# What Would You Do With a String of 25,000 Letters?

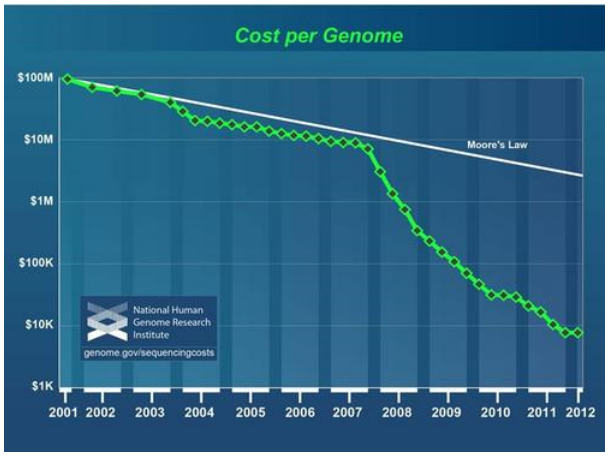
**Robert Plant**, Associate Professor, School of Business Administration, University of Miami
3/27/2013
Comment
14 comments

Login
50% 50%

The human genome is an amazingly complex and intriguing biological artifact, but one that can be broken down into just a few simple constructs, represented by four letters: A, G, C, and T. These correspond to two purines, adenine and guanine; and two pyrimidines, cytosine and thymine.

Adenine and guanine are heterocyclic aromatic organic compounds. Cytosine and thymine are aromatic heterocyclic organic compounds. In addition, nucleic acids are also present in the genetic sequence of Messenger DNA (mDNA), where uracil, represented by the letter U, replaces thymine in the sequence. These then form the basis of the genetic code, and the sequences of these, in the form of triplets, known as nucleotide triplets, or "codons," act to determine the creation of amino acids. As there are only four letters, the maximum number of codons is 64.

These make up the letter sequences that we are familiar with, such as UGA:CGG:ACG:AUG... However, this deceptively simple-looking science (for which Nobel prizes were awarded to Francis Crick and James Watson amongst others) is incredibly complex to map for a human genome, due in part to the number of genes and the need to sequence them in the correct order. In essence this is a big-data problem.




DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program
(Source: National Human Genome Research Institute, Kris Wetterstrand, May 21, 2012)

**Now what do we do with it?**
The goal of mapping the human genome was established in 1990 with $3 billion in funding from the US government, which ultimately led in 2000 to an announcement by President Bill Clinton and British PM Tony Blair that the project was complete, with a 99.9 percent mapping occurring in 2004. So, the big data question is what would you do with this very big data set?

One option is to set up shop as a genome mapping service. For example, RainDance Technologies has developed a platform for targeted sequencing of droplet digital polymerase chain reactions (ddPCRs), a process embraced by cancer researchers such as Professor Michael Griffiths, Director of the Genetics Lab at Birmingham Women's Hospital in England, where his group runs 50,000 samples annually.

The use of DNA sequencing in research hospitals is, of course, not new; however as the costs decrease and the speed of processing goes down the next step in DNA sequencing is aimed at the individual and the development of "personalized medicine." This style of medicine, which, according to the Personalized Medicine Coalition needs greater regulatory and financial systems to be developed, is based upon the use of individuals' DNA sequences or genetic profiles to help physicians more clearly make clinical determinations regarding their patients' unique molecular profiles -- for example, what makes them susceptible to certain diseases or how they would benefit from specific treatments.



**Accessible costs**
With prices for sequencing a human genome currently at around $3,000 to $4,000 this genetic approach to medicine seems to almost be in reach, and with companies such BGI-Shenzhen (which is capable of sequencing DNA through its 128 machines) this will certainly be affordable within five years.

For innovators, entrepreneurs, and owners of such DNA sequencing operations, the systems deployed produce very large amounts of data, BGI reportedly creates 10 terabytes of raw data every day from 30 sequencers running on its 500-node supercomputers, which scales to 43 terabytes for 128 sequencers when fully functioning, or more than 15 petabytes a year.

**Scaling to the problem with cloud**
Computationally this problem space is even more challenging. For example, researchers are establishing datasets for determining the genetic variances that may cause cancer, autism, or other disabilities, and therefore they will have to compare tens of thousands of sequences to identify possible genetic irregularities, which requires substantial processing capabilities.

One approach will be to use commercial cloud platforms that can scale to provide this resource. As this capability is developed it will inevitably lead to big genetic databases in the cloud, allowing researchers to access datasets for individual research projects and study. It is not unreasonable that you will, within the near future, keep your own DNA sequence in a personal genetic cloud-based dropbox for clinicians to use when you have a malady, or even via a DNA App using data compression and encryption on your smartphone for security purposes to prove who you are at airport arrivals or at your bank.

The world of personal Big Genomics Data is about to opened up a whole variety of possibilities but will require, as the PMC has identified, a strong big data framework to ensure that security and safeguards are in place to constrain access and ensure ethical use and deployment. It also marks a new frontier in innovative big data medicine, hopefully one from which we will all benefit.

**Related posts:**

- Big Data Incubator – Frost Venture Partners
- What If Your Medical Data Could Help Save Lives?
- Big Data Explained: What Is MapReduce?

