Mayu Takagi
1602-111 St. Clair Ave W
Toronto, ON
M4V 1N5

Google Report

Google Genomics

https://cloud.google.com/genomics/

- Google Genomics is a data analytics product that helps the life science community organize the world's genomic information and make it easily accessible and useful
- Connects to the Google Cloud Platform → applies the same technologies that power Google search and Maps to securely store, process, explore, and share large amounts of data
- Has the power to process complete genomic information of large research projects in seconds
- Can process as many genomes and experiments in parallel (at the same time)
- Google Genomics support open industry standards → can share your tools and data with others, if you choose to do so
- Provides security that meets or exceeds the requirements of HIPAA
- Allows you to monetize he access and usage of your genomics data
    - It hosts information in a storage and costs or using the information are easily billed to clients
- Pricing: USD$0.022/GB Per Month
    - Can enable billing and link account to a credit card or bank account for automatic billing
    - A bill is sent out at the beginning of each month for the previous month's usage
- Cloud Genomics provides a number of public datasets that you can access for free
    - You can also integrate this dataset into your applications
    - Examples of public datasets: 1000 Genomes, The Cancer Genome Atlas (TCGA)
    - Can access these files in BAM, VCF, and FASTA formats
- You can also list your research/data as a public dataset on Cloud Storage

https://www.youtube.com/watch?v=cAV0rj2_puE

- Google Genomics is an API (application programming interface) that is part of the Google Cloud platform
- Through this and other add-ons, you can apply the same technology that power Google Search and Maps to securely store, process, explore, and share large and complex data sets

https://www.youtube.com/watch?v=ExNxi_X4qug

- The cost of DNA sequencing has dropped, and the volume of data has risen
    - The first human genome sequence took 15 years and $3 billion
    - Today, it takes closer to 1 day and $1000
- Google search index is about 100 petabytes, yet search queries only take 0.25 sec

Mayu Takagi
1602-111 St. Clair Ave W
Toronto, ON
M4V 1N5

- Google joined the Global Alliance for Genomics and Health → this was their first public announcement
- Google's role is to contribute to the team that is defining a standard API to promote interoperability
- Hosting public data that is available through the API and building open-source software showing how to work with big genomic data using that API
- API can be connected to access data
- In the US, the National Center for Biotechnology Information implemented the API for sequence read archive
- The European Bioinformatics Institute also implemented the API
- At Google, they have a 3rd implementation
- All 3 of these groups store information using different internal formats BUT the same software can connect to all of them
- There are example data sets in the Genomics Team repository showing how to call the API from Python, Java etc. as well as how to analyze genomic information in 2 modes: by doing interactive queries and massively parallel processing
- Exploring genomic data can be a Challenege because …
    - You are dealing with large amounts of data
    - Many of the computing technologies to work with them were built decades ago
- BigQuery = a tool Google has for exploratory data analysis
    - The underlying technology was built to process trillions of lines of log files at interactive speed
    - How it works
        - Types a sequel query into the browser and press enter (ex. variants of population, Minor allele frequency)
    - Just wait a few seconds and then get results
    - There is also an API so you can call it from scripts or programs
- There are multiple examples in get hub that you can copy and paste or modify to ask real biological questions and get answers in seconds
- Another way to explore genomic data is using MapReduce
    - It uses the same distributing paradigm that Google uses to generate the search index
- MapReduce is a way to process data in parallel by putting lots of servers on it
- MapReduce and BigQuery provide 2 ways to work with data → can go back and forth between them
- Genomic API Broswer - https://web.ticketking.com/LionK/Online/mapSelect.asp

https://www.youtube.com/watch?v=BAAZNH-Wa6A

- As scientist query thousands and millions of genomes, they will need to have scalable technologies for manipulating, analyzing, and interpreting enormous data sets without the time or cost of moving the data from place to place
    - This is where Google believes they can help

