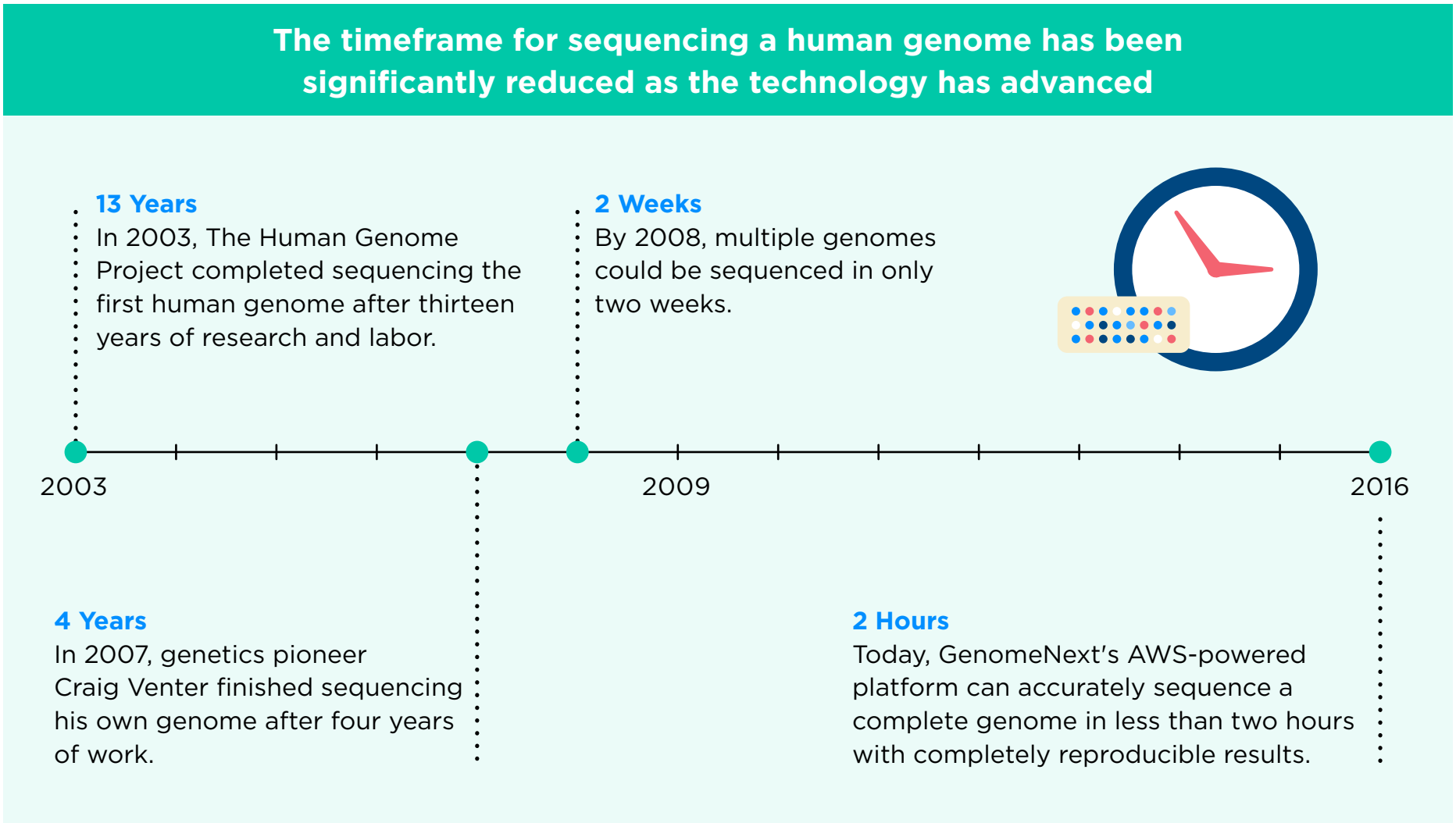
Base pairs are the hydrogen bonded nucleotides which form building blocks of the DNA double helix.

2,504 Whole Genomes

A genome contains the complete DNA of an organism. In the case of a human this corresponds to about three billion base pairs of DNA.

70 Terabytes of Data

After a single genome is sequenced—let alone a thousand—scientists are left with billions of data points to analyze.



With Amazon's help, GenomeNext has already been to able to assist researchers and health care practitioners in identifying rare pathogenic variants. But the service's most valuable innovation is just around the corner: population-scale genomic analysis, which will provide unprecedented insight into the genetic origins of cancer, obesity, Alzheimer's, and heart disease.

GenomeNext's ability to handle analysis at that scale within the cloud was demonstrated in early 2015, when the Churchill platform successfully processed a complete population data set from the 1000 Genome Project in only one week. The data set consisted of raw genomic sequence data from 2,504 individuals sampled across 26 different populations. This accomplishment represented the fastest and most accurate analysis of a data set of such magnitude to date.

As large population-scale genomic studies become routine, GenomeNext's cloud-based platform provides a reliable way for researchers to manage the data burden and share the results. And most importantly, an ever-growing repository of population-scale genome data will advance and improve the delivery of care for individual patients worldwide.

The Churchill system is incredibly fast, impressively accurate, and widely available to any researcher with a fast internet connection.

The Genome, The Cloud, And The Future

Today, thanks to GenomeNext and AWS, even small research groups can complete genomic sequencing in a matter of hours at a fraction of the previous cost.

GenomeNext's cloud-based platform represents the newest technological benchmark in the history of genomic analysis. The Churchill system is incredibly fast, impressively accurate, and widely available to any researcher with a fast internet connection. Adoption and utilization of the technology will lead to previously unimaginable scientific achievements and medical discoveries.

