



# Modeling the computing requirements and costs for genomics analysis in the cloud

---

Michael Schatz

 @mike\_schatz

July 26, 2021  
ISMB NIH/ODSS Workshop

# What is the AnVIL?

## Scalable and interoperable computing resource for the genomics scientific community

- Cloud-based infrastructure
  - Highly elastic; shared analysis and computing environment
- Data access and security
  - Genomic datasets, phenotypes and metadata
  - Large datasets generated by NHGRI programs, as well as other initiatives / agencies
  - dbGaP Authenticated sharing of primary and derived datasets
- Collaborative computing environment for datasets and analysis workflows
  - Storage, scalable analytics, data visualization
  - Security, training & outreach, with new models of data access
  - ...for both users with limited computational expertise and sophisticated data scientist users

[anvilproject.org](https://anvilproject.org)

About Data Tools Training News Events FAQ Contact NCPI

## Migrate Your Genomic Analysis Workflows to the Cloud

Analyze large, open & controlled-access genomic datasets with familiar tools and reproducible workflows in a secure cloud-based computing environment.

- Launch Terra, AnVIL's cloud computing environment.
- Create a virtual cohort in AnVIL's Gen3 Data Explorer.
- Discover and launch repeatable workflows with Dockstore.
- Explore emerging support for cross-platform data sharing and analysis via the NIH Cloud Platform Interoperability effort.


5 CONSORTIA 100 COHORTS 75K SUBJECTS 90K SAMPLES 1.1PB SIZE

[Explore AnVIL's datasets and learn how to request access.](#)

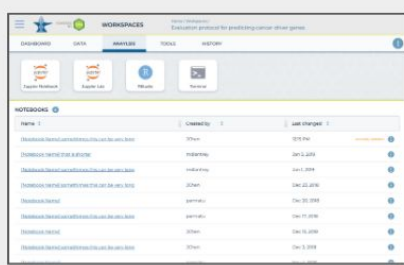
PLATFORMS Dockstore

Dockstore is an open platform used by the GA4GH for sharing Docker-based tools described with the Common Workflow Language (CWL), the Workflow...  
[Learn More >>](#)