Mayu Takagi
1602-111 St. Clair Ave W
Toronto, ON
M4V 1N5

- Google Genomics aims to help the life science community organize the world's genomic information and make it accessible and useful
- A part of Google's infrastructure that supports Youtube can handle 300 hrs of video every minute and Google's search index is over 100PB+ but it can generate search results in a quarter of a second
- That is the equivalent of raw data from 6 Whole Genomic Sequences uploaded to Google's server every minute. It also means they can index and search the equivalent of 1 million whole genomic sequences currently
- There is now increasing recognition on the importance of cloud computing for genomic data analysis
    - In early March, the NIH updated its Genomic Data Sharing policy to allow the Cloud to be used to share and analyze genomic data
- There is mounting evident that cloud based genomic data analysis is the future
- Recent articles in Nature, showcase how much cheaper, faster and flexible Cloud computing resources were compared to local data centers
    - Google Cloud is cited as a key cloud computing resource that is used increasingly by genomic researchers
- Google cloud platform provides both Infrastructure and platform as a service
- Google also built a unique genomics API that powers many genomic specific features
    - Can choose whether you want to choose the infrastructural services, platform services, genomic specific feature set that they have built
- Key benefits of Google Genomics = Store, process, explore, and share
    - Store – enable content aware storage
        - All of this data can be accessed by a web API that is an implementation of the GA4GH (Global Alliance for Genomes & Health) specifications
        - Can store many different data sets
    - Process – allow running of data processing pipelines to analyze private or public genomic data stored on Google Genomics
        - Made a deal with the Broad Institute to bring the popular GADK pipeline as a service in a cloud optimized format → GADK = an analysis tool
    - Explore – enable tertiary analysis of Genomic data using a variety of tools
        - BigQuery allows real-time analysis of big quantities of data
    - Share – enable sharing of these data sets
        - Secure – uses fine-grain access control, 2 factor authentication, security at rest and at transit
        - Security is paramount to Google – have some of the highest security standards in the industry
        - Sign BAAs that ensure the platform is compliment to customer's HIPAA compliance requirements
- How Google Genomics actually works
    - Sequenced data from the sequencer is stored in a variety of different file formats on Google Cloud

Mayu Takagi
1602-111 St. Clair Ave W
Toronto, ON
M4V 1N5

    - It is then copied from Google Cloud to Google Genomics into the Read or Variance Store → at this stage, know its genetic information
    - The API calls are then made to create a variety of queries or analysis on this data
    - If you want to do further tertiary analysis, they also enable specific data cohorts to be exported to BigQuery for further analysis
    - Enable users of varying levels of expertise to access and analyze data
    - If you don't know how to code, you can use a web URI to write simple scripts
- Already have about 1700 WGS (whole genome sequences) as of August 2015 for researchers to use
- National Cancer Institute's TCGA data set – an important data set
- Google Genomics is involved in 2 of the 3 wining bids for NCI's cancer cloud pilot
- ISB is developing a web research portal that allow users to explore the TCGA data set
- Google and the Broad Institute is collaborating → combing the Broad Institute's genomic expertise and Google's engineering expertise
    - Bring the GADK pipeline as a service on Google Genomics – this enables users to not have to worry about structure, platform, or license and simply focus on running that pipeline on their data
- Google and the research community are building a growing repository of open source codes and tools that are available on GetHub for researchers
    - They can use it on their own pipelines or modify it for their analysis
- Google is welcoming other companies to host valuable data sets as well as analytics pipelines to be hosted and shared on Google Genomics

https://www.technologyreview.com/s/532266/google-wants-to-store-your-genome/
November 2014

- Connecting and comparing genomes is what is going to propel medical discoveries and there is growing competition between Amazon, Google, IBM, and Microsoft on who will store the data
- The National Cancer Institute said in October 2014 that they would pay $19 million to move copies of the 2.6 petabyte Cancer Genome Atlas into the Google Cloud
    - Copies of the data are now stored both at Google Genomics and in Amazon's data centers
- The idea is to create "cancer genome clouds" where scientists can share information and quickly run virtual experiments quickly and easily
- Companies like Google and Amazon are encouraging other companies to genomics companies using their cloud

https://www.princeton.edu/news/2018/12/18/google-open-artificial-intelligence-lab-princeton-and-collaborate-university
December 18, 2018