Genomic analysis on the cloud will be especially valuable in the the fight against degenerative disorders—such as the spinal muscular atrophy cases addressed at the beginning of this chapter—as diagnosis can occur remotely, giving doctors a better chance to identify disorders before symptoms are even present. Gene-therapy research may even produce a cure when and if doctors are able to manipulate viruses to replace damaged genes with healthy ones.

Researchers at Children's Hospital are currently conducting trial treatments using gene-therapy techniques. If the project continues to generate positive results, gene-therapy treatments could also be explored for other diseases, including Lou Gehrig's disease and muscular dystrophy.

Greater understanding of the human genome will only lead to more such breakthroughs.

Through the scale and flexibility of the cloud, GenomeNext is able to offer fast and accurate genomic analysis on a secure and regulatory compliant platform. The service will ultimately advance and improve the delivery of care for patients worldwide and remove the barriers of population-scale genomics.

References

Arribas-Ayllon, Michael. Sarangi, Srikant. Clarke, Angus. (2014). Genetic Testing: Accounts of Autonomy, Responsibility and Blame. Routledge.

Bainbridge, William Sims. (2010). Converging Technologies for Improving Human Performance: Nanotechnology, Biotechnology, Information Technology and Cognitive Science. Springer.

Ching H. Wang, et al. (2007). Consensus statement for standard of care in spinal muscular atrophy. Journal of Child Neurology, 22(8): 1027-1049.

Crane, Misti. (June 2015). Pioneering gene therapy may save babies' lives. The Columbus Dispatch.

Davies, Kevin. (2002). Cracking the Genome: Inside the Race to Unlock Human DNA. Johns Hopkins University Press.

De Vivo, Darryl. (2011). What is SMA?. DNA Learning Center.

Ghose, Carrie. (Jan 2015). Nationwide Children's Hospital spins out fast gene analysis software. Columbus Business First.

GHR. (2015). What is a genome?. Genetics Home Reference.

Hughes, Virginia. (July 2013). Clinics Offer Expensive Whole-Genome Tests for Undiagnosed Disorders. Scientific American.

Kelly BJ, Fitch JR, Hu Y, Corsmeier DJ, Zhong H, Wetzel AN, Nordquist RD, Newsom DL, White P. (2015). Churchill: an ultra-fast, deterministic, highly scalable and highly balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. Genome Biology, 16(6).

Lee, Kristen. (2015). Vast data doesn't have to restrict population-scale genomics. TechTarget.

MDA. (2015). Overview: What is spinal muscular atrophy?. Muscular Dystrophy Association.

MDA. (2015). Spinal Muscular Atrophy: Diagnosis. The Muscular Dystrophy Association.

Mally, Julie. (April 2015). 5 Questions for Dr. Peter White, GenomeNext. Intel.

My46. (2015). Whole genome and exome sequencing. My46 / University of Washington.

NIH. (May 2015). Spinal Muscular Atrophy Fact Sheet. National Institute of Neurological Disorders and Stroke.

Ng, P. C., & Kirkness, E. F. (2010). Whole genome sequencing. In Genetic Variation (pp. 215-226). Humana Press.

Ogino, Shuji, et al. (2002). Spinal muscular atrophy genetic testing experience at an academic medical center. The Journal of Molecular Diagnostics, 4(1): 53-58.

Peeters, K., Chamova, T., & Jordanova, A. (2014). Clinical and genetic diversity of SMN1-negative proximal spinal muscular atrophies. Brain, 137(11), 2879-2896.

Pillsbury, Edmund. (ND). A History of Genome Sequencing. Yale University.

Rivard, Laura. (2015). Whole Genome Sequencing. Genetics Generation.

Rosenberg, Roger N. Pascual, Juan M. (2014). Rosenberg's Molecular and Genetic Basis of Neurological and Psychiatric Disease: Fifth Edition. Academic Press.

Topol, Eric. (2011). The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care Kindle Edition. Basic Books.

TREAT-NMD. (NA). Diagnostic Testing and Care of New SMA Patients. TREAT-NMD Neuromuscular Network.

Weisleder, Pedro. (2012). Manual of Pediatric Neurology. World Scientific Publishing Company.

2 The National Institutes of Health Human Microbiome Project

Using the Cloud to Combat Preterm Birth

By the Numbers

4 million
babies born every year in the U.S.

450,000
premature births every year in the U.S.

35%
Percent of infant mortalities that are attributed to premature birth

$26 billion
annual cost of premature birth to the U.S. healthcare system[6]

Preterm Birth in the US

"You're in labor."

These are the words any expectant parent eagerly waits to hear from her doctor. Many have tried for years to get pregnant and have taken all the precautions to encourage a healthy pregnancy—no smoking, eating right, plenty of rest. The only problem—the baby is just at 33 weeks. Although the baby was successfully delivered, he might have to spend weeks or even months in intensive care. Many babies make it through and come home with their parents to lead normal lives. Many do not.