<https://anvilproject.org>



**GEN3** Data Commons  
Data models,  
indexing, querying




**Terra** Workspaces and  
batch workflows



**Dockstore**  
Create, Share, Use

Sharing containerized tools  
and workflows



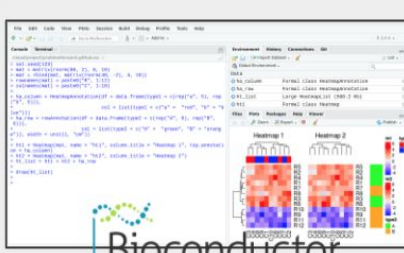
**Jupyter**

Live code, equations,  
visualizations and narratives



**Galaxy**

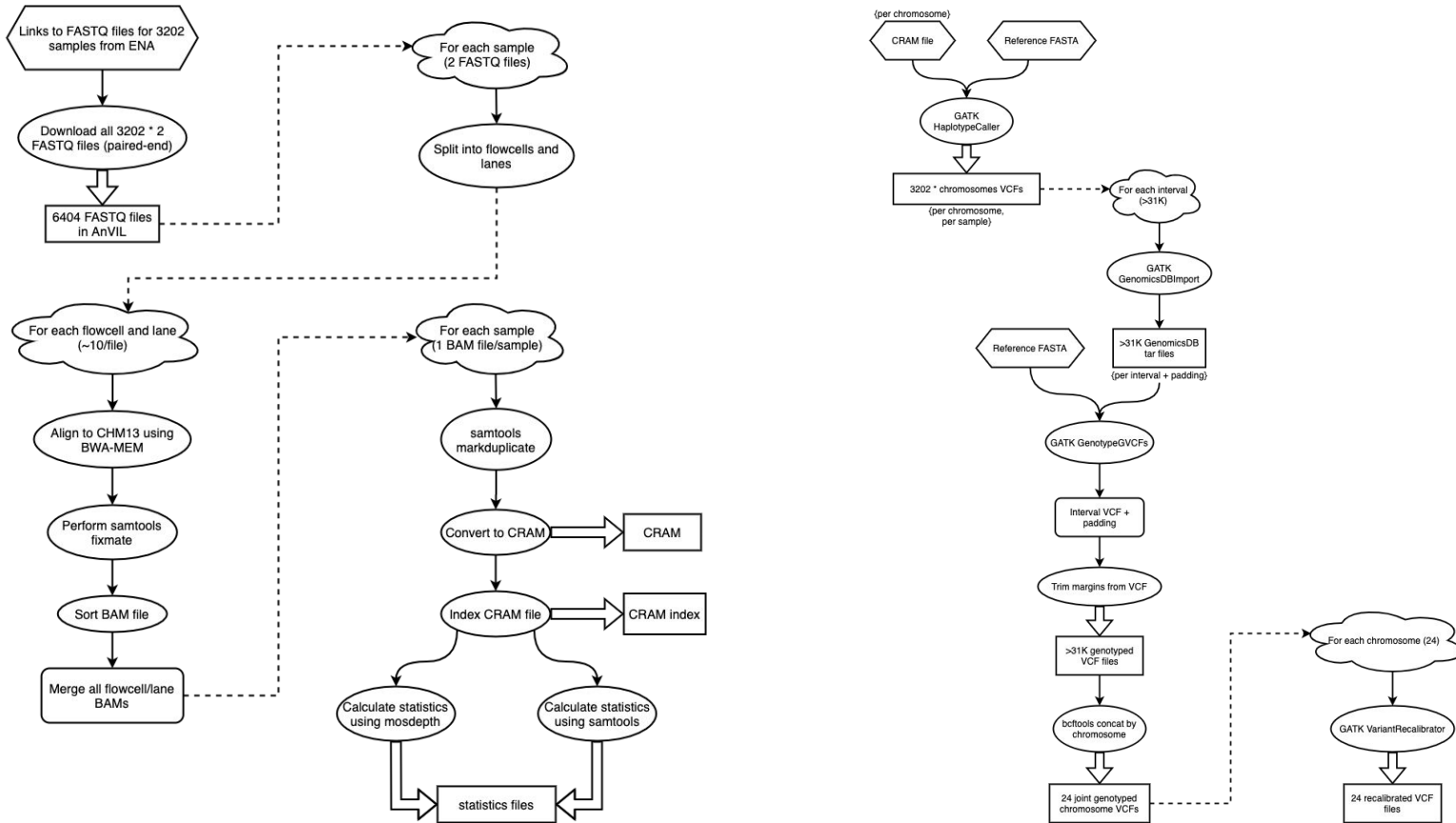
Accessible, reproducible, and  
transparent research



**Bioconductor**  
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Analysis and comprehension  
of genomic data in R

# T2T Analysis in AnVIL



*A complete reference genome improves analysis of human genetic variation*

Aganezov, S\*, Yan, SM\*, Soto, DC\*, Kirsche, M\*, Zarate, S\*, et al. (2021) *bioRxiv* doi: 10.1101/2021.07.12.452063

