



# Bioconductor - Microsoft Genomics

Jass Bagga- Software Engineer  
Erdal Cosgun, Ph.D.- Sr. Data Scientist



# Genomics is key to Precision Medicine



## Precision medicine



Early detection



Disease prediction



Personalized treatment



Targeted drugs



Gene therapy

Research institutions

Fundamental research

Academic Medical Centers

Clinical research

Pharma & Biotech

Drug development

Provider Systems

Clinical decision support

# Microsoft genomics solutions



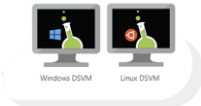
## Research & Discovery

**Genomics Notebooks for Azure**  
preconfigured Jupyter notebooks

**Bioconductor on Azure**  
bioinformatics & data science tools

**Genomics Data Science VM**  
preconfigured Azure VM templates

**Genomics Data Lake**  
collection of public genomics datasets



## Automation & Scale

**Cromwell on Azure**  
workflow & task execution engine



## Clinical Genomics

**Microsoft Genomics service on Azure**  
turnkey secondary analysis pipeline



# Microsoft Container Registry

Microsoft Container Registry (MCR) is the primary Registry for all Microsoft Published docker images that offers a reliable and trustworthy delivery of container images

<https://github.com/microsoft/ContainerRegistry>

"With Azure's global footprint, coupled with Azure CDN, customers can rest assured their image pull experience will be consistent and improve over time. Azure customers, running their workloads in Azure will benefit from in-network performance enhancements as well as tight integration with the Microsoft Container Registry (the source for Microsoft container images)"



# FAQs

## **How does MCR work with Docker Hub?**

MCR is a public registry that houses Microsoft Published images but it does not have its own catalog UI experience. Docker Hub is the official source for our customers to discover official Microsoft-published container images.

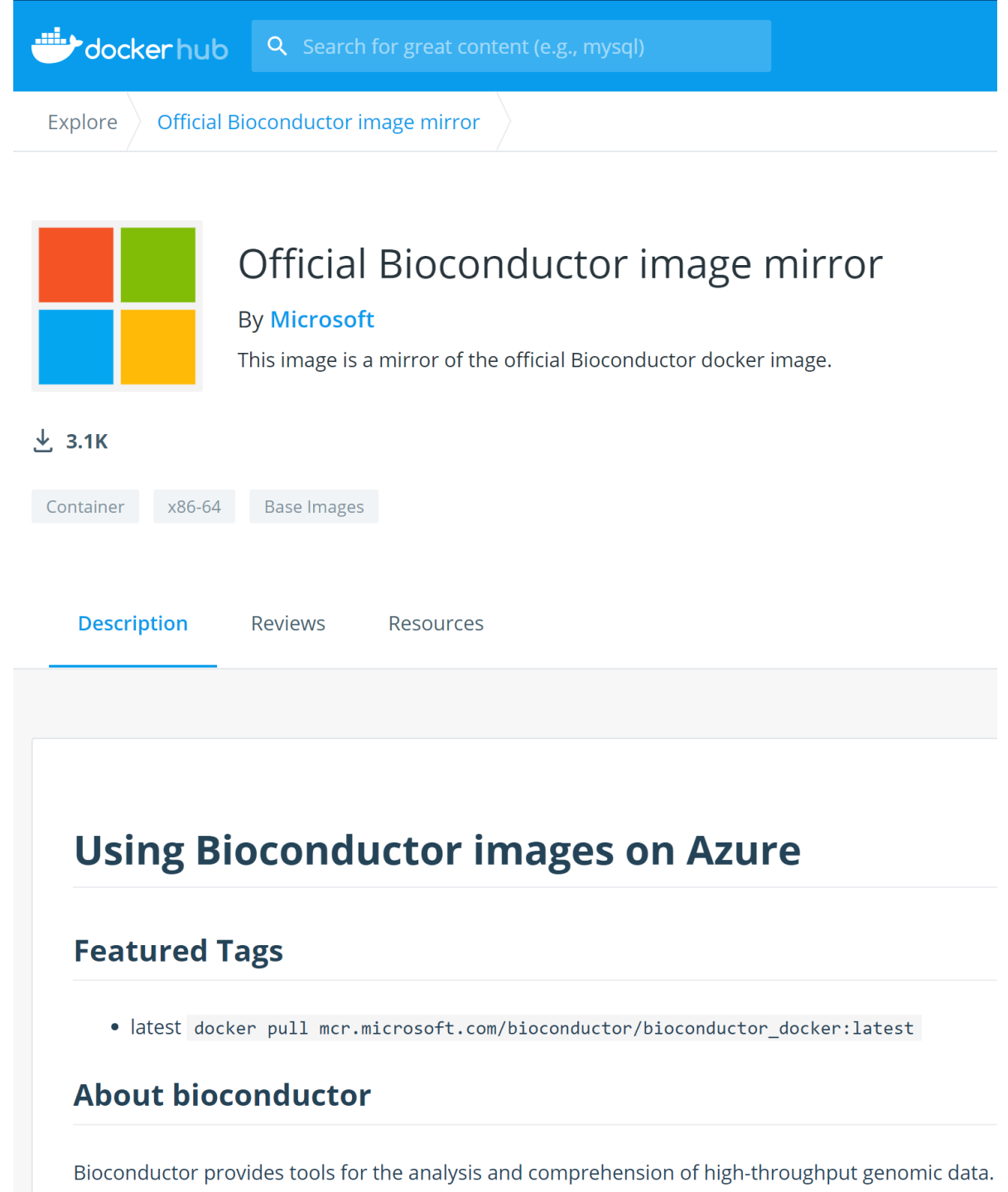
## **What is the difference between MCR and ACR(Azure Container Registry)?**