Over 4 million babies are born every year in the United States. Of those, roughly 450,000—or one in nine—are born prematurely. Any baby born before 37 weeks is considered premature, or preterm—a condition that can cause a host of symptoms, and even death. In fact, preterm birth-related causes of death account for approximately 35% of infant mortality, more than any other single cause. Though not always fatal, these effects can be devastating for families, and include low birth weight, breathing difficulties, underdeveloped organs, vision problems, and cerebral palsy.[4] Preventing this condition remains "a challenge, because the causes…are numerous, complex, and poorly understood."[5]

Microbiome

The sum genetic makeup of a person AND the microbes living in and on that person. Researchers now estimate that the human microbiome contributes 360 times more bacterial genes than human genes.[8]

The Multi-Omic Microbiome Study: Pregnancy Initiative

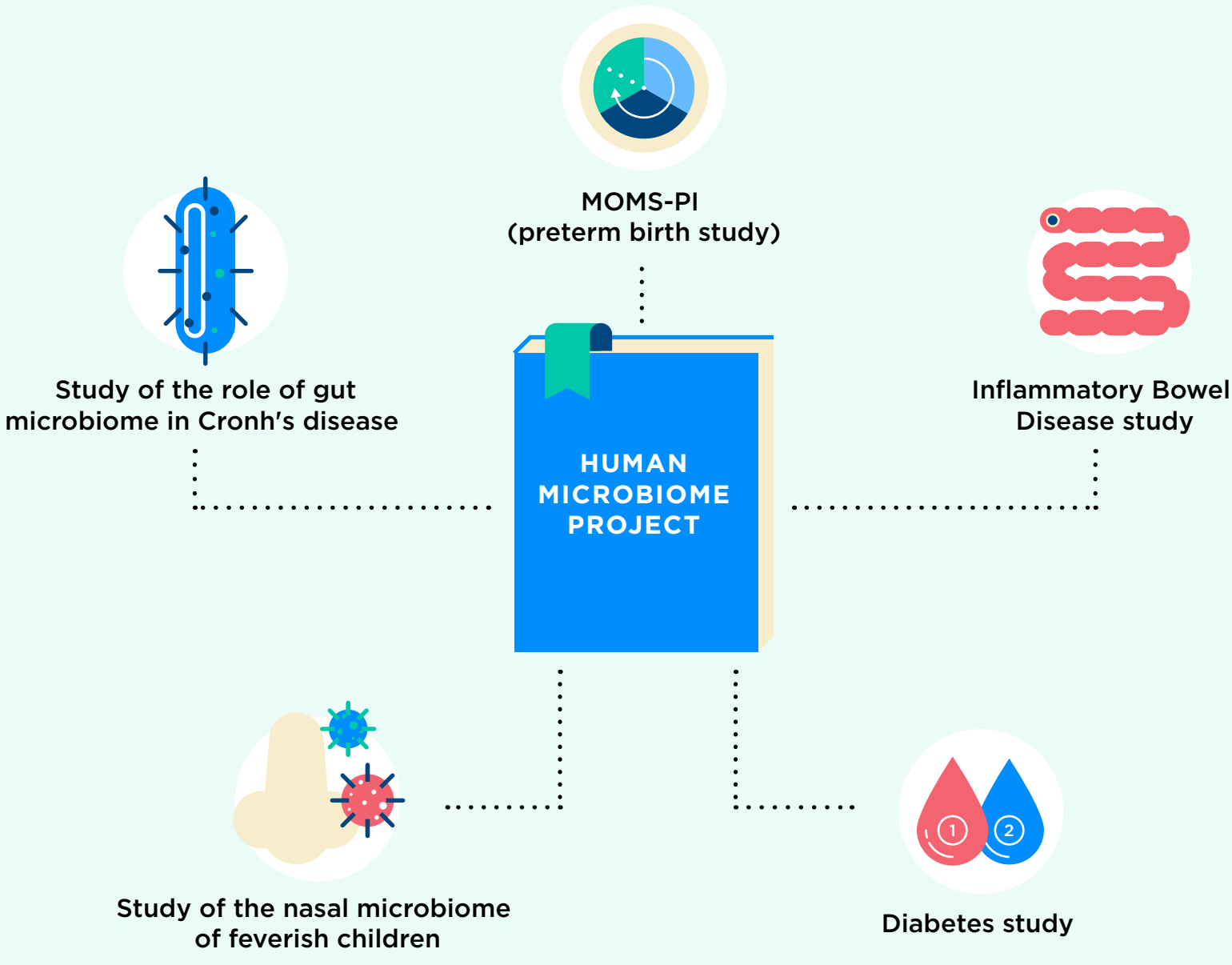
Doctors are looking for answers in an unexpected place: the microbiome, or the complex ecosystem of microbes that live within the human body. Humans have trillions of microbes living inside them. Many are helpful, like those that break down food. Sometimes, however, microbes can be harmful and lead to health complications like diabetes and preterm birth. Babies get their first microbes from their mothers, so understanding the relationship between mother, baby, and microbe may give insight into the causes of early birth.[7]

One such ongoing study is taking place at Virginia Commonwealth University. A collaborative project, the Multi-Omic Microbiome Study: Pregnancy Initiative (MOMS-PI), aims to study the impact of the vaginal microbiome on pregnancy and the fetal microbiome. The study will do this, in part, by amassing a large data set of genetic data about mothers and their particular microbiomes. Taken together, this is known as metagenomic data, and the study intends to collect information from 2,000 women.

The Human Microbiome Project

In order for a study like MOMS-PI to be viable, researchers need a baseline of healthy, normal microbiome data against which to compare results. This exists in the form of Phase I of the Human Microbiome Project (HMP), a large-scale analysis of the human microbiome that was completed four years ago by the National Institutes of Health (NIH). As NIH director Francis Collins says, HMP was established to generate a "reference database by using genome sequencing techniques to detect microbes in healthy volunteers...[laying] the foundation for accelerating infectious disease research previously impossible without this community resource."[9] The HMP Phase I can be thought of as the hub on a wheel, and MOMS-PI is one of the many spokes, or secondary studies that build on that database, as part of Phase II. Using this library of healthy metagenomic data, researchers can pose a question and have a standard to compare their results to.

