Scaling your Genomics Pipeline in the Cloud with Azure Databricks

Colby Ford

Long before "Big Data" was a buzzword in the business realm, geneticists, bioinformaticians, and computational biologists had been dealing with large-scale *-omics* data for quite some time. This data includes DNA/RNA samples, annotated variants, genotype/phenotype analyses, and more. When it comes to the large amounts of data and numerous steps it takes to get to the insights for which you're looking, processing genomic data is no small feat. Plus, biological data comes in a variety of shapes and formats, each adding its own bit of complexity to your analysis process.

Azure Databricks, as I'm sure you're familiar with, is the premier platform for performing massively parallel processing tasks in the cloud. This platform serves as an optimized Spark service for users looking to scale up their ETL and Machine Learning pipelines. However, recent efforts in the life science development space have made some common bioinformatics tools available on the Spark platform.

Today, we'll introduce a specialized runtime for Health and Life Sciences soon to be available on Databricks and highlight a few Spark-based libraries that you can begin using today.

Databricks Runtime for Health and Life Sciences

The Databricks Runtime for Health and Life Sciences is a specialized version of Databricks that has been optimized for working with genomic and biomedical data. It is a component of Databricks' Unified Analytics Platform for Genomics.

POWER YOUR PIPELINES **<1.5 hours** Run your alignment and variant calls in less than an hour and a half	**RAPID RESULTS** **60-100X faster** Tertiary analytics 60-100x faster on Databricks compared to open source Apache Spark^{TM}	**MORE EFFECTIVE TEAMS** **30% + productive** Leading healthcare company improved productivity 30% with Databricks' unified analytics
Source: https://databricks.com/product/genomics		

To sign up for the HLS Runtime Preview, click here.

Included in the HLS Runtime:

- A fast, scalable DNASeq pipeline
- Spark SQL optimizations for common query patterns
- Hail 0.2 integration
- Popular open-source libraries, optimized for performance and reliability
  - ADAM
  - GATK
  - Hadoop-BAM
- Reference data (GRCh37 or 38, known SNP sites)

In addition to the support for a few Spark-based genomics libraries (which we'll discuss in a bit), this runtime also includes support for various file types seen in genomics data.

For example, just as you would use `spark.read.format("csv").load("file.csv")` to easily read in CSV files, you can use a very similar approach to read and write VCF and BGEN files.

VCF and BGEN

```
## Read in VCF data
df = spark.read.format("com.databricks.vcf").load("file.vcf")

## Write out VCF data

df.write.format("com.databricks.vcf").save("newfile.vcf")

## Read in BGEN data
df = spark.read.format("com.databricks.bgen").load("file.bgen")

## Write out BGEN data

df.write.format("com.databricks.bgen").save("newfile.bgen")
```

An example Databricks notebook for working with variant data can be found here.

DNASeq Pipeline

A common pipeline for genomic analysis is the Genome Analysis Toolkit (GATK) by the Broad Institute. GATK creates best practice workflows for various tasks from data pre-processing to variant discovery and beyond. Using these best practices allows for research labs to have a standardized operation pipeline for performing analyses.

In the HLS Runtime, Databricks now includes a GATK-compliant DNASeq pipeline for short read alignment, variant calling, and variant annotation and an RNASeq pipeline for handling short read alignments and quantification.

These pipelines make it easy to get started analyzing genomics data using popular techniques such as SnpEff annotation, STAR alignments, and ADAM. Plus, this allows for the use of a variety of other common input formats such as SAM, BAM, CRAM, Parquet, and FASTQ.

An example Databricks notebook for using the DNASeq pipeline can be found here.

Hail 0.2

Hail is an open-source, scalable framework for genomic data analysis and exploration. This project is supported by the Neale Lab out of Harvard Medical School. In the most recent edition of Hail (0.2), support for Spark (and thus Databricks) has been enabled.

hail-logo-cropped

Hail allows for the many different types of analyses from Genome-Wide Association Studies (GWAS), annotation, expression analysis, and visualization. Hail is designed to scale from a single laptop to a cluster with little to no code changes and is also meant for use on datasets that do not fit in memory.

Once you have the HLS runtime enabled in Databricks, getting started with Hail is quite simple.