# T2T Analysis on Google Cloud Platform

Preview

1 hour

4 hours

1 day



● instance/cpu/reserved\_cores: 11,552.00

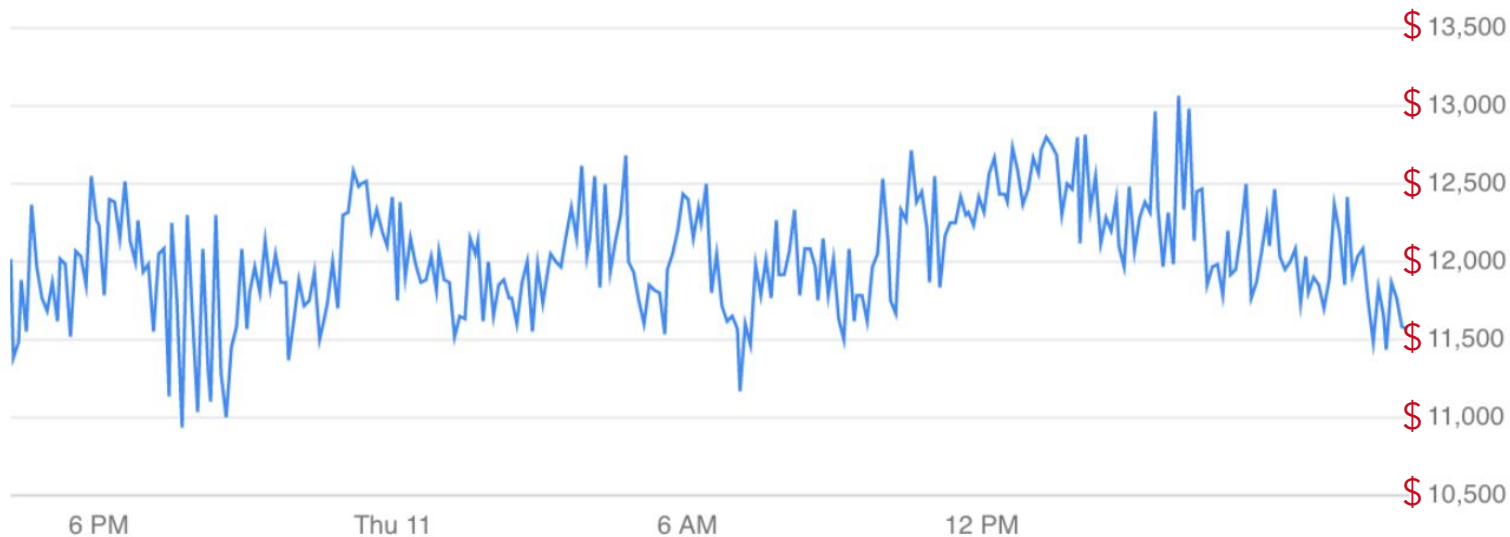
# T2T Analysis on Google Cloud Platform

Preview

1 hour

4 hours

1 day



● dollars/hour

# Cloud Costs are complicated

## E2 standard machine types

The following table shows the calculated cost for standard predefined machine types in the E2 machine family. The vCPUs and memory from each of these machine types are billed by their individual [predefined vCPU and memory prices](#), but these tables provide the cost that you can expect using a specific machine type.

Standard machine types have 4 GB of memory per vCPU.

Iowa (us-central1)				
Monthly <input type="radio"/> Hourly <input checked="" type="radio"/>				
Machine type	Virtual CPUs	Memory	Price (USD)	Preemptible price (USD)
e2-standard-2	2	8GB	\$0.067006	\$0.020102
e2-standard-4	4	16GB	\$0.134012	\$0.040204
e2-standard-8	8	32GB	\$0.268024	\$0.080408
e2-standard-16	16	64GB	\$0.536048	\$0.160816
e2-standard-32	32	128GB	\$1.072096	\$0.321632
Custom machine type	If your ideal machine shape is in between two predefined types, using a custom E2 machine type could save you as much as 40%. For more information, see <a href="#">E2 custom vCPUs and memory</a> .			

## N2 standard machine types

The following table shows the calculated costs for standard predefined machine types in the N2 machine family. The vCPUs and memory from each of these machine types are billed by their individual [predefined vCPU and memory prices](#), but these tables provide the cost that you can expect using a specific machine type.

Standard machine types have 4 GB of memory per vCPU.

Iowa (us-central1)				
Monthly <input type="radio"/> Hourly <input checked="" type="radio"/>				
Machine type	Virtual CPUs	Memory	Price (USD)	Preemptible price (USD)
n2-standard-2	2	8GB	\$0.097118	\$0.02354
n2-standard-4	4	16GB	\$0.194236	\$0.04708
n2-standard-8	8	32GB	\$0.388472	\$0.09416
n2-standard-16	16	64GB	\$0.776944	\$0.18832
n2-standard-32	32	128GB	\$1.553888	\$0.37664
n2-standard-48	48	192GB	\$2.330832	\$0.56496
n2-standard-64	64	256GB	\$3.107776	\$0.75328
n2-standard-80	80	320GB	\$3.88472	\$0.9416
Custom machine type	If your ideal machine shape is in between two predefined types, using a custom machine type could save you as much as 40%. For more information, see <a href="#">Custom vCPU and memory</a> .			

