6M.2.DEMO

Data Analysis: Run Scalable, Cost-Controlled Analysis for Data Enrichment

Rapid annotation of GTEx genomic variants with the TOPMed Annotation Pipeline

We demonstrate rapid, reliable, secure, cost-effective FAIR genome annotation

Goal:

Rapidly and reliably aggregate variant annotations for large volumes of whole genome sequence data, with results that are findable, accessible, interoperable, and reusable: FAIR.

Method:

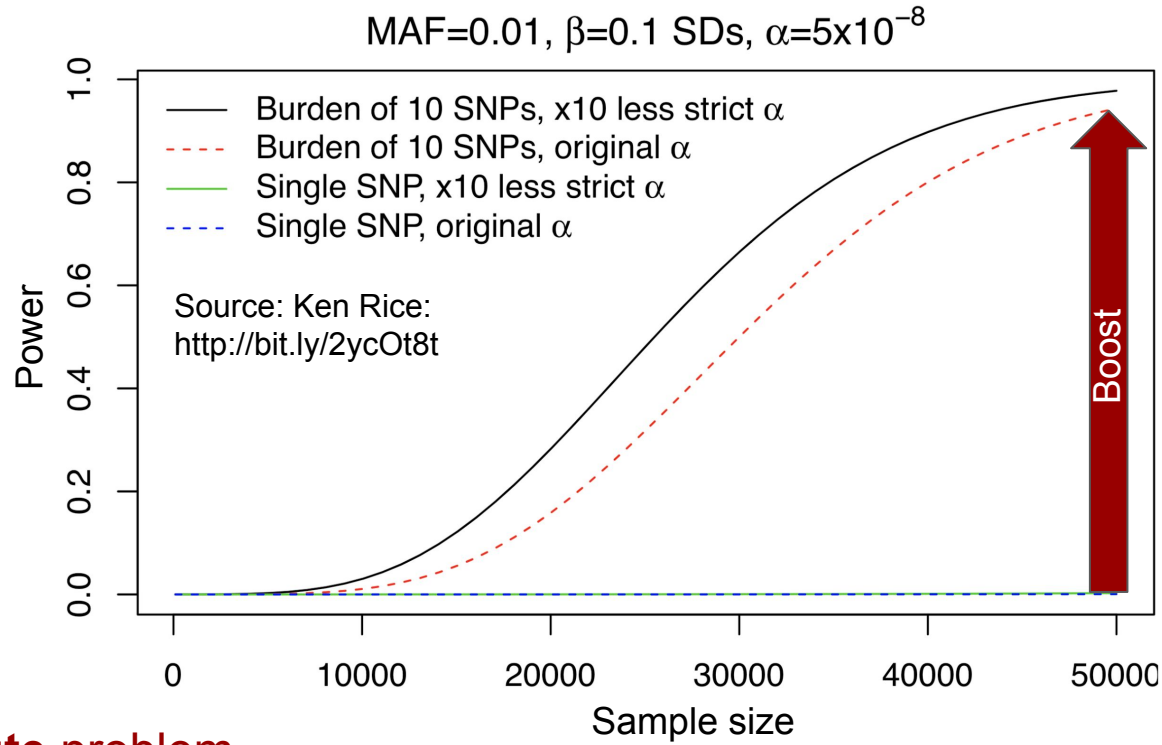
Leverage cloud computing and Commons tools to run the Whole Genome Sequence Annotator (WGSA) on many GTEx genomes, with identifiers, and reproducible pipelines used pervasively to ensure FAIRness.