MCR is a public registry for housing only Microsoft's official Container images. ACR is a private container Registry for housing Our customers container Images.

# Docker Hub UI to explore the MCR image

- Updated within a week of new official Bioconductor images
- "devel" image updated every Monday
- Basic instructions on how to pull images from MCR and when an image was updated on Docker Hub

[https://hub.docker.com/\\_/microsoft-bioconductor-bioconductor-docker](https://hub.docker.com/_/microsoft-bioconductor-bioconductor-docker)



The screenshot shows the Docker Hub interface for the 'Official Bioconductor image mirror' image. The header includes the Docker Hub logo and a search bar. The breadcrumb trail shows 'Explore' and 'Official Bioconductor image mirror'. The image is represented by a 2x2 grid of colored squares (red, green, blue, yellow). The title is 'Official Bioconductor image mirror' by Microsoft. A description states: 'This image is a mirror of the official Bioconductor docker image.' The download count is 3.1K. The image is categorized as 'Container', 'x86-64', and 'Base Images'. The 'Description' tab is selected, showing a section titled 'Using Bioconductor images on Azure'. Below this is a 'Featured Tags' section with a single tag: 'latest docker pull mcr.microsoft.com/bioconductor/bioconductor\_docker:latest'. The 'About bioconductor' section is partially visible at the bottom, stating: 'Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data.'

# How to use instructions on Bioconductor.org

<https://bioconductor.org/help/docker/#msft>

- Mount Azure File Share to persist analysis data between sessions
- How to use with Azure CLI

## Microsoft Azure Container Instances

If you are a Microsoft Azure user, you have an option to run your containers using images hosted on [Microsoft Container Registry](#).

Microsoft Container Registry (MCR) is the primary Registry for all Microsoft Published docker images that offers a reliable and trustworthy delivery of container images with a syndicated catalog

### Using containers hosted on Microsoft Container Registry

You can learn more about the `bioconductor_docker` image hosted on Microsoft Container Registry [here](#).

