Scaling Genomic Data Management with TileDB

A cloud-native solution for population-scale variant analysis

[tile]DB
DESIGNED FOR DISCOVERY™

- ✔ Organize
- ✔ Structure
- ✔ Collaborate
- ✔ Analyze

Up to 97% cost reduction
vs VCF file-based approaches

Run complex variant queries fast
<30 seconds

A trusted research environment
enables federated queries across organizations

At a Glance

Population genomics is transforming precision medicine through national biobank initiatives and large-scale sequencing efforts, but faces significant technical challenges in managing massive variant call data. Traditional VCF formats become unsustainable beyond a few thousand samples due to their monolithic structure and inability to handle large-scale queries. The "N+1" problem and limitations in joining phenotypic data create additional bottlenecks. TileDB offers a solution through its multidimensional array-based architecture, enabling efficient storage and analysis of genomic data while facilitating secure data sharing and collaborative research through its trusted research environment.

[tile]DB Carrara is the platform for fast & scalable analysis of variant datasets

Power variant analysis at biobank scale

TileDB offers a highly scalable computational platform, combining an efficient storage format with the distributed power of the cloud.

Store petabyte-scale data cost-effectively

TileDB can store any data modality without compromising performance, including tables, variants, single-cell, imaging, proteomics and any modality that emerges in the future—all while reducing costs by up to 97% compared to file-based approaches.

Run complex variant queries in less than 30 seconds

TileDB delivers analysis of newly sequenced genomes at unprecedented speed, achieving a seven-hour clinical turnaround of diagnoses that once took days.

Support federated queries between namespaces & organizations

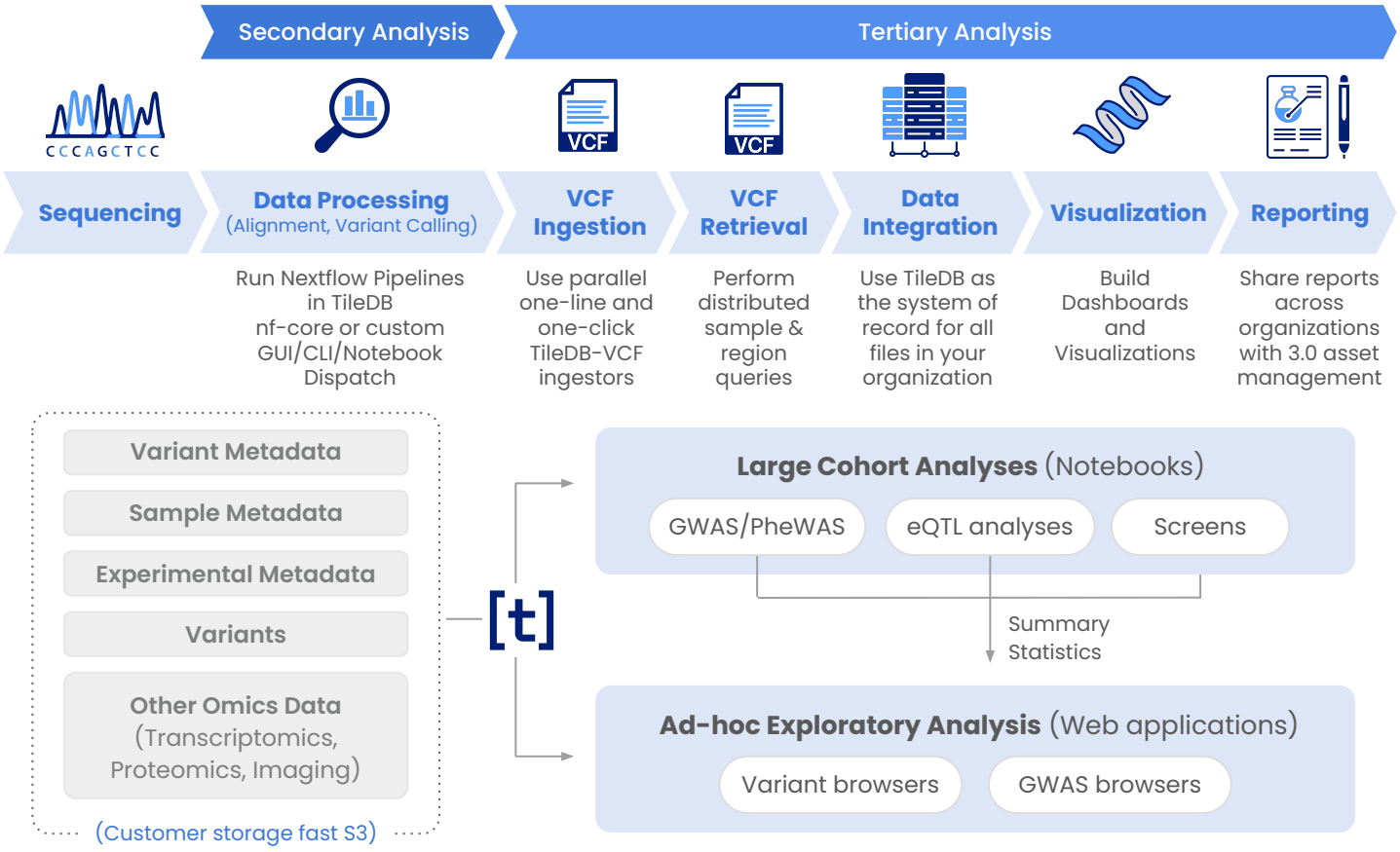
TileDB's trusted research environment facilitates precise analysis of comprehensive genomic repositories, such as UK Biobank, while maintaining rigorous privacy protocols, safeguarding sensitive data like sample identities and individual genetic profiles.