Numerous secondary studies have been spawned from the data compiled in the original Human Microbiome Project



Nephele uses built-in logic to automate much of the remote data analysis process

STEP 1

Clients gather genomic data from subject microbiome ("metagenomic data")

STEP 2

Clients answer a few questions about what their goals are, click "submit"

STEP 3

Data is sent through one of 15 standardized pipelines

STEP 4

Built in logic determines case-specific need to install libraries, applications, etc.

STEP 5

Nephele initiates only as much server space as necessary through AWS cloud architecture.

STEP 6

Clients receive automatically generated results which are customized to their needs: charts, annotation files, etc.

Under the Hood: Intel Solutions

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases.

For NIH: C3 instances are based on 2.4 GHz Intel Xeon® E5-2676 v3 (Haswell) processors and provide a balance of compute, memory, and network resources.

Nephele

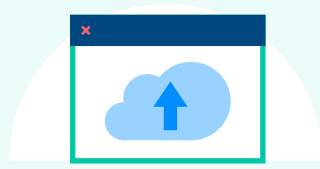
Preterm birth is just one of three core studies funded by NIH as part of Phase II, including research into the possible microbial causes of diabetes and Inflammatory Bowel Disease. Fortunately, the Office of Cyber Infrastructure and Computational Biology within the National Institute of Allergy and Infectious Diseases, a division of NIH, has devised a platform, running on Amazon Web Services, that enables researchers from all over the globe to tap into data within HMP.

Named after Nephele, the Greek cloud goddess, the platform leverages AWS S3 cloud architecture to facilitate remote upload of data that researchers have gathered and have it rapidly, reliably analyzed. This is facilitated by Lamba Functions, which also automatically manage initialization of Amazon Machine Instances, based on the internal logic that Nephele's creators baked into the system. Nephele provides researchers with tools, data, and a highly functional infrastructure, making what was previously a very technical, difficult process much more affordable, accessible, and automated. The Nephele team calls this a democratization of scientific research—researchers who wouldn't have the resources to do this sort of work at their home institutions are now able to access the tools to do so.

Rather than remotely accessing a server and having to wait in queue while one application finishes, receiving that data, and then doing it again, Nephele boasts an automated interface predicated on AWS's ability to scale computing power based on demand. The platform initiates server space as needed, capitalizing on EC2's parallel processing to drastically shorten the "time to science" for researchers. Nephele is the tool that unlocks the potential of the HMP—providing research teams posing novel

questions a platform to answer those questions. By looking at the human microbiome at scale, medical science will gain insight into myriad health issues—from dry skin to Irritable Bowel Syndrome to liver cancer to preterm birth. The AWS-supported pipeline application allows researchers to focus resources on scientific investigation, working toward diagnosis and treatment. Being able to analyze extremely large data sets rapidly, accurately, and remotely means the MOMS-PI team may be able to isolate a genetic or metagenetic cause of preterm birth. If they can isolate these causes, we are one step closer to finding treatment—and one step closer to bringing those babies home healthy.

Nephele Cloud Architecture



WEB FRONT END (UI)
The user uploads data and selects the data outputs they need.

INPUT/OUTPUT HANDLING
Internal logic interprets user requests and translates them for AWS.
This is the “Template” layer, enabled by S3.

LAMBDA FUNCTION
Code initializes Amazon Machine Instances via EC2,
including type selection based on Template logic.

S3
Stores user pipeline parameters, submitted data, and logs in cloud.

SQS
Manages queueing of messages

EC2
Processes data within analysis environment.
Elastic Cloud Compute can be scaled up and down rapidly as needed.

3 Inova Translational Medicine Institute

Personalized Treatment and Preventative Care Through the Cloud

By the Numbers

Genetic alterations in the two main breast cancer susceptibility genes (BRCA1 and BRCA2) account for:

5% to 10%
of female breast cancers

5% to 20%
of male breast cancers

15% to 20%
of familial breast cancers.

SOURCE: ACR. (2015). Cancer Facts and Figures: 2015. American Cancer Society.

A Cancer Clue in the Genetic Code

After a long career enduring Washington, D.C.'s harsh winters, Jack V. retired to sunny Pinehurst, North Carolina to spend his golden years golfing, relaxing, and finally enjoying time with his family. But Jack's retirement dream was interrupted by a sudden nightmare: a strange lump on his cheek, which he first noticed during his morning shave.

The lump was diagnosed as malignant cancer in the salivary gland—a very rare cancer that makes up less than one percent of all cancers in the United States.

Jack's treatment involved standard medical procedures: a nine hour surgery to remove the tumor followed by chemotherapy and several courses of radiation. But when the cancer recurred, his doctors decided against automatically repeating chemo and radiation until they could learn more about Jack's DNA and its relationship to his tumor.

The procedure they employed instead is called "tumor profiling," and it involves a genomic assessment of a tumor to learn more about its root causes. In cancer cells, small mutations in the genetic code can cause the cell to make a protein that prevents the cell from properly functioning. Such abnormal proteins can cause cells to multiply uncontrollably and damage neighboring cells.

By studying the cancer genome, scientists can discover what specific mutations are causing a cell to become a cancer, and to distinguish one type of cancer from another. Understanding the cancer genome can also help doctors develop a personalized treatment plan for each patient.

Jack's tumor was tested for up to forty-seven different genetic mutations that could be driving the cancer growth. Surprisingly, his genome assessment came back positive for mutations normally associated with breast cancer. Even though he didn't have that specific disease, his doctors knew what genetic mutation Jack carried.

Armed with this critical piece of information, his doctors prescribed the breast cancer drug Herceptin.

This treatment was successful for two and a half years until the cancer mutated. But by again using genetic analysis, Jack's doctors were able to

Understanding the cancer genome can help doctors develop a personalized treatment plan for each patient.

detect a mutation—this time, it was related to melanoma genes. Again, they were able to use pharmaceuticals to turn off the cancer mutations' "on switch"—and allow Jack to continue to enjoy his well-deserved retirement.

