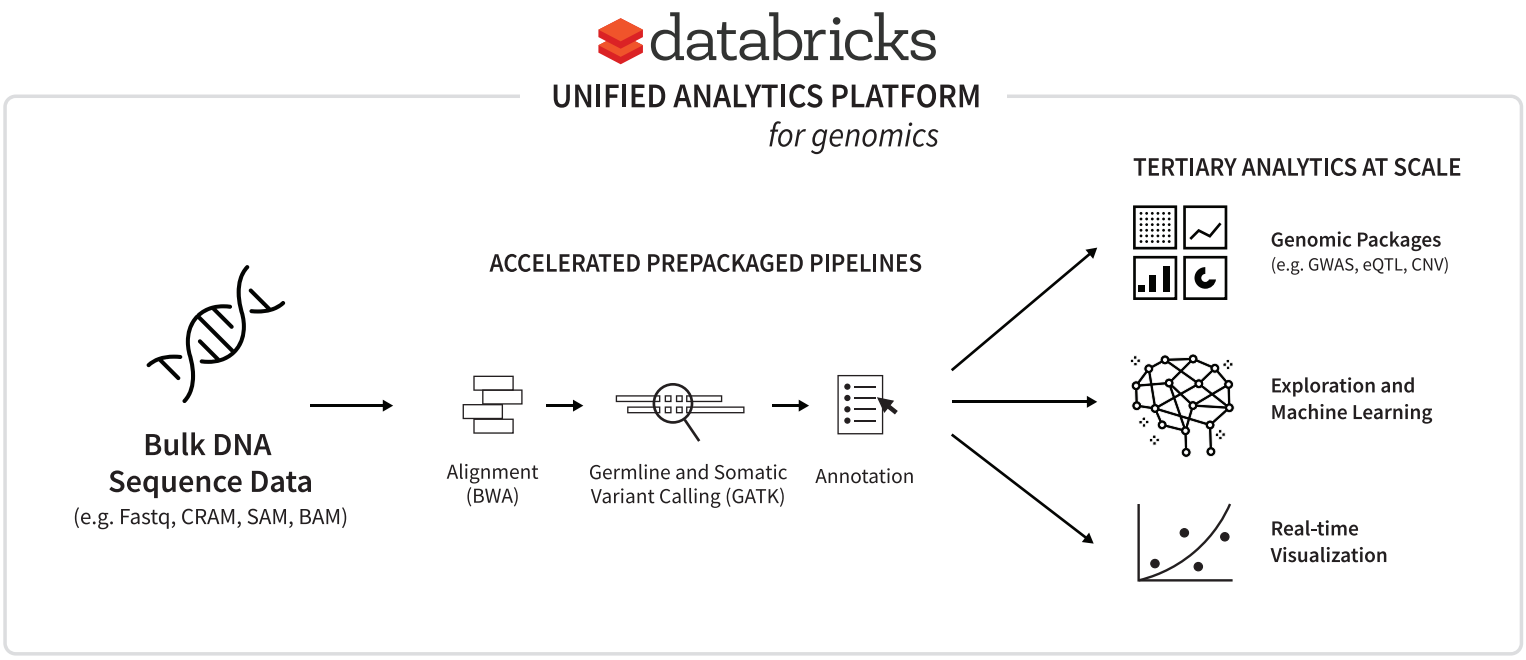
databricks

Unified Analytics Platform for Genomics

Over the last decade, the time and cost of sequencing a genome has dropped exponentially enabling researchers to generate petabytes of genomic data. Yet, transforming this data into life changing discovery and targeted treatments has never been more challenging. Bioinformatics teams are handicapped by infrastructure and pipelines too complex to manage while downstream analytics teams are unable to interactively explore or scale their genomic studies, slowing down critical research. The Databricks Unified Analytics Platform for Genomics provides the scale and speed bioinformatics teams need to accelerate discovery with a collaborative platform for genomic data processing and interactive tertiary analytics at petabyte-scale.

Accelerate Discovery with Unified Analytics

The Databricks Unified Analytics Platform provides prebuilt genomics pipelines seamlessly connected with cutting edge tertiary analytics and AI packages in a collaborative, easy-to-use and fully-managed cloud platform that scales for any size study.