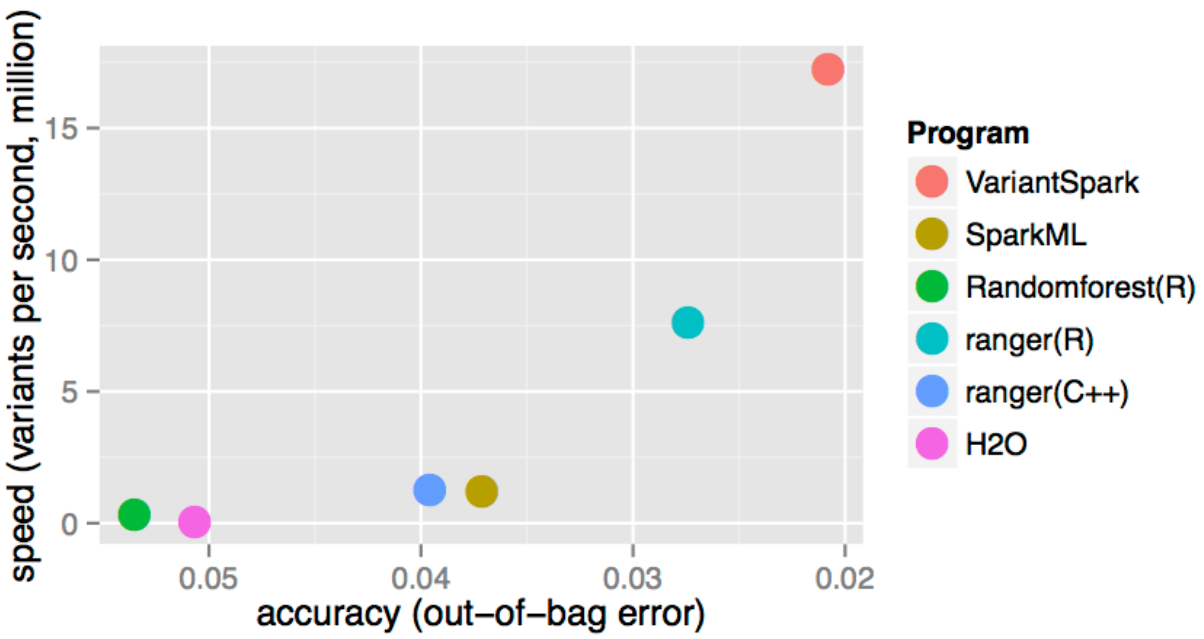
```
## Set the environment variable: ENABLE_HAIL=true

import hail as hl
hl.init(sc, idempotent=True)
```

An example Databricks notebook for using Hail can be found here.

VariantSpark

VariantSpark, by O'Brien et al. (2015), is an interesting library for Spark. While other genomics packages provide general bioinformatics analysis of genetic datasets, this library provides a machine learning analysis framework for analyzing genomic variants using the Spark engine.

VariantSpark prides itself in being an efficient (fast) and accurate contender against other machine learning implementations, such as Spark's own MLlib, randomForest in R, H2O, and more.



Runtime vs. accuracy of six available implementations showing that VariantSpark has the highest accuracy and is substantially faster than its competitors, enabling point-of-care diagnostics within 30 minutes instead of 24h.

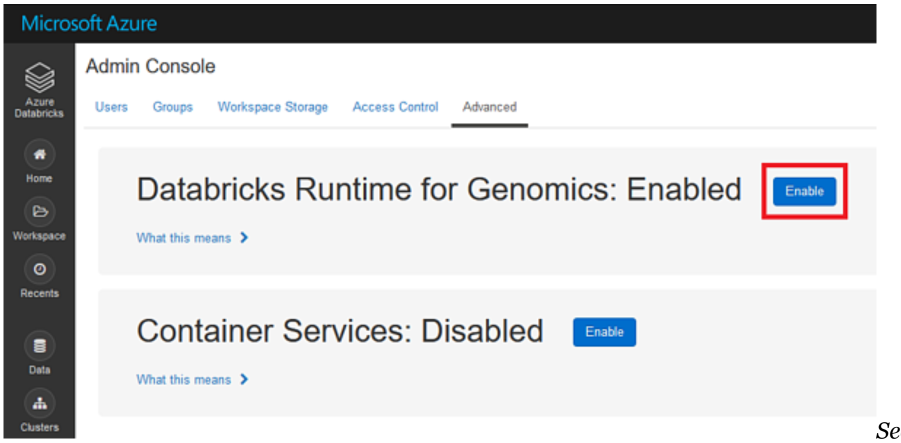
Source

VariantSpark is developed for data with many samples and many features. It includes machine learning methods for clustering (k-Means) and classification (Cursed Forest). Though VariantSpark was originally developed for genomic variant data, it can cater to any feature-based dataset, such as methylation, transcription, and even non-biological applications.

An example Databricks notebook for using VariantSpark can be found here.

Getting Started

Enabling the Genomics Runtime is easy. Simply go into the Admin Console in your Databricks workspace, click the Advanced tab, then enable the Databricks Runtime for Genomics.



Databricks Runtime for Genomics: Enabled

Container Services: Disabled

See the Azure Databricks Documentation for genomics pipeline examples here.

Whether your bioinformatics practice is completely on-premise today or is growing into the Azure cloud, BlueGranite can help you get started using Azure Databricks. In addition to scaling up your analysis pipelines, setting up additional services, such as a flexible storage and visualization solutions, is also important. Since Azure Databricks easily integrates with Azure Storage (such as blob storage or Data Lake Store) and Power BI, using the Azure cloud from end-to-end is the best way to scale your Health and Life Science practice for faster, deeper insight.