Give Your Genomics Pipeline a *Glow* Up in Azure Databricks

Colby Ford

In some of my previous posts, I've talked about using the Databricks Runtime for Genomics for scaling up common bioinformatics analyses using Apache Spark. Today, I want to highlight a rather new package that further enhances our ability to perform genomics-based workloads in Azure Databricks: Glow. Project Glow began out of a partnership between the life sciences team at Databricks and the at Genetics Center at Regeneron.

About Glow

Glow is an open-source and independent Spark library that brings even more flexibility and functionality to Azure Databricks. This toolkit is natively built on Apache Spark, enabling the scale of the cloud for genomics workflows.

Glow allows for genomic data to work with Spark SQL. So, you can interact with common genetic data types as easily as you can play with a *.csv* file. *Learn more about Project Glow at projectglow.io*.

What's Included?

Glow already includes easy-to-use functions for reading and writing common file formats like VCF, BGEN, Plink, or GFF3. In addition, there are tools for performing the following secondary and tertiary analyses:

Secondary Analyses	Tertiary Analyses
• Perform variant quality control • Perform *liftOver* genomic conversions • Perform variant normalization • Split multiallelic variants • Prepare genomics data for machine learning	• Parallelize common bioinformatics tools with *Transformer* • Utilize Python statistics libraries • Perform GWAS regression tests

Learn more about the features of Glow here: https://glow.readthedocs.io/

Why Do We Need Scale?

Genome-wide association studies (GWAS) correlate genetic variants with a trait or disease of interest. These types of studies are effective in the identification of particular mutations and how they affect the disease in question. Traditionally, these analysis are performed by bioinformaticians and genetics and workstations, which have a limit in their processing power.

As genetic sequencing becomes cheaper and more prevalent and as study cohorts have increased in size to millions, there is a need to robustly engineer GWAS to work at scale. Luckily, the Azure cloud, Apache Spark, and Databricks are built for just that!

Demo Video

In the following video, I give a quick overview of some nice features from the Genomics Runtime in Azure Databricks and how to get started using the Glow package.

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.