Rationale: We can boost the statistical power of association studies by aggregating variants

Problem: Large quantities of genome data are now available. But genome-wide association studies that depend on rare variants lack statistical power

Solution: Aggregate rare variants to boost power

Approach: Combine annotations from many sources to provide a comprehensive “genome map”



This is a **big data** and **big compute** problem

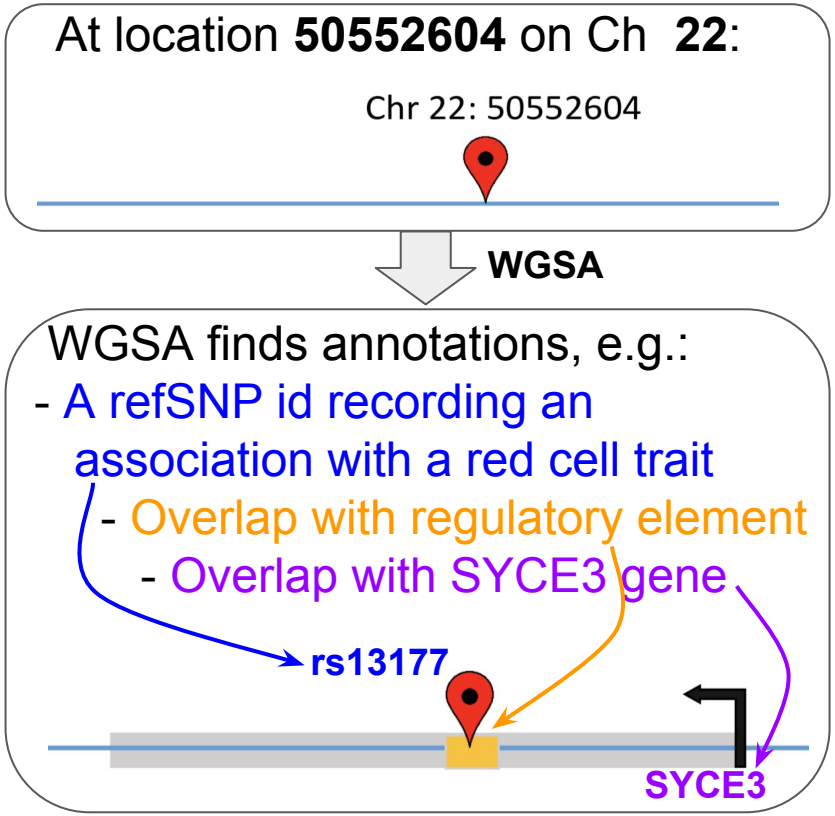
Rationale (2): We can boost the statistical power of association studies by aggregating variants

70 databases are used for annotating variants: e.g., NCBI, Ensemble, UCSC, ENCODE, Roadmap, dbSNP

TOPMed's Annotation pipeline uses WGSA to identify and assignment annotations

We **address 3 big challenges** to enable reliable use on big data:
- Scaling to big cloud data and compute
- Reliable, secure, and inexpensive execution
- FAIR execution and results: Findable, Accessible, Interoperable, Reusable