Reduce risk with secure global collaboration

TileDB enables secure collaboration across international databases, transcending individual projects and geographic boundaries all while implementing FAIR practices and adhering to GDPR, HIPAA and SOC 2 Type 2 compliance.

Get started with genomics at:
tiledb.com

Why [tile]DB Carrara for Genomics



Technical Benefits

TileDB's population genomics solution is architected around the TileDB-VCF library, which drives efficient and lossless storage, access and exporting of variant data and provides advanced security, management, scalable compute, and visualization features.TileDB-VCF is built on top of the TileDB array engine, and models population VCF data as 3-dimensional sparse arrays. TileDB-VCF offers a range of benefits:

Performance
Optimized for quickly slicing variant records by genomic regions across multiple samples, with features implemented in C++ for speed.

Solves the N+1 problem
Rapidly adds new samples, scaling storage and update time linearly. gVCFs are recommended for better handling of reference/no-call blocks.

Multiple APIs
Offers C++, Java, and Python APIs alongside a command-line interface.

Integration with other omics data
Links genomic data with transcriptomes for GxE or genotype-to-phenotype studies, compatible with TileDB SOMA for multiomics experiments.

Compressibility
Efficiently stores samples in a compressed, lossless manner, using columnar format to apply different compressors based on data types.

Cohort-level stats
Provides allele counts and zygosity for internal allele frequency calculations, allowing for summary transformations.

Notebooks for reports
Jupyter notebooks in TileDB Cloud allow reproducible variant analyses, stored as TileDB arrays.

Dashboards
Custom dashboards can be built in RShiny or ipywidgets, with options for self-hosted applications.

Optimized for cloud
Inherits features from the TileDB core array engine, ensuring speed and optimization for cloud storage like Amazon S3 and Google Cloud.

Data-annotation separation
Supports external annotation tables for efficient queries and updates without revisiting original VCF files.

AI & ML support
Facilitates AI/ML workflows, saving models as TileDB arrays and enabling vector search capabilities.

Learn more about TileDB population genomics solutions on **TileDB Academy**: https://shorturl.at/YF10m
Diagnose critical illness within hours in the NICU with rapid variant analysis at scale: https://shorturl.at/5KZEh