- A new Google AI is scheduled to open in January in Princeton → 2 Princeton University computer science professors will lead it – Elad Hazan and Yoram Singer
- The lab will focus on machine learning

Mayu Takagi
1602-111 St. Clair Ave W
Toronto, ON
M4V 1N5

- Primary focus being developing efficient methods for faster training of learning machines

https://techcrunch.com/2018/12/03/deepmind-claims-early-progress-in-ai-based-predictive-protein-modelling/

DeepMind claims early progress in AI-based predictive protein modelling

November 2018

- Google-owned AI specialist, DeepMind has claimed a milestone in demonstrating the usefulness of AI to help with the task of predicting 3D structures of proteins based soley on its genetic sequence
- Understanding protein structures is important to disease diagnosis and treatment, and could improve scientists; understanding of the human body
- Modelling the 3D structure of protein is a highly complex task
- DeepMind's approach rests on years of prior research in using big data to try to predict protein structures → they are applying deep learning approaches to genomic data
    - Their methods rely on using deep neural networks trained to predict protein properties from its genetic sequence
    - Properties include:
        - Distance between pairs of amino acids
        - Angles between chemical bonds that connect those amino acids
- AlphaFold (DeepMind's AI) was submitted to CASP (Community Wide Experiment on the Critical Assessment o Techniques for Protein Structure Prediction) this year
    - CASP has deemed their work "unprecedented progress in the ability of computational methods to predict protein structure"
    - They placed first among the teams that entered
- They first trained a neural network to predict a separate distribution of distances between every pair of residues in a protein
- These probabilities were then combined into a score that estimates how accurate a proposed protein structure is
- A separate neural network was then trained to use all distances as a collection to estimate how close the proposed structure is to the right answer
- A new method was then used to try to construct predictions of protein structures and search known structures that matched its predictions
- They trained a generative neural network to invent new protein fragments, which were used to continually improve the score of the proposed protein structure
- Scores were them optimized through gradient descent → this resulted in highly accurate structures
- This technique was applied to entire protein chains
- DeepMind describes the results they have achieved as "early signs of progress in protein folding"
- However, they state that it is still early for the deep learning approach to have quantifiable impact on treating diseases

Mayu Takagi
1602-111 St. Clair Ave W
Toronto, ON
M4V 1N5

Blogpost by DeepMind: https://deepmind.com/blog/alphafold/

https://www.teslarati.com/googles-neural-network-takes-a-step-closer-to-predicting-disease-using-dna/
December 5, 2018

- AlphaFold – a neural network designed by Google's AI company DeepMind was declared winner out of 98 AI competitors in the biannual protein folding prediction contest called the Critical Assessment of Structure Prediction (CASP)
    - For predicting 25 out of 43 proteins shapes given using genetic sequence alone
    - Second place predicted only 3

https://cloud.google.com/customers/desktop-genetics/
How Desktop Genetics uses the Google Cloud Platform

- In genomic analysis, a single cell typically generates around a terabyte of raw data
- London-based startup, Desktop Genetics is a frontrunner t in the field of bioinformatic software development
- Desktop Genetics uses the Google Cloud Platform to enable massive data processing on-demand with rapid scaling on Google Computer Engine
- They specialize in handling and analyzing large amount of data for laboratories
- Google Cloud Storage saves machines and standard references flies, keeps automatic backups, and allows servers to work across each other, reducing the need to replicate functionality
- The company can guarantee security and privacy to 3rd parties – follows HIPAA compliance
- The development team can easily switch between each other's servers without a need to provision SSH keys every time
- Google Cloud Platforms allows companies to immediately take on large-scale bioinformatics projects
- Google Cloud Platform's multizone architecture can also be used to set up demo servers and present services o global clients without connection issues

Deep Variant

https://www.technologyreview.com/s/609647/google-has-released-an-ai-tool-that-makes-sense-of-your-genome/
December 4, 2017