WGSA: https://sites.google.com/site/jpopgen/wgsa

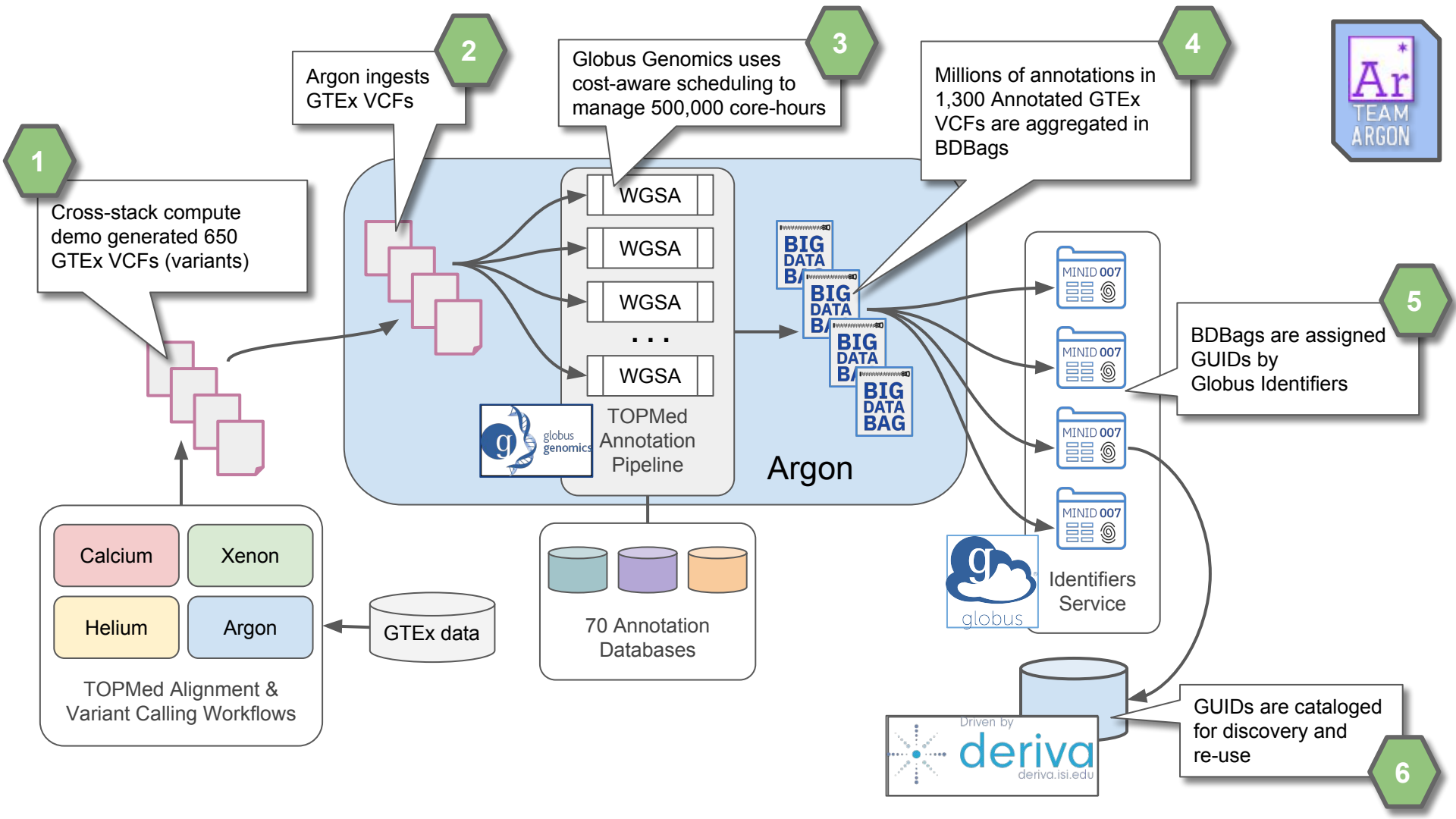


Approach

We create a high performance, parallel implementation of TOPMed's Annotation pipeline that uses WGSA to annotate variant (VCF) files from GTEx

We use this pipeline to create annotated variant files that can then be used in genome-wide association studies

We leverage Commons tools to perform these tasks efficiently, reliably, cost-effectively, and FAIRly


1
Cross-stack compute demo generated 650 GTEx VCFs (variants)
2
Argon ingests GTEx VCFs
3
Globus Genomics uses cost-aware scheduling to manage 500,000 core-hours
4
Millions of annotations in 1,300 Annotated GTEx VCFs are aggregated in BDBags
5
BDBags are assigned GUIDs by Globus Identifiers
6
GUIDs are cataloged for discovery and re-use
Ar TEAM ARGON
WGSA
WGSA
WGSA
WGSA
globus genomics
TOPMed Annotation Pipeline
Argon
BIG DATA BAG
MINID 007
MINID 007
MINID 007
MINID 007
globus
Identifiers Service
Calcium
Xenon
Helium
Argon
TOPMed Alignment & Variant Calling Workflows
GTEx data
70 Annotation Databases
Driven by
deriva
deriva.isi.edu