Managed in the cloud to provide 10x-100x performance improvements

Find Answers Faster with Genomic Analysis at Scale

Fast, Easy-to-Use Pipelines

- Reduce complexity with 1-click setup, prebuilt bioinformatics pipelines on a fully-managed platform
- Securely connect to read, variant, and feature data in AWS or Azure and run variant calls at speeds nearly 4x faster than Edico with workflows running in parallel
- Launch and scale clusters in <5min with a few simple clicks in a user friendly interface

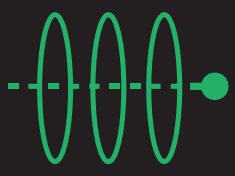
Interactive Tertiary Analytics

- Interactively explore large read and variation datasets in real-time with queries optimized to run at high speed
- All your analytics in one platform with prepackaged genomic analytics (such as Joint Variant Calling, GWAS, PheWas and eQTL), popular machine learning libraries and visualizations
- Analyze genomic variants across hundreds of thousands of samples with tools like Hail which are preinstalled and optimized for the cloud
- Seamlessly connect processed genomic data with downstream analytics in one platform for faster results

Collaborative, Flexible Workspaces

- Enable researchers, computational biologists and bioinformaticians to iterate together in real-time with collaborative workspaces
- Explore data efficiently using your favorite languages: SQL, R, Python, Java, and Scala
- Improve reproducibility across teams with standardized genomic workflows and detailed revision histories to track changes to computational design

POWER YOUR PIPLINES



<1.5 hours

Run your alignment and variant calls in less than an hour and half

RAPID RESULTS

60-100X faster

Tertiary analytics 60-100x faster on Databricks compared to open source Apache Spark™

MORE EFFECTIVE TEAMS

30% + productive

Leading healthcare company improved productivity 30% with Databricks' Unified Analytics

databricks

Industry Leading Benchmarks

In benchmarks against industry leading solutions our pipelines have shown to achieve faster processing speeds while remaining concordant with GATK4 at roughly the same compute cost.

PERFORMANCE

30x Coverage Whole Genome

PLATFORM	REFERENCE CONFIDENCE CODE	CLUSTER	RUNTIME	APPROX COMPUTE COST	SPEED IMPROVEMENT
Databricks	VCF	13 c5.9xlarge (416 cores)	24m29s	$2.88	3.6x
Edico	VCF	1 f1.2xlarge (fpga)	1h27m	$2.40	—
Databricks	GVCF	13 c5.9xlarge (416 cores)	39m23s	$4.64	3.8x
Edico	GVCF	1 f1.2xlarge (fpga)	2h29m	$4.15	—

30x Coverage Whole Exome

PLATFORM	REFERENCE CONFIDENCE CODE	CLUSTER	RUNTIME	APPROX COMPUTE COST	SPEED IMPROVEMENT
Databricks	VCF	13 c5.9xlarge (416 cores)	6m36s	$0.77	3.0x
Edico	VCF	13 c5.9xlarge (416 cores)	19m31s	$0.54	—
Databricks	GVCF	13 c5.9xlarge (416 cores)	7m22s	$0.86	3.5x
Edico	GVCF	1 f1.2xlarge	25m34s	$0.71	—

300x Coverage Whole Exome

PLATFORM	REFERENCE CONFIDENCE CODE	CLUSTER	RUNTIME	APPROX COMPUTE COST	SPEED IMPROVEMENT
Databricks	GVCF	50 c5.9xlarge (1600 cores)	2h34m	$69.30	(No competitive solutions at this scale)

ACCURACY

	PRECISION	RECALL	F SCORE
SNP	99.34%	99.89%	99.62%
INDEL	99.20%	99.37%	99.29%

Concordance vs GIAB NA24385 high confidence calls on PrecisionFDA Truth Challenge dataset (according to hap.py)

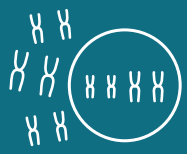
For more details on our benchmarks visit: www.databricks.com/DNASeq

Use Cases

Databricks Unified Analytics Platform for Genomics helps healthcare and life sciences organizations accelerate genomic workflows by delivering on a wide range of use cases

BEST PRACTICE BIOINFORMATICS PIPELINES

Databricks provides Apache Spark™-accelerated pipelines for analyzing germline variants and bulk and single-cell RNA-seq data that drop single-sample analysis times from hours down to tens of minutes.



SCALABLE, CROSS-COHORT STUDIES

Our optimized platform enables order-of-magnitude performance improvements when running genotype/phenotype associations on hundreds of thousands of samples and supports joint genotyping across large cohorts.

ENABLING ML FOR GENOMICS

Whether using ML to filter out erroneous variant calls or using deep learning to predict the effect of non-coding variation, Databricks seamlessly connects best-of-breed ML libraries to genomic data.

INTERACTIVE GENOMIC ANALYSES

Run queries in real-time across large read and variation datasets to accelerate discovery while integrating your genomic data with large clinical datasets and population databases to improve the richness of your study.

Healthcare and Life Sciences Customers

REGENERON science to medicine®

HUMAN LONGEVITY, INC.

SANFORD HEALTH

Contact us to learn more about the Databricks Unified Analytics Platform for Genomics or start a POC today at databricks.com/genomics.

databricks