- In December of 2017 Google released a tool called DeepVariant – a machine that uses the latest AI techniques to build a picture of a person's genome from sequencing data
    - Turns high-throughput sequencing readouts into a picture of a full genome
    - Automatically identifies small insertion and deletion mutations and single-base-pair mutations in sequencing data

Mayu Takagi
1602-111 St. Clair Ave W
Toronto, ON
M4V 1N5

- High-throughput sequencing became widely available in the 2000s and has made genome sequencing more accessible HOWEVER, data produced using this system offered only a limited- error-prone snapshot of a full genome
- Tools such as GATK, VarDict, and FreeBayes can be used to interpret readouts from high-throughput sequencing but these software programs typically use simpler statistical and machine-learning approaches to identify mutations
    - They do so by attempting to rule out read errors
- DeepVariant was developed by researchers from the Google Brain team that focuses on developing an applying AI techniques and Verily, an Alphabet subsidiary that focuses on life sciences
- The team collected millions of high-throughput reads and fully sequences genomes from the Genome in a Bottle (GIAB) project
- They then fed the data to a deep-learning system and tweaked the parameters of the model until it learned to interpret sequences data
- In 2016, DeepVariant won first place in the Precision FDA Truth Challenge – a contest run by the FDA to promote more accurate genetic sequencing
- This shows that deep learning can be used to automatically train systems and boost progress in genomics
- DeepVariant will be available on the Google Cloud Platform
- Google is adding machine-learning features to heir cloud platform to attract anyone who might want to tap into the latest AI techniques

https://github.com/google/deepvariant

- DeepVariant maintains high accuracy even for error-prone sequencing conditions
- DeepVariant is fast – using Google Cloud Platform, a whole human genome analysis can be completed in as little as 70 minutes
- DeepVariant is cost-efficient – Using Google Cloud Platform, storing a holw genome costs $2-3
- DeepVariant can be used for non-human species
- DeepVariant is easy to use
- DeepVariant currently requires Python 2.7
- It can also be built and run on any standard Linux computer – don't have to be run on Google Cloud Platform

https://www.forbes.com/2007/09/12/genomics-wojcicki-brin-biz-sci-cx_mh_0912_23andme.html#3677bb3067bc
September 12, 2007

- Google invested $3.6 million into 23 and me when it was still a start-up
- $2.6 million of 23andMe's funding then went to paying back a loan from Sergey Brin (co-founder of Google)

https://www.forbes.com/sites/stevensalzberg/2017/12/11/no-googles-new-ai-cant-build-your-genome-sequence/#52cbf1bf5774

Mayu Takagi
1602-111 St. Clair Ave W
Toronto, ON
M4V 1N5

- In contrast to what the *Wired* article (https://www.wired.com/story/google-is-giving-away-ai-that-can-build-your-genome-sequence/), Deep Vairant does not assemble genomes
    - It is a program for identifying small mutations – mostly changes of a single letter (called SNPs) → knowns as variant calling or SNP calling
    - There are many programs that can do this with accuracy of 99.9%
- Deep Variant is a variant caller
- In 2016, the Google team released a prepint on bioRxiv (https://www.biorxiv.org/content/early/2016/12/21/092890) that shows that their method is more accurate (used only a limited data set) than an earlier method called GATK
    - This is only a pre=print so it was NOT published in a peer-reviewed journal
- Google didn't compare their program to other variant calling programs so it is hard to say if it better or worse than all of the others
- While other programs run on commodity hardware, Google's DeepVariant requires a large, dedicated grid of computers working in parallel
    - Some companies have had to invest in new GPU-based computer hardware to run DeepVariant

https://blog.dnanexus.com/2017-12-05-evaluating-deepvariant-googles-machine-learning-variant-caller/

