Introducing the Databricks Unified Analytics Platform for Genomics

Colby Ford

At the recent Spark+AI Summit 2018, Databricks unveiled a few amazing platform enhancements. One of the most exciting ones (in my opinion) is the new Unified Analytics Platform for Genomics. As a computational biologist, much of my research time is spent waiting on something to run. This is due to the fact that many bioinformatics tools simply aren't built for the modern data platform, not to mention the cloud. Thanks to the amazing people at Databricks, the Azure cloud is ready for scalable genomics workloads!

Accelerating Discovery

Since the first human genome was sequenced over a decade ago, the cost and time to sequence subsequent genomes has drastically reduced. This means that the amount of data being generated from genomics-related work has grown exponentially (and is expected to continue to grow).

This also means that the potential for scientific discovery is increasing. From specialized drug treatments to combating infectious diseases to making genetically modified foods (albeit controversial), the area of genomics is definitely gaining momentum.

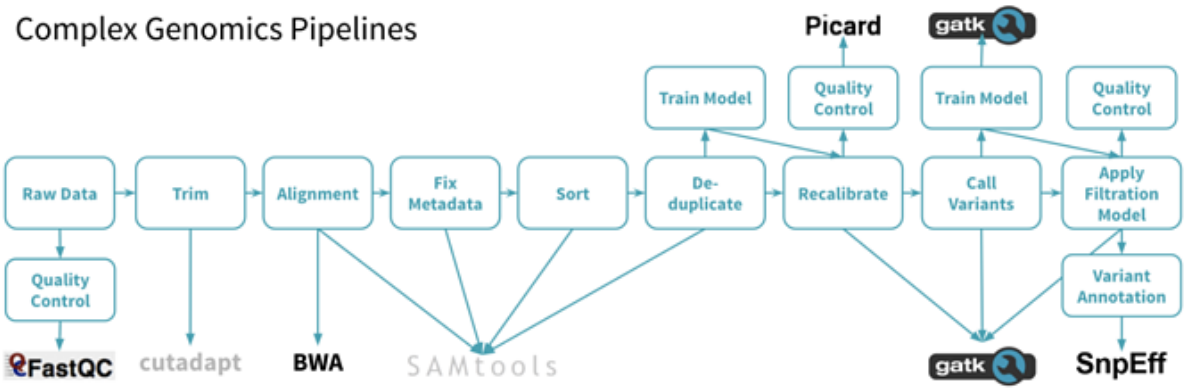
The rise in popularity of genomics calls for an overhaul of the traditional methodology in order to meet the research demands. Databricks has answered the need with its Unified Analytics Platform for Genomics.

Traditional Challenges

When it comes to the modern genomic workflow, there are plenty of challenges that come into play:

- **Complex Pipelines**

  To effectively process terabytes or petabytes of data, there are many steps that go into the pipeline to transform the data, align sequences, call variants, annotate them, and then analyze the output. This inherently produces bottlenecks due to the complex nature of the workflow.



- **Antiquated Analytics Tools**

  In the sample pipeline above, you'll notice that there are plenty of tools that are used along the way. For example, SnpEff, BWA, GATK4, and others are used to perform very specialized tasks on the genetic data. Many of these tools are command line-based. Some are only single-threaded and therefore very slow. And, even if the software has some notion of parallelism, it's often only been

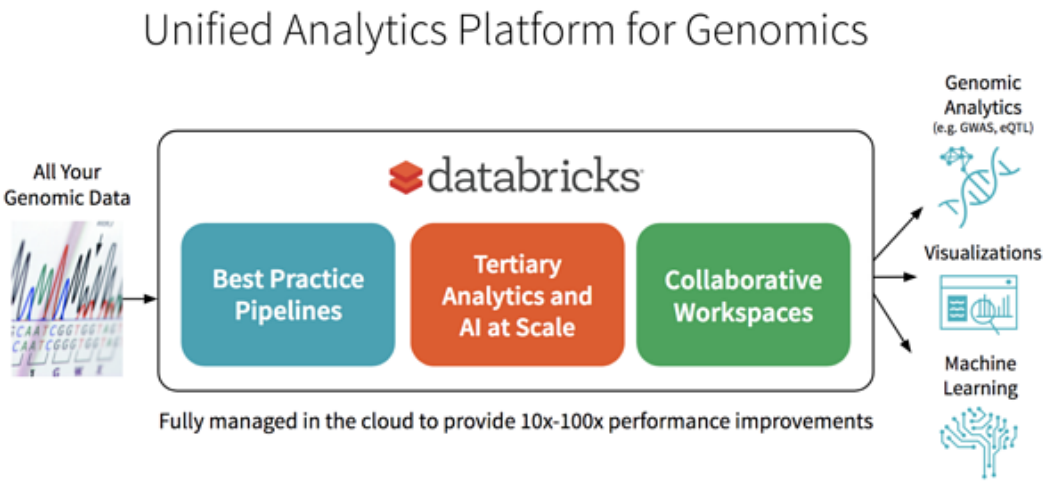
implemented using Message Passing Interface (MPI). This is certainly the case at many universities. Furthermore, it's quite rare that any of these important packages have been fully implemented in Spark and therefore are not "embarrassingly parallel".

Links:						
FastQC	cutadapt	BWA	SAMtools	Picard	GATK	SnpEff

- **Diverse Teams**

  From the white coat-wearing geneticists/biologists/lab monkeys to the data science/bioinformatics team to the data engineering group, there are lots of moving parts to a research project. Many times, each of these groups use different technologies and therefore use different types of files or programming languages to complete their work. This makes it hard to work together and transport results back and forth.

The Databricks Unified Analytics Platform addresses each of these aforementioned struggles in very innovative ways.



- **Simplified Genomics Pipelines**

  Using pre-built pipelines created by industry standards, you can easily pull in data, analyze it, visualize it, and output the results.

- **Interactive, Scalable AI**

  New enhancements to traditional libraries allow for scalable analysis using the power of Databricks. Currently, distributed versions of Joint Variant Calling, GWAS, PheWAS, eQTL, and machine learning frameworks are available in the unified platform. Databricks reports that these are optimized to run in parallel 60x-100x faster than the open-source (traditional) versions. Plus, because of this speed, this allows you to interactively work with your data rather than having to wait hours or days for something to finish running.

- **Team Collaboration**

  Since the Databricks notebook environment can be shared with anyone you wish, this means that it's easy to collaborate among your teammates. Plus, you can have notebooks for each part of the pipeline (in varying languages) for each subset of your team. This will all work on a single storage location, which can help you keep your datasets tidy and all in one place.

Resources

- To view sample genomics notebooks from Databricks, click here.
- To view genomics pipeline examples from the Azure Databricks documentation, click here.
- Read the blog post from Dr. Ion Stoica from the Spark Summit here.
- View BlueGranite's collection of free Azure Databricks resources here.

Want to learn more about how BlueGranite can help with your data and AI needs? Contact us today!

About The Author

Dr. Colby Ford is the Principal of Life Sciences at BlueGranite. Coming from a background in mathematics, data science, and computational biology, he combines this expertise to architect scalable solutions in the Azure cloud. Using R, Python, and Spark, he puts AI and bioinformatics to work to provide valuable insights from data. Outside of BlueGranite, Colby is an avid researcher in infectious diseases and human genomics. Check out Colby's website at www.colbyford.com.