Pull the `bioconductor_docker` image from Microsoft Container Registry, specifying your `tag` of choice. Check [here](#) for the list of tags under "Full Tag Listing":

```
docker pull mcr.microsoft.com/bioconductor/bioconductor_docker:<tag>
```

To pull the latest image:

```
docker pull mcr.microsoft.com/bioconductor/bioconductor_docker:latest
```

### Example: Run RStudio interactively from your docker container

To run RStudio in a web browser session, run the following and access it from `127.0.0.1:8787`. The default user name is "rstudio" and you can specify your password as the example below (here, it is set to 'bioc'):

```
docker run --name bioconductor_docker_rstudio \  
-v ~/host-site-library:/usr/local/lib/R/host-site-library \  
-e PASSWORD='bioc' \  
-p 8787:8787 \  
mcr.microsoft.com/bioconductor/bioconductor_docker:latest
```

To run RStudio on your terminal:


```
docker run --name bioconductor_docker_rstudio \  
-it \  
-v ~/host-site-library:/usr/local/lib/R/host-site-library \  
-e PASSWORD='bioc' \  
-p 8787:8787 \  
mcr.microsoft.com/bioconductor/bioconductor_docker:latest R
```

# Azure Container Instances

Jass Bagga



erdalcosgun Update README.md

Latest commit 95e6deb 3 days ago [History](#)2 contributors [microsoft/genomicsnotebook: Jupyter Notebooks on Azure for Genomics Data Analysis \(github.com\)](https://github.com/microsoft/genomicsnotebook)

60 lines (40 sloc) | 5.07 KB

Raw

Blame



# Genomics Data Analysis with Jupyter Notebooks on Azure



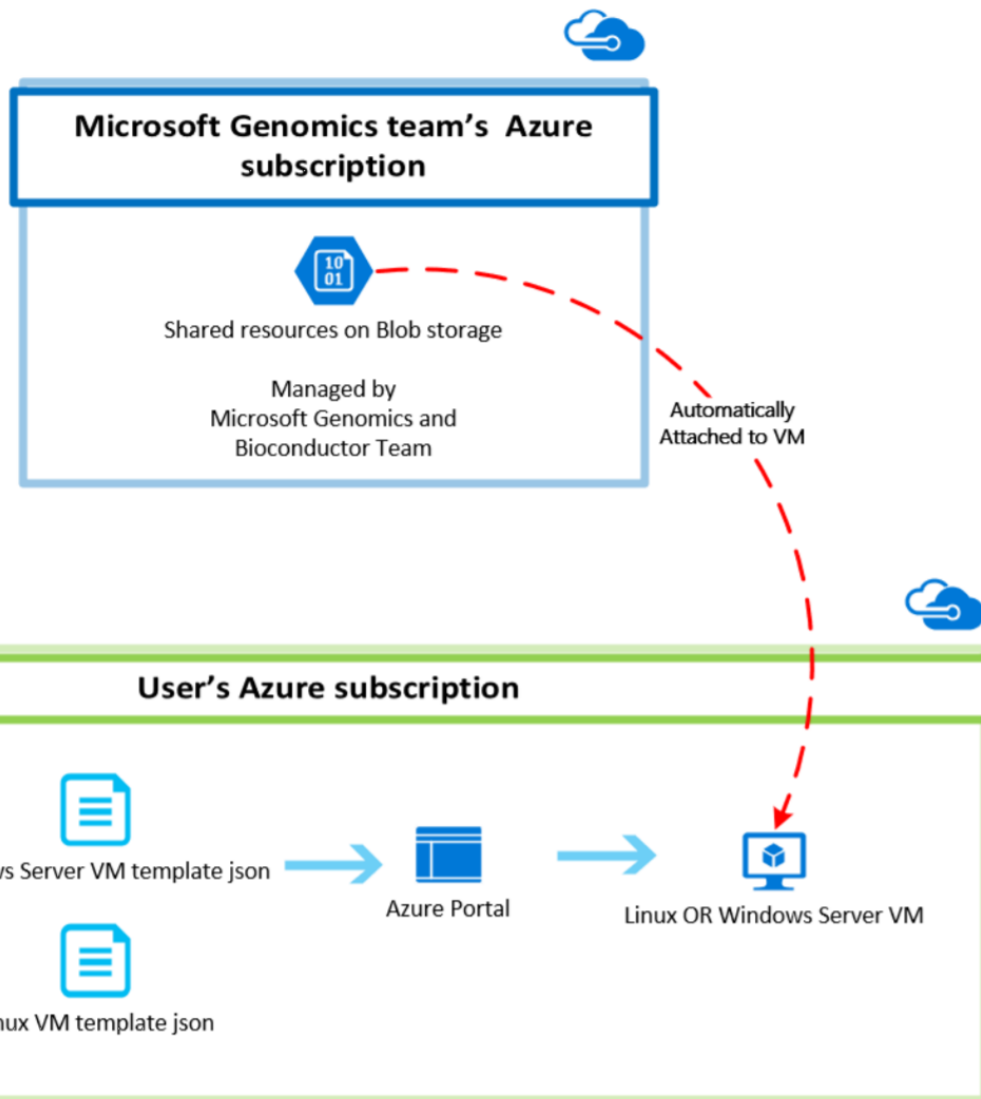
Jupyter notebooks are a great tool for data scientists who are working on genomics data analysis. In this repo, we demonstrate the use of [Azure Notebooks](#) for genomics data analysis via GATK, Picard, Bioconductor and Python libraries.