Results*

- 3 billion variants from 605 GTEx genomes processed against 70 databases
- 42,000 compute jobs using 500,000 core-hours on Amazon cloud
- Millions of annotations generated and made available to community in BDBags named by Minid GUIDs

Total cost of computation performed $32k vs. $78k
if our cost-aware scheduling had not been employed

* Expected: Computations are completing

We leverage Commons technologies

Lightweight digital identifiers used to identify data products through the lifecycle

Interactive exploration and cohort formation across Commons data

Scalable data bundles, based on Library of Congress standards, used for data exchange

Cost-optimized, reliable, **cloud computation** with parallel pipelines for scale

DATS-based KC7 Crosscut Metadata Model and GTEx databases for data ingest

Infrastructure for auth, data management, discovery across clouds and resources

BDBags & Minds

Minids provide a simple and well-defined identification mechanism that allows a scientist to create a reference to a BDBag (or any other type of data) on the web as a single, easily shared URL. Minid URLs dereference to a “landing page” that provides basic metadata about the published entity, such as the author, publication date, title/description, location of the data, and a checksum of the data that can be used to verify the data integrity. Minids are implemented by the Globus Identifier service.

BDBags provide a file container mechanism that ensures dataset integrity, completeness, and provenance. BDBags also provide a mechanism for ensuring that privacy restrictions and data use agreements can be honored. A BDBag provides a *blueprint* of what a complete data set should look like. Scientists can share BDBag instances that contain only references to restricted data, with confidence that only those parties with proper access to the restricted data can fully reconstitute the bag. BDBags can be named by Minids and can reference other data via Minids.

Minimal Viable Identifiers (Minids)

Lightweight identifiers that support simple creation/use

- Unique identifier (ARK)
  - E.g., ark:/57799/b9040f
  - Or compact identifier (minid:b9040f)
- Standard minting/resolution services
- KC2 core metadata (creator, date, name)
- Checksum ensures data is verifiable
- BDBags for multi-file datasets

Easy to use: CLI, Python SDK, R SDK, JSON-LD REST API

minid

Minimal Viable Identifiers (minids) provide a lightweight way of uniquely and unambiguously identifying research data products. Find out more at https://fair-research.org/tools/minid.

Identifier	ark:/57799/b9040f
Created On	2018-07-03 19:06:48.965758
Locations	https://raw.githubusercontent.com/DataBiosphere/identifier-interoperability/master/README.md https://github.com/DataBiosphere/identifier-interoperability/blob/master/README.md
Checksums	690a921e4a076fe889fdee13791b50a79e9a9d636cdb3ac1cd015b1991e43d01 (sha256)

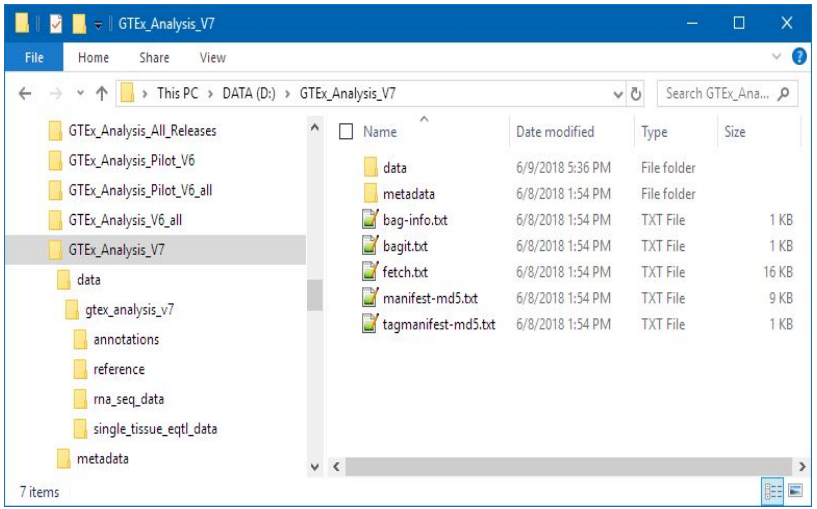
Metadata	{
	"title": "Identifier Interoperability"
	}