## E2 high-memory machine types

The following table shows the calculated cost for the E2 high-memory predefined machine types. The vCPUs and memory from each of these machine types are billed by their individual [predefined vCPU and memory prices](#), but these tables provide the cost that you can expect using a specific machine type.

High-memory machine types have 8 GB of memory per vCPU. High-memory instances are ideal for tasks that require more memory relative to virtual CPUs.

Iowa (us-central1)				
Monthly <input type="radio"/> Hourly <input checked="" type="radio"/>				
Machine type	Virtual CPUs	Memory	Price (USD)	Preemptible price (USD)
e2-highmem-2	2	16GB	\$0.09039	\$0.027118
e2-highmem-4	4	32GB	\$0.18078	\$0.054236
e2-highmem-8	8	64GB	\$0.36156	\$0.108472
e2-highmem-16	16	128GB	\$0.72312	\$0.216944
Custom machine type	If your ideal machine shape is in between two predefined types, using a custom E2 machine type could save you as much as 40%. For more information, see <a href="#">E2 custom vCPUs and memory</a> .			

## N2 high-memory machine types

The following table shows the calculated cost for the N2 high-memory predefined machine types. The vCPUs and memory from each of these machine types are billed by their individual [predefined vCPU and memory prices](#), but these tables provide the cost that you can expect using a specific machine type.

High-memory machine types have 8 GB of memory per vCPU. High-memory instances are ideal for tasks that require more memory relative to virtual CPUs.

Iowa (us-central1)				
Monthly <input type="radio"/> Hourly <input checked="" type="radio"/>				
Machine type	Virtual CPUs	Memory	Price (USD)	Preemptible price (USD)
n2-highmem-2	2	16GB	\$0.131014	\$0.03178
n2-highmem-4	4	32GB	\$0.262028	\$0.06356
n2-highmem-8	8	64GB	\$0.524056	\$0.12712
n2-highmem-16	16	128GB	\$1.048112	\$0.25424
n2-highmem-32	32	256GB	\$2.096224	\$0.50848
n2-highmem-48	48	384GB	\$3.144336	\$0.76272
n2-highmem-64	64	512GB	\$4.192448	\$1.01696
n2-highmem-80	80	640GB	\$5.24056	\$1.2712
Custom machine type	If your ideal machine shape is in between two predefined types, using a custom machine type could save you as much as 40%. For more information, see <a href="#">Custom vCPU and memory</a> .			

## E2 high-CPU machine types

The following table shows the calculated cost for E2 high-CPU predefined machine types. The vCPUs and memory from each of these machine types are billed by their individual [predefined vCPU and memory prices](#), but these tables provide the cost that you can expect using a specific machine type.

High-CPU machine types have one vCPU for every 1 GB of memory. High-CPU machine types are ideal for tasks that require moderate memory configurations for the needed vCPU count.

Iowa (us-central1)				
Monthly <input type="radio"/> Hourly <input checked="" type="radio"/>				
Machine type	Virtual CPUs	Memory	Price (USD)	Preemptible price (USD)
e2-highcpu-2	2	2GB	\$0.049468	\$0.01484
e2-highcpu-4	4	4GB	\$0.098936	\$0.02968
e2-highcpu-8	8	8GB	\$0.197872	\$0.05936
e2-highcpu-16	16	16GB	\$0.395744	\$0.11872
e2-highcpu-32	32	32GB	\$0.791488	\$0.23744
Custom machine type	If your ideal machine shape is in between two predefined types, using a custom E2 machine type could save you as much as 40%. For more information, see <a href="#">E2 custom vCPUs and memory</a> .			