Here is the list of sample notebooks on this repo:

1. [genomics.ipynb](#) : Analysis from 'uBAM' to 'structured data table' analysis.
2. [genomicsML.ipynb](#) : Train Machine Learning models with Genomics + Clinical Data
3. [genomics-platinum-genomes.ipynb](#) : Accessing Illumina Platinum Genomes data from [Azure Open Datasets](#) and to make initial data analysis.
4. [genomics-reference-genomes.ipynb](#) : Accessing reference genomes from [Azure Open Datasets](#)
5. [genomics-clinvar.ipynb](#) : Accessing ClinVar data from [Azure Open Datasets](#)
6. [genomics-giab.ipynb](#) : Accessing Genome in a Bottle data from [Azure Open Datasets](#)
7. [SnpEff.ipynb](#) : Accessing SnpEff databases from [Azure Open Datasets](#)
8. [Bioconductor.ipynb](#) : Pulling Bioconductor Docker image from [Microsoft Container Registry](#)
9. [simtable.ipynb](#) : Simulate NGS data, use Cromwell on Azure OR Microsoft Genomics service for secondary analysis and convert the gVCF data to a structured data table.



# Azure R-Bioconductor Data Science VM



## High Level Design

# Genomics Public Datasets on Azure

## Genomics Data Lake on Azure Open Dataset

- Collection of commonly used genomics datasets
- Azure Notebooks implementations for data access and analysis
- Easily integration into genomics workflows
- Pay only for Azure services consumed while using the datasets
- 1 petabyte and growing

### Genomics Datasets

- Human Reference Genomes
- Illumina Platinum Genomes
- ClinVar variant annotation database
- Genome in a Bottle
- SnpEff variant annotation-prediction database
- gnomAD

## Azure Open Datasets

Easily access curated datasets and accelerate **machine learning**



Quickly build more accurate models



Promote community collaboration



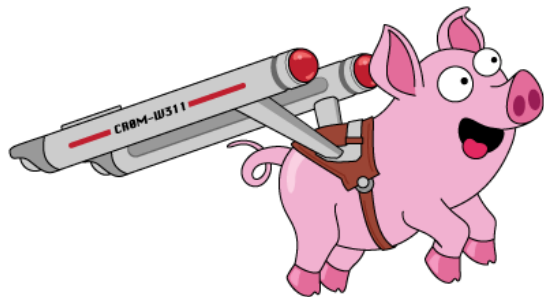
Speed insights with Azure scale



<https://azure.microsoft.com/en-us/services/open-datasets/catalog/genomics-data-lake/>

# Workflow management on Azure

[Cromwell](#) is a workflow management system for scientific workflows, orchestrating the computing tasks needed for genomics analysis. Originally developed by the [Broad Institute](#), Cromwell is also used in the **GATK Best Practices genome analysis pipeline**. Cromwell supports running scripts at various scales, including your local machine, a local computing cluster, and on the cloud.