- DeepVariant applies the InceptionTensorFlow framework (https://www.tensorflow.org), which was originally developed to perform image classification.
- It converts a BAM into images similar to genome browser snapshots and then classifies the positions as variant or non-variant.
- The first part is to **make examples** that represent candidate sites.
    - This involves finding all of the positions that have even a small chance of being variants with a **very sensitive caller**.
    - Finally, multi-dimensional pileup images are produced for the image classifier.
- The second part is to **call variants** using the TensorFlow framework.
    - passes the images through the Inception architecture that has been trained to recognize the signatures of SNP and Indel variant positions.
- Both components are computationally intensive.
- The **call variants** step can be accomplished much faster if a GPU machine is available.
- Using Google's specially designed TPU hardware (https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu) will make this step much faster and cheaper
- DeepVariant is much more accurate compared to others methods in SNPs
    - However, its accuracy comes at the price of computational intensity

https://www.genengnews.com/topics/omics/dnanexus-platform-offers-google-developed-deepvariant/

Mayu Takagi
1602-111 St. Clair Ave W
Toronto, ON
M4V 1N5

DNA nexus Platform offers Google-Developed DeepVariant
December 13, 2017

- Cloud-based DNAnexus Platform is now offering DeepVariant bioinformatics tool in a pilot program

https://www.businessinsider.com/google-calico-ancestry-dna-genetics-aging-partnership-ended-2018-7
August 1, 2018

- Ancestry had been partnering with Google-owned company Calico, that studies aging and longevity.
- In July 2015, the companies announced their partnership in a press release, however both companies have since then disclosed little about what the research partnership did
- The assumption is that Calico was interested in Ancestry's genetic data to identify commonalities among people who live a long time
- Neither company has published any research from the collaboration
- Calico states that some of the results of its research with Ancestry will be published in a peer-reviewed journal soon

https://beta.canada.com/technology/the-under-the-radar-google-company-chasing-immortality/wcm/4973ae9e-2702-48e5-b0f1-5a01ebc06d05/html
December 29, 2018

- Calico is an acronym for California Life Company
- The company is a division of Google's parent company Alphabet
- It is 5 years old
- Operations remain highly secretive
    - Its website only lisyt a few studies and research tie-ups
- Calico Labs' office is an hour north of Google's headquarters, up in Oyster Point, south San Francisco
- Calico was founded by Art Levinson, a biotech entrepreneur and former CEO of Silicon Valley pharmaceutical Genentech, now Apple's chairman
    - He convinced Google co-founder Larry Page to open a company that had one mission: achieve immortality
- Calico is looking for a magic pill to cure aging
- Last January, they demonstrated that naked mole rats are not more likely to die as they get older → this research appeared to back up Calico's goal but no therapies have emerged
- A year after Calico was founded, the company announced a partnership with AbbVie – a Chicago-based maker of arthritis drug Humira – a $1.5 billion partnership on developing anti-aging drugs
    - This summer, the partnership was extended

Mayu Takagi
1602-111 St. Clair Ave W
Toronto, ON
M4V 1N5

    - Each company pledges an extra US$500 million ($1 billion combined) to "discover and bring to market new therapies for patients with age-related diseases"
    - They will be conducting research until 2022 when it will enter Phase 2 – human testing
    - The 2 companies have agreed to share costs and profit equally
    - In November, they hired Dr. Garret FitzGerald, an Irish professor who is known for his theory about "molecular clocks" and the idea that genes operate on 24 hour cycles
- Other prominent figures in the company include Cynthia Henyon, Dr. David Botstein (chairman of genetics at Stanford University), Robert Cohen (from Genentech – developed groundbreaking cancer drugs)
- In March, Calico made headlines when AI expert Daphne Koller quit, the second executive to quite in several months
    - When outsiders asked what is going on at Calico, the company had no response

https://endpts.com/a-monster-discovery-deal-between-abbvie-and-googles-calico-gets-a-new-lease-on-the-lab-with-1b-more-to-back-aging-research/
December 30, 2018

- Calico has built a big team of 150-plus around an HQ base in South San Francisco with plan to add more

http://fortune.com/2017/12/14/google-alphabet-anti-aging-calico/
December 14, 2017