From Cancer to Congenital Disorders

The underlying breast cancer-related genes in Jack's cancer were identified through rapid genomic analysis, a revolutionary diagnostic tool that is finally becoming widely accessible to doctors and researchers. Genomic analysis facilitates highly individualized care for patients. Jack's entire course of treatment, from the prescription of Herceptin to the reassessment after the cancer's mutation, was guided by the results of his cancer's unique genome sequencing.[10]

The type of personalized treatment plan that saved Jack's life is being developed and implemented through the Inova Translational Medicine Institute (ITMI), a non-profit research institute that applies genomic and clinical information from individuals to develop innovative methods for personalized healthcare.[11]

ITMI is collecting and analyzing data on a scale unmatched by other genomic research institutes and labs, and using that data to make important discoveries for the future of healthcare.

ITMI's research on premature births, childhood genomic health, and congenital disorders is especially significant, as they demonstrate how genomic analysis allows for a shift from a reactive to a predictive care model that improves patient outcomes:

Premature Births

Premature birth is a complicated phenomenon linked to a range of variables—including genomic, clinical, environmental, and behavioral factors—that have not received enough analysis to reveal satisfying answers for patients and doctors. ITMI collaborated with Fairfax Hospital and GNS Healthcare in 2011 to identify factors that may be linked to genetic or other biologic changes. The Premature Birth Study involved complete DNA sequencing of ~1,000 families—reflecting an approximately equal mix of preterm and full term deliveries—which produced a massive database of information. The study is now in the analysis phase to examine the role of specific genetic differences in mothers, fathers and babies

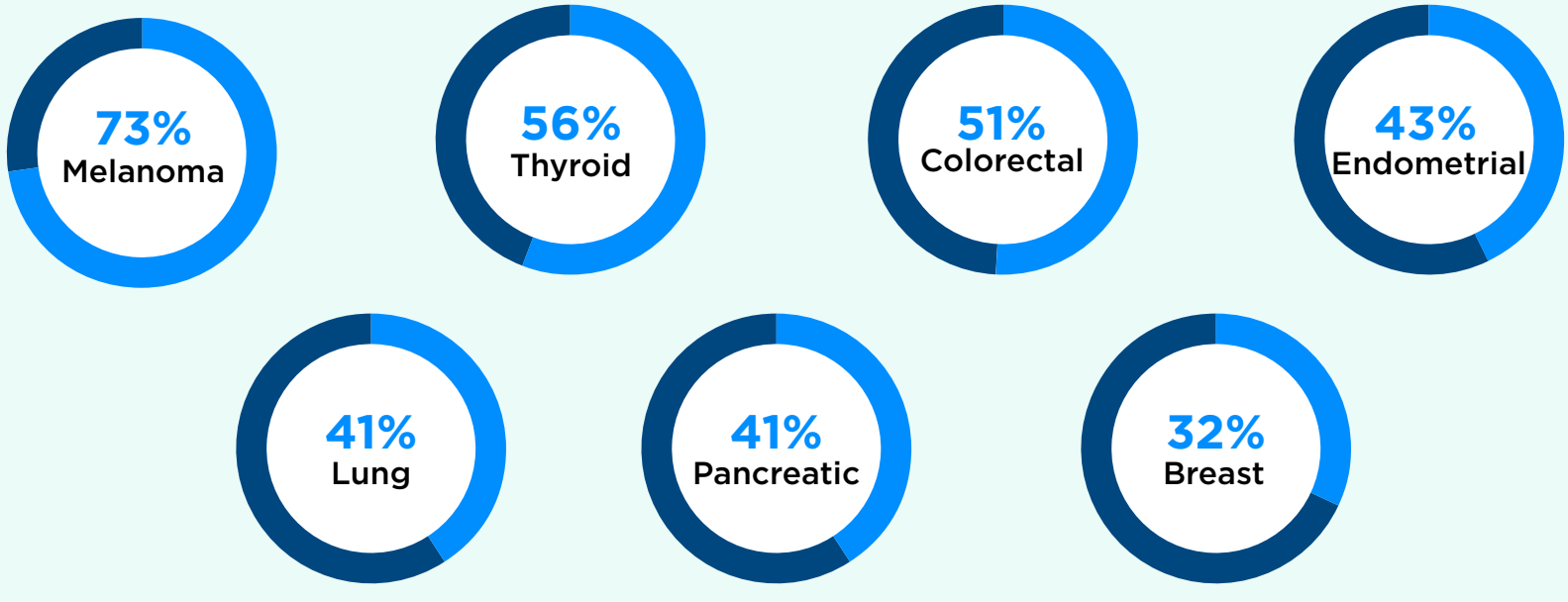
when the baby is born before and after 36 weeks. The results will help researchers pinpoint common genetic traits associated with early births, which will ultimately help doctors predict and prevent them.[12]

Longitudinal Study
In 2012, ITMI launched the Longitudinal Childhood Genome Study (also known as the First 1,000 Days of Life and Beyond study) as an ongoing, 18-year study of up to 10,000 families—including grandparents, parents, children and other relatives. The goal of this massive, multigenerational study is to generate more than 30,000 whole genomes, with the earliest samples collected during pregnancy and continuing through gestation to 18 years of age.[13] This data will enable ITMI researchers to look at several aspects of neonatal well-being and childhood health, such as augmenting the newborn screening process.

Congenital Disorders
ITMI performs whole genome sequencing on newborns in the neonatal intensive care unit when conventional genetic testing fails to provide a diagnosis for their conditions. Whole genome sequencing currently yields diagnosis in more than half of fully analyzed cases for study participants, and allows medical practitioners to assess these methods in real-time clinical practice. Around 150 families have been enrolled in the congenital disorder study since it began in 2012.