## N2 high-CPU machine types

The following table shows the calculated cost for N2 high-CPU predefined machine types. The vCPUs and memory from each of these machine types are billed by their individual [predefined vCPU and memory prices](#), but these tables provide the cost that you can expect using a specific machine type.

High-CPU machine types have one vCPU for every 1 GB of memory. High-CPU machine types are ideal for tasks that require moderate memory configurations for the needed vCPU count.

Iowa (us-central1)				
Monthly <input type="radio"/> Hourly <input checked="" type="radio"/>				
Machine type	Virtual CPUs	Memory	Price (USD)	Preemptible price (USD)
n2-highcpu-2	2	2GB	\$0.071696	\$0.01736
n2-highcpu-4	4	4GB	\$0.143392	\$0.03472
n2-highcpu-8	8	8GB	\$0.286784	\$0.06944
n2-highcpu-16	16	16GB	\$0.573568	\$0.13888
n2-highcpu-32	32	32GB	\$1.147136	\$0.27776
n2-highcpu-48	48	48GB	\$1.720704	\$0.41664
n2-highcpu-64	64	64GB	\$2.294272	\$0.55552
n2-highcpu-80	80	80GB	\$2.86784	\$0.6944
Custom machine type	If your ideal machine shape is in between two predefined types, using a custom machine type could save you as much as 40%. For more information, see <a href="#">Custom vCPUs and memory</a> .			

# Cloud Costs are complicated

The image displays the Google Cloud Pricing Calculator interface, which is used for estimating cloud costs. The interface is divided into several sections:

- Navigation and Search:** The top navigation bar includes links for "Why Google", "Solutions", "Products", "Pricing", and "Getting Started". A search bar is located on the right.
- Product Categories:** A row of icons represents different Google Cloud services: Compute Engine, GKE Standard, GKE Autopilot, Cloud Run, VMware Engine, App Engine, Cloud Storage, Networking Egress, and Cloud Load Balancing.
- Search and Filter:** A search bar prompts the user to "Search for a product you are interested in." Below it, a list of filters for "Instances" includes "Number of instances", "Operating System / Software", "Machine Class", "Machine Family", "Series", "Machine type", "Datacenter location", and "Instances using ephemeral/static public IP".
- Configuration Form:** The main form is titled "Sole-tenant nodes" and includes fields for:
  - Number of nodes:** A dropdown menu.
  - Node type:** A dropdown menu showing "n1-node-96-624 (vCPUs: 96, RAM: 624 GB)".
  - Local SSD:** A dropdown menu showing "0".
  - Datacenter location:** A dropdown menu showing "Iowa (us-central1)".
  - Committed usage:** A dropdown menu showing "None".
  - Average hours per day each server is running:** A dropdown menu showing "24".
  - hours:** A dropdown menu showing "hours".
  - per day:** A dropdown menu showing "per day".
  - Average days per week each server is running:** A dropdown menu showing "7".
  - Persistent Disk:** A section with a "Location" dropdown showing "Iowa (us-central1)" and a list of disk options: "Zonal standard PD", "Regional standard PD", "Zonal balanced PD", "Regional balanced PD", "Zonal SSD PD", and "Regional SSD PD".
- Summary and Chat:** On the right side, there is a summary section with fields for "Extreme PD", "Extreme PD IOPS", "Snapshot Storage", "Multi-regional snapshot Storage", and "Cloud TPU". It includes an "ADD TO ESTIMATE" button and a "Ready to get started? Chat with us" button.
- FAQ:** A section titled "FAQ" provides information about price estimates, estimated fees, usage timeframes, and pricing data.

<https://cloud.google.com/products/calculator>



# Cloud Costs can cause “Range Anxiety” (h/t Jeff Leek)

---



## Purchased car

- You buy the car
- You fill up at a station
- You pay less per mile



## ZipCar

- You don't buy the car
- You pay by the mile
- You may pay more per mile



What can we do?