- In December of 2017, Calico's chief computing officer announced in a conference hosted by CB Insights that current research, while is in its earliest stage, is being done on mouse models
    - 750 mice separated into 5 groups with different diets
    - Idea is to explore how calric intake influences overall health
- They are also tracking the growth of yeast cells to probe how cellular aging affects the behavior of cells and how they begin to break down

https://www.vox.com/science-and-health/2017/4/27/15409672/google-calico-secretive-aging-mortality-research

- Even after 5 years, Calico has done only a little over a dozen (12) press releases → which only outlined broad descriptions of collaborations with outside labs and pharmaceutical companies
- Multiple sources have tried to reach out for interviews, however they were met with no response
- Even when a reporter tired to talk to people who used to work at Calico, or Calico's outside collaborators, no one provided descriptive information about what the company is doing → only vague details
    - Everyone seems to be highly secretive about what the company is doing
- This seems to go against Google's pride for being a leader on transparency and open culture

Mayu Takagi
1602-111 St. Clair Ave W
Toronto, ON
M4V 1N5

http://www.bio-itworld.com/2018/03/07/wuxinextcode-announces-genomics-partnership-google-cloud.aspx

WuXi NectCODE Announces Genomics Partnership with Google Cloud

March 7, 2018

- Wuxi NextCODE's programs and applications such as GORdb, QuXi NextCODE secondary analysis, the Sequence Miner case-control research application, and the Clinical Sequences Analyzer clinical interpretation system will now be hosted on Google Cloud and available on the Google Cloud Launcher marketplace
- Key Google genomics and research tools will also be integrated and accessible on the WuXi NextCODE platform
    - This will include the DeppVarian tsecondary analysis pipeline, BigQuery, and other analysis pipelines and tools available through Google Cloud Platform

https://ai.googleblog.com/2018/04/deepvariant-accuracy-improvements-for.html

DeepVariant Accuracy Improvements for Genetic Datatypes

April 19, 2018

- Google launched DeepVariant v0.6 earlier this year → included major accuracy improvements
    - Increased accuracy for whole exome sequencing and polymerase chain reaction sequencing
-

https://www.medscape.com/viewarticle/904705

CEO Feinberg Leaving Geisinger to Manage Healthcare for Google

November 9. 2018

- David T. Feiberg, MD, MBA, president and CEO of Geisinger is joining Google to lead its healthcare ventures which includes Google Fit app and Google Genomics
- They were able to improve accuracy for new datatypes by including representative data in the training process
- The majority of DeepVariant's training data is from the first benchmarking genome released by DIAB, HG001 (this is commercially available)
    - By using many replicates and different datatypes of HG001, they generated millions of training examples which helped DeepVariant leanr to accurately classify many datatypes and generalize to datatypes it had never seen before
- In V0.5, they focused on exome data (subset of genome that directly codes for proteins)
    - Increased exome accuracy by adding a variety of WES datatypes provided by DNAnexus to DeepVariant's training data
- The newest release of DeepVariant, v0.6, focuses on improved accuracy for data that has undergone DNA amplification via polymerase chain reaction (PCR) prior to sequencing.

Mayu Takagi
1602-111 St. Clair Ave W
Toronto, ON
M4V 1N5

- Prior to v0.6, training data for DeepVariant was exclusively PCR-free data
- By adding PCR+ examples to DeepVariant's training data, provided by DNAnexus, there have been significant accuracy improvement for this datatype (60% reduction in indel errors)
- DeepVariant has been released as an open source software to encourage collaboration and to accelerate the use of this technology to solve real world problems

https://www.biorxiv.org/content/biorxiv/early/2018/03/20/092890.full.pdf
March 20, 2018
Creating a universal SNP and small indel variant caller with deep neural networks

https://www.youtube.com/watch?v=goBFt3B976A
July 25, 2018
Speeding up Research in Genomics

- Announcement: New partnership with the National Institute of Health –making many f the high value public and controlled data sets funded by the NIH available to users on Google Cloud
- Aim is to democratize access to data to researchers
- Variant Transforms: a tool that lets you take your VCF data, process genome data, and import it directly to BigQuery
- BigQuery – manage data warehouse solutions