Big Data Bags (BDBags)

Standards-based, portable file container that stores hashed manifests of both local and remote content

- Data consistency guarantees via checksum algorithms
  - MD5, SHA1, SHA256, SHA512
- Multiple file transfer protocol support
  - HTTP, FTP, S3/GCS, Globus Transfer
- Multiple identifier resolution support
  - Ark/Minid, DOI, DataGUID
- Secure access to protected data
- Integrated provenance metadata (RO)

Easy to use: CLI, Python API, GUI



DERIVA

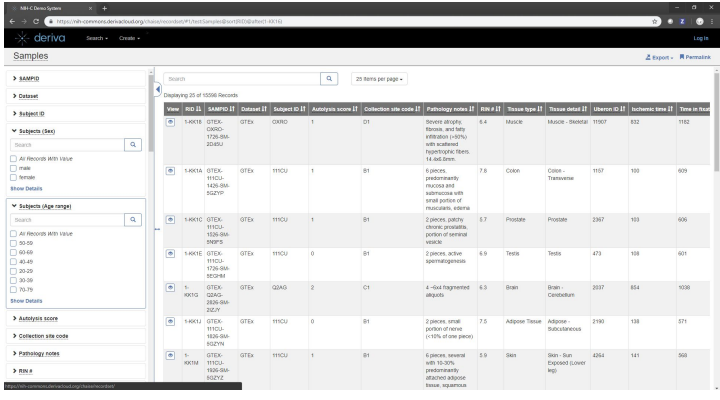
Navigation and discovery of all data from source through secondary and tertiary derived data.

- Integrated management of all data
  - CCMM instances, derived data, user defined collections
- Powerful data discovery and organization with rich models
- Rapid definition and BDBag export of virtual cohorts
- Rich policy with fine-grained access control
- Dynamically adapts to changing data collections

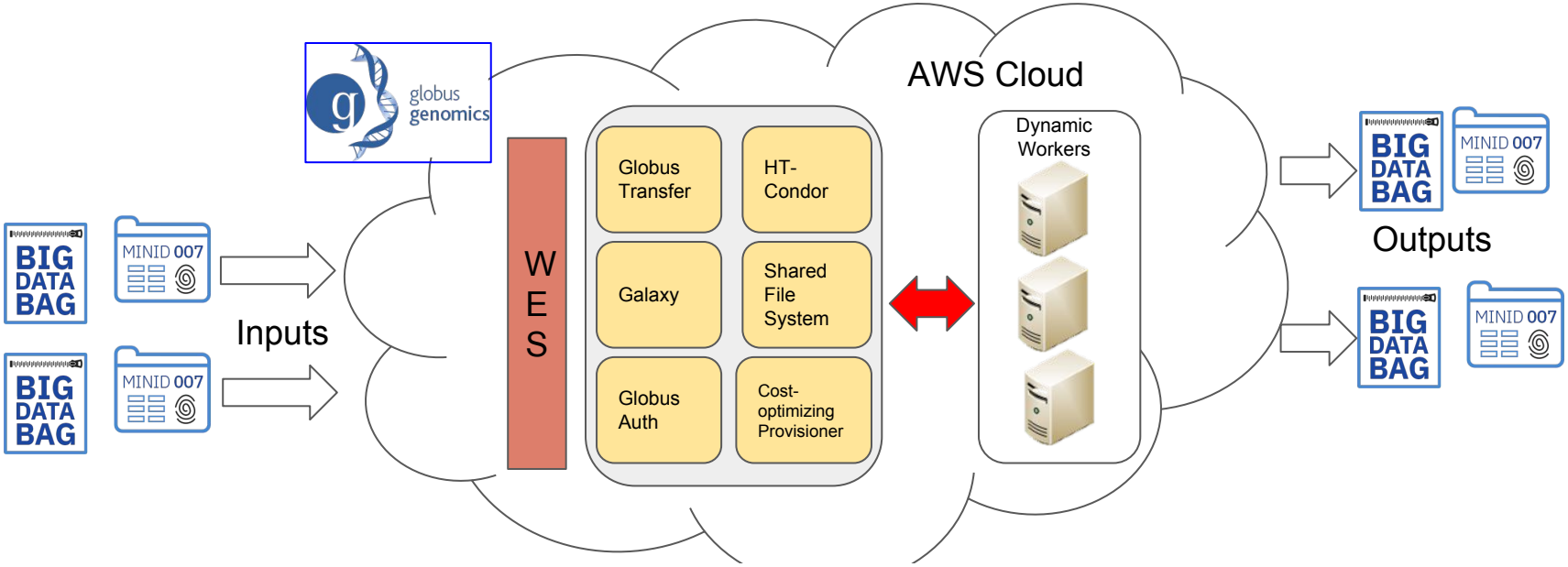
Easy to use: Browser GUI, CLI, Python SDK, Javascript SDK, REST API