# NIH/ODSS STRIDES Initiative

data-science.nih.gov

U.S. Department of Health & Human Services | National Institutes of Health | Division of Program Coordination, Planning, and Strategic Initiatives (DPCPSI)

NIH National Institutes of Health  
Office of Data Science Strategy

Home Strategic Plan Resources Research Funding News & Events About

COVID-19

- Get the latest public health information from CDC
- Get the latest research information from NIH | Español
- NIH staff guidance on coronavirus (NIH Only)
- NIH and other federal agencies have made COVID-19 data available through several Open-Access Data and Computational Resources

STRIDES Initiative

Office of Data Science Strategy » Resources » STRIDES Initiative

About Cloud Preparing to Use the Cloud Partner Offerings Success Stories

### About the STRIDES Initiative

Data generated via biomedical research continues to outpace the ability to process, store, and analyze in many local environments.

The **NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative** allows NIH to explore the use of cloud environments to streamline NIH data use by partnering with commercial providers. NIH's STRIDES Initiative provides cost-effective access to industry-leading partners to help advance biomedical research. These partnerships enable access to rich datasets and advanced computational infrastructure, tools, and services.

The STRIDES Initiative is one of many NIH-wide efforts to implement the **NIH Strategic Plan for Data Science**, which provides a roadmap for modernizing the NIH-funded biomedical data science ecosystem.

By leveraging the STRIDES Initiative, NIH and NIH-funded institutions can begin to create a robust, interconnected ecosystem that breaks down silos related to generating, analyzing, and sharing research data. NIH-funded researchers with an active NIH award may take advantage of the STRIDES Initiative for their NIH-funded research projects. Eligible investigators include awardees of NIH contracts, other transaction agreements, grants, cooperative agreements, and other agreements.

Benefits of using the STRIDES Initiative as a vehicle to access STRIDES Initiative partners include:

- Discounts on STRIDES Initiative partner services**—Favorable pricing on computing, storage, and related cloud services for NIH Institutes, Centers, and Offices (ICOs) and NIH-funded institutions.
- Professional services**—Access to professional service consultations and technical support from the STRIDES Initiative partners.
- Training**—Access to training for researchers, data owners, and others to help ensure optimal use of available tools and technologies.
- Potential collaborative engagements**—Opportunities to explore methods and approaches that may advance NIH's biomedical research objectives (with scope and milestones of engagements agreed upon separately).

At this time, the STRIDES Initiative supports programs/projects who want to prepare, migrate, upload, and compute on data in the cloud. In the future, the ability to access data across NIH and NIH-funded institutions from various research domain repositories will become available.

Is cloud right for me?

- What is Cloud?
- How Can Cloud Services Be Used for Research?

To learn more details about the STRIDES Initiative, enroll in training, or opt in to receive newsletters, visit the [STRIDES Initiative website](#). Extramural institutions may find these [Frequently Asked Questions](#) useful. For further questions, the [STRIDES Initiative team](#) is available to help facilitate connection to cost-effective, cloud-based computing resources.

The first STRIDES Initiative partnership was established with **Google Cloud** in July 2018; a second partnership was established with **Amazon Web Services (AWS)** in September 2018.