**Cromwell on Azure** configures all Azure resources needed to run workflows through Cromwell on the Azure cloud, and uses the [GA4GH TES](#) (Task Execution Schemas) backend for orchestrating the tasks that create a workflow. A VM host runs Cromwell server and uses Azure Batch to spin up VMs that run each task in a workflow. Cromwell workflows can be written using either the Workflow Description Language ([WDL](#)) or the Common Workflow Language ([CWL](#)) scripting languages.

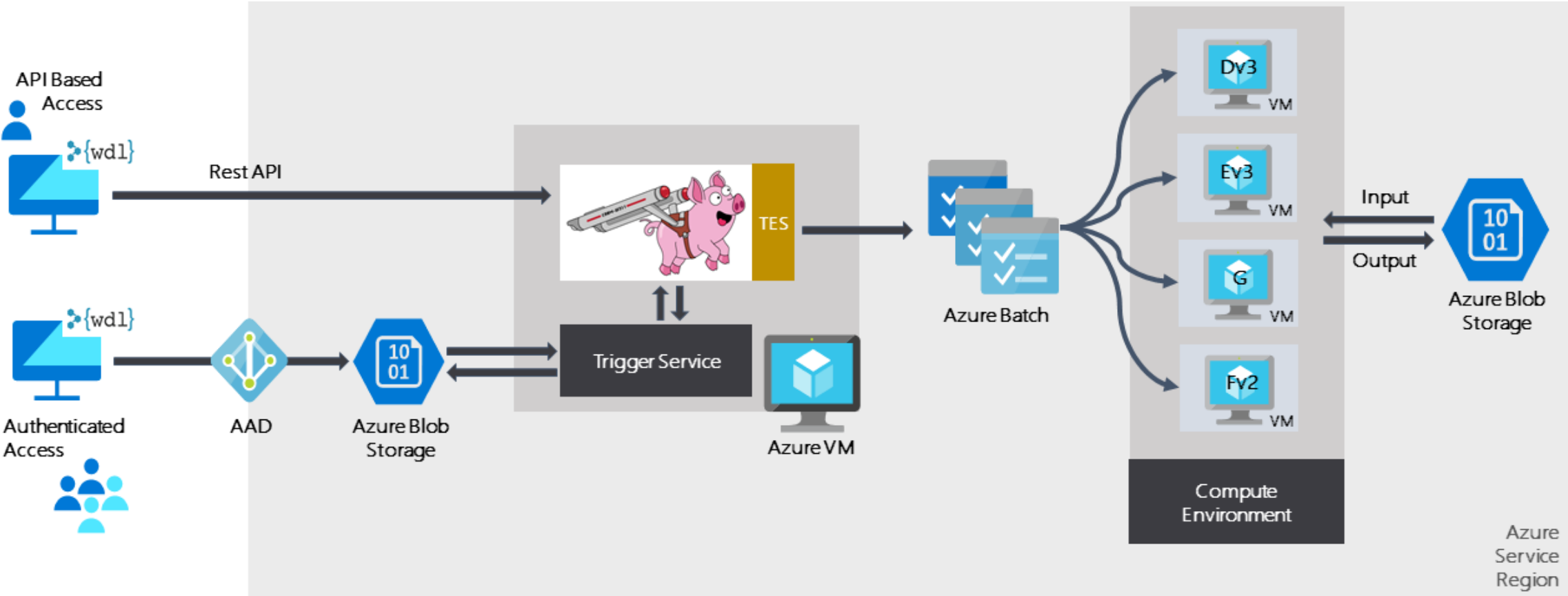


Cromwell



Azure

# Cromwell on Azure overview



<https://github.com/microsoft/CromwellOnAzure>

## [Broad Institute and Verily partner with Microsoft to accelerate the next generation of the Terra platform for health and life science research - Stories](#)