Driven by deriva
deriva.isi.edu

Interactive exploration and cohort formation across Commons data

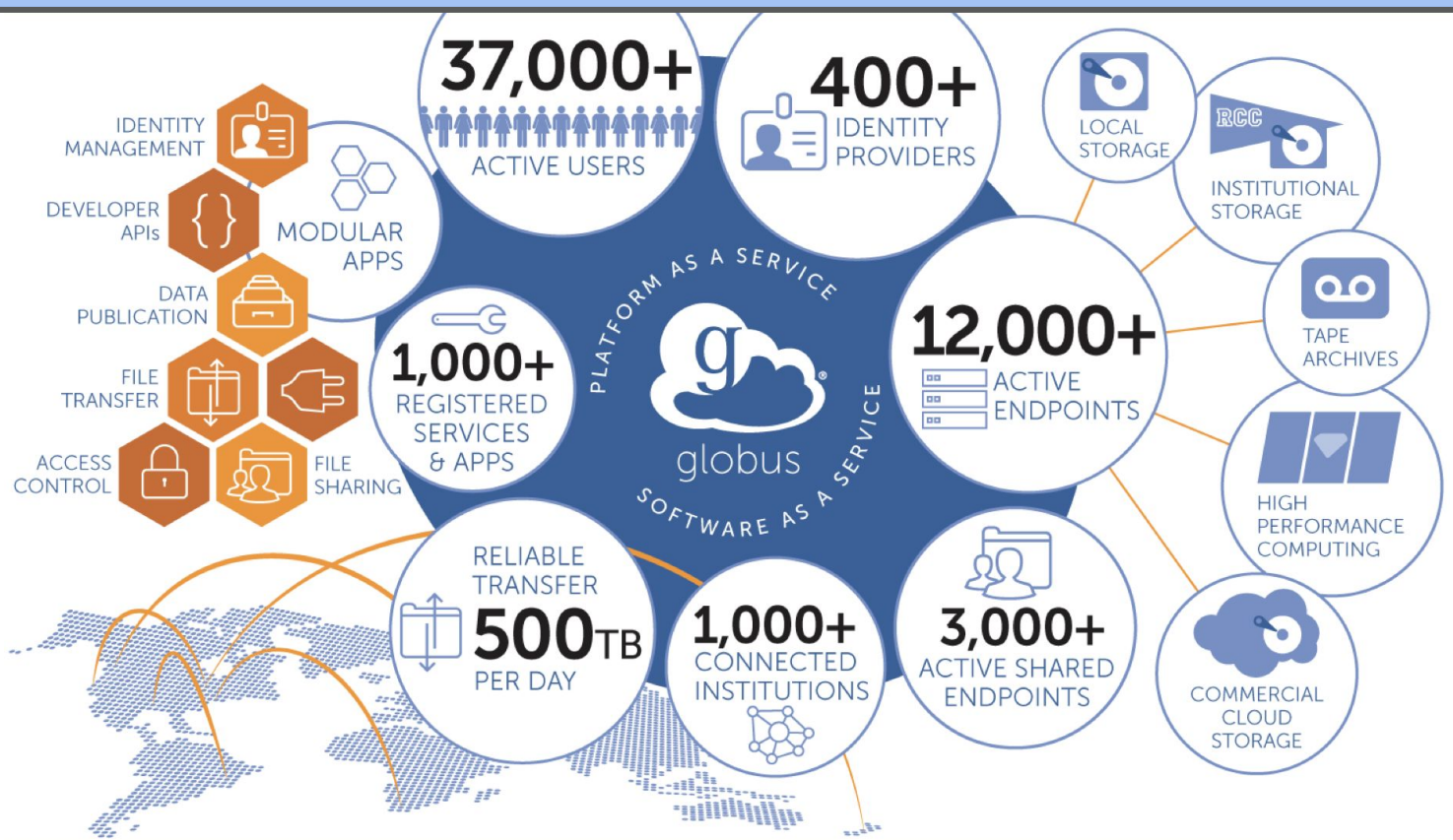


Globus Genomics



We leverage **cloud computation methods** developed by the Globus Genomics team. These enable analysis pipelines to be scheduled securely and reliably onto many cloud computers (high performance), selected to minimize cost (cost optimization). Inputs and outputs are packaged in BDBags and referenced by Minids, providing FAIRness.

Globus


37,000+
ACTIVE USERS
400+
IDENTITY PROVIDERS
LOCAL STORAGE
RCC
INSTITUTIONAL STORAGE
IDENTITY MANAGEMENT
DEVELOPER APIs
DATA PUBLICATION
FILE TRANSFER
ACCESS CONTROL
FILE SHARING
MODULAR APPS
1,000+
REGISTERED SERVICES & APPS
PLATFORM AS A SERVICE
globus
SOFTWARE AS A SERVICE
12,000+
ACTIVE ENDPOINTS
TAPE ARCHIVES