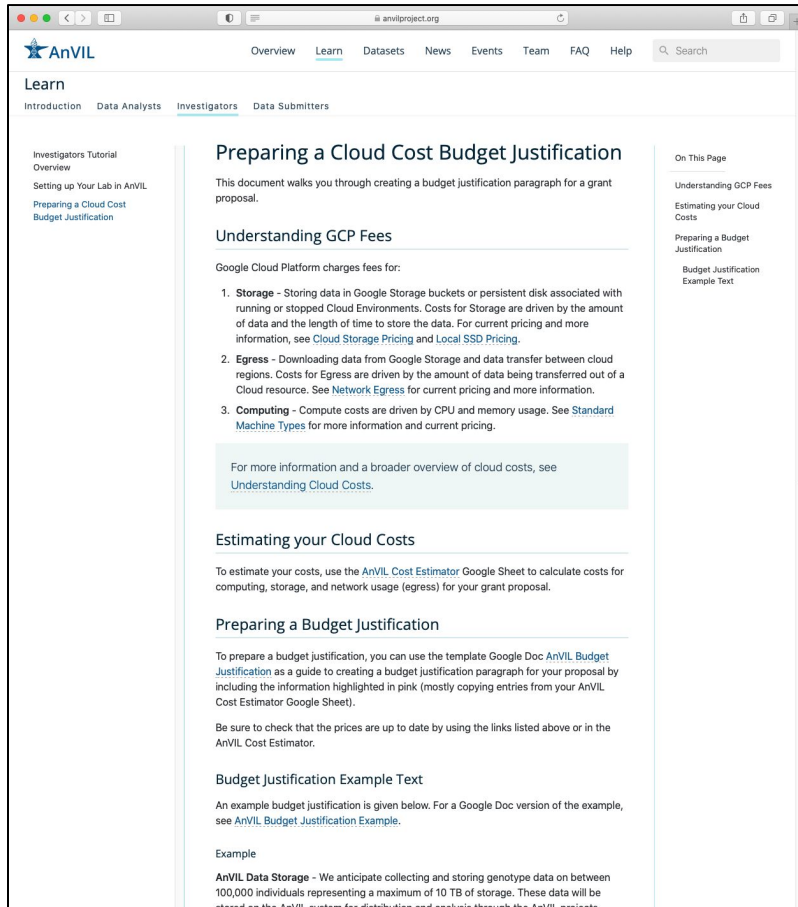
Read more

Get started with the STRIDES Initiative

## STRIDES Benefits

- **Discounts** (typically 10%-25%) on computing, storage, and related cloud services for NIH Institutes, Centers, and Offices (ICOs) and NIH-funded institutions & investigators.
- **Professional services** — Access to professional service consultations and technical support from the STRIDES Initiative partners.
- **Training** — Access to training for researchers, data owners, and others to help ensure optimal use of available tools and technologies.
- **Potential collaborative engagements** — Opportunities to explore methods and approaches that may advance NIH's biomedical research objectives

# AnVIL Cloud Cost Budget Templates



The screenshot shows a web browser window displaying the AnVIL website. The page is titled "Preparing a Cloud Cost Budget Justification" and is part of the "Learn" section under "Investigators". The page content includes a navigation menu, a sidebar with a table of contents, and the main body text. The main body text is divided into sections: "Understanding GCP Fees", "Estimating your Cloud Costs", "Preparing a Budget Justification", and "Budget Justification Example Text".

**Learn**

Introduction Data Analysts **Investigators** Data Submitters

Investigators Tutorial Overview  
Setting up Your Lab in AnVIL  
Preparing a Cloud Cost Budget Justification

## Preparing a Cloud Cost Budget Justification

This document walks you through creating a budget justification paragraph for a grant proposal.

### Understanding GCP Fees

Google Cloud Platform charges fees for:

- Storage** - Storing data in Google Storage buckets or persistent disk associated with running or stopped Cloud Environments. Costs for Storage are driven by the amount of data and the length of time to store the data. For current pricing and more information, see [Cloud Storage Pricing](#) and [Local SSD Pricing](#).
- Egress** - Downloading data from Google Storage and data transfer between cloud regions. Costs for Egress are driven by the amount of data being transferred out of a Cloud resource. See [Network Egress](#) for current pricing and more information.
- Computing** - Compute costs are driven by CPU and memory usage. See [Standard Machine Types](#) for more information and current pricing.

For more information and a broader overview of cloud costs, see [Understanding Cloud Costs](#).

### Estimating your Cloud Costs

To estimate your costs, use the [AnVIL Cost Estimator](#) Google Sheet to calculate costs for computing, storage, and network usage (egress) for your grant proposal.

### Preparing a Budget Justification

To prepare a budget justification, you can use the template Google Doc [AnVIL Budget Justification](#) as a guide to creating a budget justification paragraph for your proposal by including the information highlighted in pink (mostly copying entries from your AnVIL Cost Estimator Google Sheet).