All of ITMI's studies are designed to build genetic models that help answer questions about individual predispositions to diseases, treatments, and preventive measures. But collecting, storing, and processing this vast genetic data presents unique technological challenges for researchers and practitioners working within limited operational budgets.

The percentage of tumors driven by genetic mutations that could be targets for specific drugs, by type of cancer



Why ITMI uses AWS

- To migrate biological data from vendors
- To easily deploy web applications
- To quickly perform proof of concepts
- To lower data storage costs
- To share data with collaborators
- To utilize a flexible number of Amazon EC2 instances

Under the Hood: Intel Solutions

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases.

For Inova: C3 instances are based on high frequency Intel Xeon E5-2680 v2 ("Ivy Bridge") processors, and are designed for running compute-intensive applications.

Partnership with AWS

ITMI is assembling what is expected to be one of the world's largest whole genome sequences databases, and already has 8,300 genomes stored. With this massive data set researchers will be able to track more than 30 billion genetic variants. But the storage and management of this data presents major technological challenges.[14]

Constructing an on-premise storage infrastructure for petabyte-scale data sets would quickly exhaust any organization's budget. Furthermore, physically moving data around through uploads, downloads, and transfers from server to server incurs a lot of latency, which delays research significantly. Extremely large file sizes also make data durability a concern, as even minor levels of data decay could jeopardize analysts' ability to reproduce results.

The architecture provided by Amazon Web Services (AWS) addresses and resolves all of these logistical hurdles:

The cloud provides Inova with elastic capability—their capital outlay only needs to match the task at hand. In a traditional model, the equipment and infrastructure necessary to run the largest possible study is always the limiting factor. With AWS, Inova can initiate server space and compute instances only as needed, allowing them to work much faster, at a smaller cost, and with greater quality control.

The cloud architecture enables Inova to consolidate all of its genomes into one place. This, in turn, facilitates effective collaboration wherein doctors and researchers from all over the globe can access and process genetic data.

Inova's analysis pipeline, also built on AWS, is optimized to process full genomes and deliver panel results in a matter of days. This computational power is coupled with access to a vast database of baseline information, which allows doctors to meaningfully place into context the data they receive.

AWS also allows Inova to secure data, as the platform meets all Health Insurance Portability and Accountability Act (HIPAA) compliance requirements. This is an ethical and business necessity when dealing with something as intimately personal as genomic data.

Inova estimates that storing data using Amazon S3 (Simple Storage Service) and Amazon EC2 (Elastic Compute Cloud) has saved the organization more than $10 million dollars in upfront costs. And just as importantly, the power and scalability of the AWS platform enables Inova researchers to start unlocking the potential of genomic analysis in developing personalized treatments and predictive care. The results are already helping scientists understand the causes of premature birth, newborns overcome congenital disorders, and Jack get his cancer in check so he can get back onto the golf course.

Sources

ACR. (NA). Genes and Cancer. American Cancer Society.

ACR. (2015). Cancer Facts and Figures: 2015. American Cancer Society.

ASCO. (Oct 2015). The Genetics of Cancer. American Society of Clinical Oncology.

Avere. (2014). Secure Access for Genomics in the Cloud. Avere Systems.

Black, Aaron. (Nov 2014). Bursting to the Cloud: Deploying a Hybrid Cloud Storage Solution with AWS. Amazon Web Services.

Botstein, David. (NA). Why tumor typing is important. Cold Springs Harbor Laboratory.

Cancer Genome Atlas. (NA) What is Cancer Genomics?. National Institutes of Health.

Cancer Genome Atlas. (NA) The Genetic Basis of Cancer. National Institutes of Health.

Fischer, Ben. (Aug 2013). Ahead of its time: Inova health targets premature births with software partnership. Washington Business Journal.

Inova. (NA). Congenital Disorders Study. Inova Translational Medicine Institute.

Inova. (NA). First 1,000 Days of Life and Beyond. Inova Translational Medicine Institute.

Kodoldt, Dan. (Nov 2014). Brace Yourself for Large-Scale Whole Genome Sequencing. MassGenomics.

McMullan, Dawn. (2014). What is Personalized Medicine?. Genome.

NCI. (April 2015). The Genetics of Cancer. National Cancer Institute.

Offit, K. (2011). Personalized medicine: new genomics, old lessons. Human genetics, 130(1), 3-14.

Regalado, Antonio. (Nov 2014). Google Wants to Store Your Genome. MIT Technology Review.

Tsimberidou, A. M., Iskander, N. G., Hong, D. S., Wheler, J. J., Falchook, G. S., Fu, S., ... & Orcutt, Mike. (April 2012). Bases to Bytes. MIT Technology Review.

Tsimberidou, A. M., Iskander, N. G., Hong, D. S., Wheler, J. J., Falchook, G. S., Fu, S., ... & Kurzrock, R. (2012). Personalized medicine in a phase I clinical trials program: the MD Anderson Cancer Center initiative. Clinical Cancer Research, 18(22), 6373-6383.

Kurzrock, R. (2012). Personalized medicine in a phase I clinical trials program: the MD Anderson Cancer Center initiative. Clinical Cancer Research, 18(22), 6373-6383.

Vence, Tracy. (Aug 2013). Predicting Preterm Birth. The Scientist.

Winslow, R. (2011). Major shift in war on cancer. The Wall Street Journal.

Wheeler, D. A., & Wang, L. (2013). From human genome to cancer genome: the first decade. Genome research, 23(7), 1054-1062.