January 11, 2021 | Microsoft News Center



*Multiyear partnership brings together advanced technology, industry expertise and scale to help researchers interpret an unprecedented amount of biomedical data and derive insights to advance the treatment of human diseases*

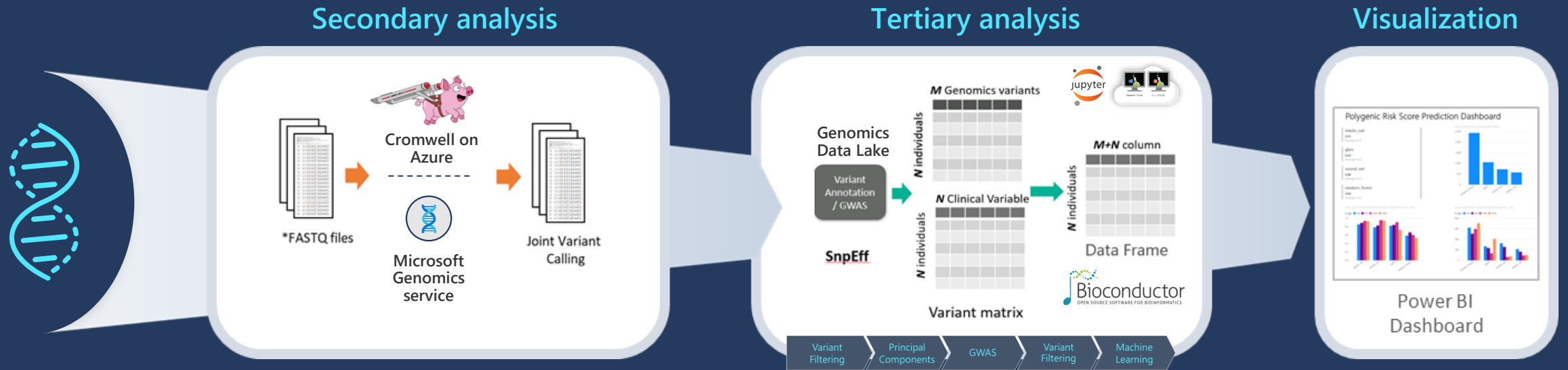
# terra.bio

"Access data, run analysis tools, and collaborate in **Terra: a scalable platform for biomedical research.**"



**CAMBRIDGE, Mass., SOUTH SAN FRANCISCO, Calif., and REDMOND, Wash. — Jan. 11, 2021** — On Monday, Broad Institute of MIT and Harvard, Verily, an Alphabet company, and Microsoft Corp. announced a strategic partnership to accelerate new innovations in biomedicine through the Terra platform. Terra, originally developed by Verily and the Broad Institute, is a secure, scalable, open-source platform for biomedical researchers to access data, run analysis tools and collaborate. Terra is actively used by thousands of researchers every month to analyze data from millions of participants in important scientific research projects.

# End-2-end genomics analysis workflow on Azure



AZURE VM



AZURE BATCH



AZURE CONTAINER SERVICE



docker



AZURE BLOB STORAGE



DATA LAKE STORE



DATA FACTORY



SQL DB



SQL DW



DATA SCIENCE VM



AZURE ML



POWER BI



jupyter



R

# Q&A

Thank you!

[Microsoft Genomics](#)

[Dissertation Grant - Microsoft Research](#)

[Academic Programs - Microsoft Research](#)

[Create your Azure free account today | Microsoft Azure](#)

For questions: [genomics@microsoft.com](mailto:genomics@microsoft.com)