HIGH PERFORMANCE COMPUTING
RELIABLE TRANSFER
500TB
PER DAY
1,000+
CONNECTED INSTITUTIONS
3,000+
ACTIVE SHARED ENDPOINTS
COMMERCIAL CLOUD STORAGE

Commons advantages demonstrated

- Reuse of workflows, robust sharing of results, and reproducibility of every element (KC1)
- Naming of data via GUIDs (KC2)
- High-performance parallel computation and cost-aware cloud provisioning (KC4)
- Secure data access and analysis (KC6)
- All via well-defined APIs (KC3)

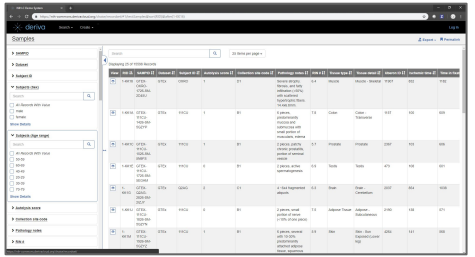
Output Data

Individual results are assigned a GUID (Minid)

`https://identifiers.globus.org/ark:/57799/b95E1DXjoyity0`

GUIDs will be indexed in DERIVA

`https://nih-commons.derivacloud.org`



Complete set is being added to the Full Stacks repo for reference by other DCPPC Teams.

`https://github.com/dcppc/full-stacks/pull/41`