WUSA9. (Feb 2015). Genomics: The Power to Predict -- Jack's Story. Inova Health System.

4 University of California, San Diego Center for Computational Biology and Bioinformatics

Using Technology to Unlock the Promise of Personalized Medicine

An Abiding Belief in the Transformative Power of Technology

Steve Jobs is one of history's most famous innovators and designers. His name is not just synonymous with a brand—Apple—but with a particular style of personalized technology. The products he and his team introduced—including the iPod, iPhone, and MacBook—changed not only the way we use devices but also how we can interact with the world around us.

Most people know about his place in the history of personal technology, but they might not know about his place in the history of personal medicine. After being diagnosed with pancreatic cancer in 2003, Jobs turned to technology in hopes of finding the cause of his cancer and therefore perhaps a treatment. He went to the Broad Institute at MIT to have his entire genome sequenced, a procedure that at the time cost nearly $100,000.[15] He hoped doctors would be able to find some mutation within his genome, or within his tumor tissue, that would suggest a course of treatment.

Personalized medicine

An emerging practice of medicine that uses an individual's genetic profile to guide decisions made in regard to the prevention, diagnosis, and treatment of disease.[16]

Jobs survived for several years after his initial diagnosis, before eventually succumbing to the disease in 2011. Jobs believed, again, in the power of technology. He felt, as many researchers do today, that technology would enable us to treat individuals based on the particular genetic underpinnings of their condition. This is what is known today as personalized or precision medicine.

At the time Jobs was seeking insight into his cancer, just four years ago, it was only those with his resources who had access to genetics-based personalized medicine. The concept of personalized medicine dates back hundreds of years, but it remained more theory than practice until the last century, as science was able to identify the underlying causes of many diseases. Recent advances in genomics, as well as developments in computational biology, medical imaging, and cancer immunotherapy are enabling scientists to truly personalize diagnosis and treatment.

The cost of genomic sequencing has dropped dramatically in the last few years. This gives doctors and researchers unprecedented access to the genetic causes of certain illnesses. The cost will soon become low enough that sequencing can be employed as a standard diagnostic procedure, much like taking a blood sample or a throat culture.

A patient with pancreatic cancer, or their doctors, could take that diagnosis to a sequencing lab and undergo a complete genomic analysis, which might identify key markers that could point the way toward targeted therapies.

Next Generation Sequencing (NGS)

NGS involves sequencing millions of small fragments of DNA in parallel. These fragments are analyzed by mapping the individual results against the reference human genome. Each of the three billion bases in the human genome is sequenced multiple times, providing insight into unexpected DNA variation.[17]

The Center for Computational Biology & Bioinformatics

Hospitals and research institutions around the country are developing and implementing technical innovations in genomic sequencing. Foremost among these is the Center for Computational Biology & Bioinformatics, or CCBB, at the University of California, San Diego. CCBB provides expertise in managing and analyzing large genomic data sets to provide insights to clinical research labs. CCBB has seven core AWS-supported analysis pipelines which are able to provide next-generation sequencing (NGS) data.

CCBB customizes the logic for analysis for each particular customer as well—each project is uniquely designed for a given set of parameters. Each pipeline is targeted at identifying molecular differences, be they in the genetic material present in a tumor or in the microbiome of patients. These small changes can provide insights into molecular pathways and mechanisms affected in diseased tissues.

NGS data require analysis of "shotgun" DNA chunks, which are read in small segments before being stitched back together like a jigsaw puzzle in order to detect differences between normal samples and disease-carrying samples. CCBB's "boutique" service does not just provide the results of analysis—which many doctors have little to no experience dealing with. Rather, they work to guide customers through the iterative process of analyzing data, giving recommendations, answering questions, and preparing publication materials. CCBB operates like a full-service consultancy, taking on analysis cases from each customer lab.

CCBB also specializes in systems biology, in which they use network analysis on biological data. By running data through AWS-supported applications, CCBB is able to look at vast data sets to analyze, for instance, interactions between genes, thus ascertaining a form of "guilt by association". This is another way to identify gene variants that can cause disease. AWS is particularly adept at the scalable computing required for Big Data processing programs like Apache Spark.

One way this is being applied is tumor neo-antigen candidate selection. Custom analysis pipelines are designed and executed to look for small variant changes within coding genes and rank which are most likely to be immunogenic, or able to actually fight the cancer. This process finds which mutations are expressed by a tumor that might be targetable by

Using Neo-antigens to fight cancer

1. Patient is diagnosed with cancer

2. Patient has blood drawn and tissue samples taken

3. Samples are genetically sequenced

4. Sequence data is uploaded to CCBB AWS servers and run through a pipeline for analysis

5. Variant changes in the coding genes are identified and ranked for likelihood of immunogenic properties

6. If particular neo-antigens are identified in the tumor's genetic makeup, the patient's immune system can be "trained" to recognize and fight those neo-antigens

the patient's own immune system. The goal is to train specific immune cells to recognize tumor neo-antigen expressing cells and to attack it—essentially using the body's own immune system to treat the tumor.[18] This is similar to the way in which the body can be trained to resist disease through vaccination.

Such targeted, highly specific treatment is the promise of molecular-guided precision medicine. Advances in computing are enabling researchers to analyze much more data, much faster, at a much smaller cost. Cloud computing, especially, has enabled parallel processing, massive data storage, and remote data access, all of which provide researchers with unprecedented opportunities for collaborative analysis.