— Robert Plant, *Associate Professor, School of Business Administration, University of Miami*

Email This Print Comment

## COMMENTS

**Newest First** | Oldest First | Threaded View

PAGE 1 / 2 > >>

Keith.Grinsted, User Rank: Petabyte Pathfinder
4/29/2013 | 7:50:56 PM

**Re: what i don't want to know...**
@ Saul yes, I guess we can make our decisions accordingly but where is the mystery in that?

Login
50% 50%

Do you really want to known about everything before it happens/ Surely dealing with stuff after it has happened if half the fun?

Reply | Post Message | Edit/Delete | Messages List | Start a Board

a.saji, User Rank: Megabyte Messenger
4/12/2013 | 1:21:04 AM

**Re: Personalized medicine**

*Sure the technology isn't quite mature yet, but that is more so because of the costly implementation*

Login
50% 50%

@Kiran: No, technology is matured enough, it's the users who are not matured enough.

Reply | Post Message | Edit/Delete | Messages List | Start a Board

Saul Sherry, User Rank: Blogger
4/8/2013 | 7:41:50 AM

**Re: what i don't want to know...**
I wonder if there is scope in that environment for a get out clause SharCo... if you get a null result or something which can't be fixed, they just tell you you are fine. If you get something that can be treated, you get told about it... these systems never work in healthcare though. It needs to be all or nothing.

Login
50% 50%

Reply | Post Message | Edit/Delete | Messages List | Start a Board

Saul Sherry, User Rank: Blogger
4/8/2013 | 7:40:27 AM

**Re: what i don't want to know...**
Keith - that's discounting the real benefit here, early catching and treatment. This can save lives, not ruin them. The threat you have to weigh up is whether you have something that can't be fixed vs something you that can be. It would be a horrible feeling to have missed something that could have saved and prolonged your life!

Login
50% 50%

Reply | Post Message | Edit/Delete | Messages List | Start a Board

SharCo, User Rank: Petabyte Pathfinder
4/5/2013 | 11:32:15 PM

**Re: what i don't want to know...**
I feel the same way. If you get a particularly bad result, then what? You'd be living life, dreading and counting down the days until it's over for you. There are better uses for this data; I just don't think it's this.

Login
50% 50%

Reply | Post Message | Edit/Delete | Messages List | Start a Board

Saul Sherry, User Rank: Blogger
4/3/2013 | 5:55:32 AM

**Big Data Republic**@BigDataRepublic 15 Apr
AllJoyn: A Common Language For Internet Of Things - Open-source lingua franca aims to let IoT devices and services... ow.ly/2FHbY7
Retweeted by Yakamoto
Expand

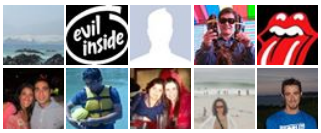
**Hajime Ozaki**@hajozaki 22h
BDR Top 20 #BigData David Smith David Smith List of top tweeters on Big Data from @BigDataRepublic wp.me/p4Aavi-2
Show Summary

**Marcus Borba**@marcusborba 28 Apr
#BigData #Analytics #BusinessIntelligence #FS @informationweek @AllAnalytics @BigDataRepublic @BigDataNetwork @BigDataClub

## LIKE US ON FACEBOOK

Big Data Republic
Like

4,168 people like Big Data Republic.




Facebook social plugin

## ACCOLADES

MEDIA PIONEER AWARDS WINNER 2013
Media Pioneer Award Winner

**Re: Personalized medicine**
@UKjontech, let's not forget the legal issues with medical data.... those guys who are freeing up medical data sets will bascially be able to 'speed up' this process.

Login
50% 50%

Reply | Post Message | Edit/Delete | Messages List | Start a Board

---

Saul Sherry, User Rank: Blogger
4/3/2013 | 5:53:59 AM

**Re: Personalized medicine**
Agreed, we need to look at big data as just the latest tools in this search. Standing on the shoulder of giants etc... Great to see this tech decreasing costs as part of that progress, opening up new opportunities to a wider section of humanity (but we still have a long way to go).

Login
50% 50%

Reply | Post Message | Edit/Delete | Messages List | Start a Board

---

Keith.Grinsted, User Rank: Petabyte Pathfinder
3/31/2013 | 8:30:02 PM

**what i don't want to know...**
Is what my life expectancy is or what diseases I may develop in later life! I'd just rather not know some things.

Login
50% 50%

Reply | Post Message | Edit/Delete | Messages List | Start a Board

---

Keith.Grinsted, User Rank: Petabyte Pathfinder
3/31/2013 | 7:48:50 PM

**what i don't want to know...**
Is what my life expectancy is or what diseases I may develop in later life! I'd just rather not know some things.

Login
50% 50%

Reply | Post Message | Edit/Delete | Messages List | Start a Board

---

Keith.Grinsted, User Rank: Petabyte Pathfinder
3/31/2013 | 7:45:56 PM

**Re: Personalized medicine**
I think we should find every opportunity we can to draw together medical research data. The cures to diseases lie out there somewhere. My mother died of leukemia back in 1975 and there was very little research to go on. We must use whatever means at our disposal to progress this research.

Login
50% 50%

Reply | Post Message | Edit/Delete | Messages List | Start a Board

---

PAGE 1 / 2 > >>