HIPAA

Health Insurance Portability and Accountability Act (HIPAA). The primary goal of this federal law passed by Congress in 1996 is to make it easier for people to keep health insurance, protect the confidentiality and security of healthcare information, and help the healthcare industry control administrative costs.[19]

Partnership with AWS

CCBB relies on the AWS platform for many facets of their service—from the ability of customers to upload sample data remotely, to data storage, data analysis, and the delivery of results. AWS cloud architecture enables CCBB to manage huge amounts of data. Whole genome sequences (WGS) can often weigh in at 150GB per sample, making traditional computing and storage options prohibitively bandwidth-consuming and expensive—especially for large-sample studies. AWS S3 also facilitates remote data access, meaning CCBB's research customers can upload genetic sequence data from anywhere on AWS's HIPAA-compliant secure servers.

Because AWS cloud architecture is scalable on demand, CCBB need only maintain the server space and compute instances that their current workload requires. This means that rather than capital being funneled into server maintenance, it can go toward research and development. Using EC2, CCBB can customize their analysis pipelines to fit each customer's needs. Though there are seven base pipelines, within each of those are many variations. Further, it is the custom statistical analysis in the pipeline that CCBB truly tailors to each customer, setting them apart from many other analysis providers.

“I’m either going to be one of the first to be able to outrun a cancer like this, or I’m going to be one of the last to die from it.”

—Steve Jobs

AWS meets CCBB’s needs for raw computing power, via EC2. But it also meets an often overlooked necessity for any lab doing genetic analysis: storage, and lots of it. The limiting factor in the advent of personal medicine, at this point, is cost. Cost is driven in large part by the realities of storage, which AWS is quickly changing via S3.

With that shifting economic reality comes the emergence of affordable, rapid personal medicine. Though Steve Jobs was unable to win his fight against cancer, he understood that the field was on the cusp of major breakthroughs. “I’m either going to be one of the first to be able to outrun a cancer like this,” he predicted, “or I’m going to be one of the last to die from it.”

As technology advances, many patients in the near future will have much better odds, thanks to labs like CCBB and the window into human genetics they can provide.

Under the Hood: Intel Solutions

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases.

For CCBB: C3 instances are based on high frequency Intel Xeon E5-2680 v2 (“Ivy Bridge”) processors, and are designed for running compute-intensive applications.

M1 instances are based on Intel Xeon Processors.

M3 instances are based on High Frequency Intel Xeon E5-2670 v2 (Ivy Bridge) Processors and provide a balance of compute, memory, and network resources.

M4 instances are based on 2.4 GHz Intel Xeon® E5-2676 v3 (Haswell) processors and provide a balance of compute, memory, and network resources.

R3 instances are based on High Frequency Intel Xeon E5-2670 v2 (Ivy Bridge) Processors and are optimized for memory-intensive applications.

T2 instances are based on High Frequency Intel Xeon Processors with Turbo up to 3.3GHz and are good for workloads that don’t use the full CPU often or consistently, but occasionally need to burst (e.g. web servers, developer environments and small databases).

Endnotes

1. Before the widespread adoption of genetic testing, diagnosis of genetic disorders involved complicated and indirect methods, such as electrophysicological testing.
2. Each genome contains all of the information necessary to build and maintain that organism.
3. A nucleotide is one of the structural components of DNA and RNA.
4. Mayo Clinic Staff. (NA). Diseases and Conditions: Preterm labor. The Mayo Clinic.
5. CDC. (Dec 2015). Preterm Birth. The Center for Disease Control.
6. March of Dimes. (2015). The Impact of Premature Birth on Society. The March of Dimes Foundation.
7. VMC. (2015). About MOMS-PI. Virginia Commonwealth University.
8. NIH. (June 2012). NIH Human Microbiome Project defines normal bacterial makeup of the body. National Institutes of Health.
9. NIH. (June 2012). NIH Human Microbiome Project defines normal bacterial makeup of the body. National Institutes of Health.
10. Cancer research and genomic research are inextricably linked, as a desire to better understand the role of gene variations in cancer progression was one of the impetuses for the Human Genome Project in 1986. Only with a complete human genome sequence available for reference could researchers properly understand the full spectrum of somatic changes that lead to cancer.
11. ITMI is part of Inova Center for Personalized Health (ICPH), which connects researchers, clinicians and consumers to integrate genomic research for patient care, prevention and wellness.
12. One of every nine babies in the United States is born before reaching full gestation, but medical practitioners still know surprisingly little about what causes such premature births—and are thus unable to properly identify mothers at risk.
13. Throughout the study, parents provide updated information about their child, including health status, immunizations, growth, and development measures. This data will be contextualized with home and community information related to nutrition, the environment, potential toxin exposure from parental workplaces, and psychological issues.
14. The 3.2 billion DNA base pairs contained within a human genome requires about 800 megabytes to store in isolation. But additional data *about* each base is also collected during sequencing, and genes are often sequenced many times to ensure accuracy. The final raw data file for a single whole genome may end up closer to 100 gigabytes—although a polished version may be much smaller.

15. Regalado, Antonio. “Steve Jobs Left a Legacy on Personalized Medicine.” *MIT Technology Review.* N.p., 27 Sept. 2013. Web. 30 Nov. 2015.

16. “Personalized Medicine.” *Genetics Home Reference.* U.S. National Library of Medicine, 23 Nov. 2015. Web. 30 Nov. 2015.

17. Behjati, Sam, and Patrick S. Tarpey. What Is next Generation Sequencing? *Archives of Disease in Childhood. Education and Practice Edition.* BMJ Publishing Group, 28 Aug. 2013. Web. 30 Nov. 2015.

18. Schumacher, T., and R. Schreiber. “NEOANTIGENS IN CANCER IMMUNOTHERAPY *Science.* N.p., 3 Apr. 2015. Web. 30 Nov. 2015.

19. “The Definition of Disclosure - HIPAA.com.” *HIPAAcom.* N.p., 10 May 2009. Web. 30 Nov. 2015.