Be sure to check that the prices are up to date by using the links listed above or in the AnVIL Cost Estimator.

### Budget Justification Example Text

An example budget justification is given below. For a Google Doc version of the example, see [AnVIL Budget Justification Example](#).

Example

**AnVIL Data Storage** - We anticipate collecting and storing genotype data on between 100,000 individuals representing a maximum of 10 TB of storage. These data will be stored on the AnVIL system for distribution and analysis through the AnVIL projects.

**On This Page**

- [Understanding GCP Fees](#)
- [Estimating your Cloud Costs](#)
- [Preparing a Budget Justification](#)
- [Budget Justification Example Text](#)

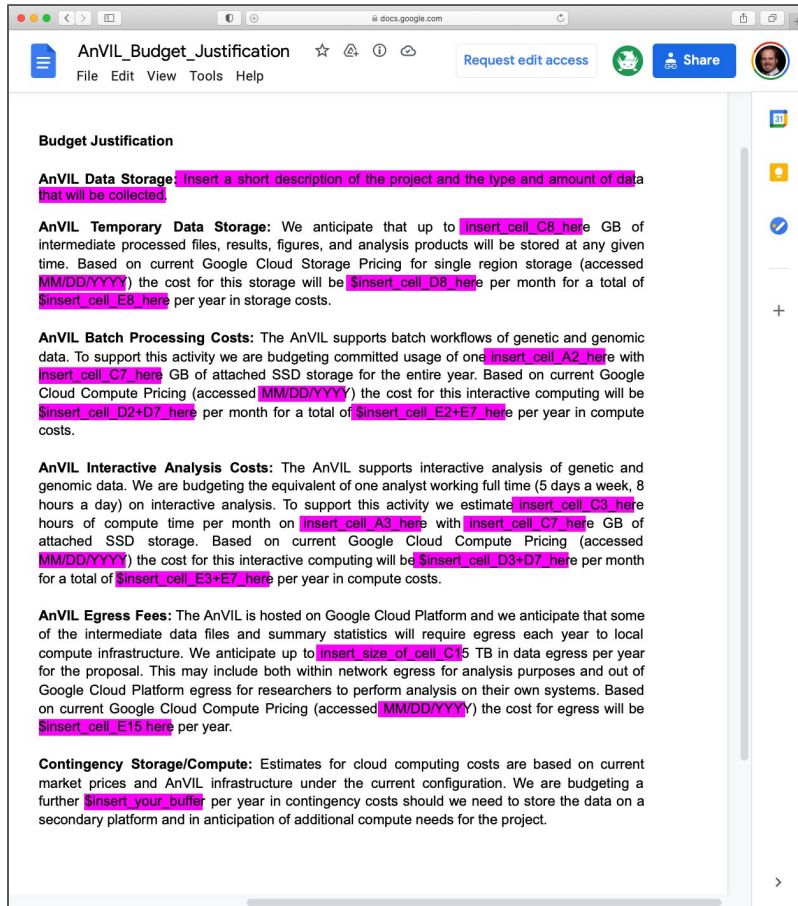
# AnVIL Cloud Cost Budget Templates

The screenshot shows the AnVIL website's 'Learn' section. The main heading is 'Preparing a Cloud Cost Budget Justification'. Below it, a paragraph states: 'This document walks you through creating a budget justification paragraph for a grant proposal.' A sub-section titled 'Understanding GCP Fees' follows, with the text: 'Google Cloud Platform charges fees for:'. A list of three items is provided: 1. **Storage** - Storing data in Google Storage buckets or persistent disk associated with running or stopped Cloud Environments. Costs for Storage are driven by the amount of data and the length of time to store the data. For current pricing and more information, see [Cloud Storage Pricing](#) and [Local SSD Pricing](#). 2. **Egress** - Downloading data from Google Storage and data transfer between cloud regions. Costs for Egress are driven by the amount of data being transferred out of a Cloud resource. See [Network Egress](#) for current pricing and more information. 3. **Computing** - Compute costs are driven by CPU and memory usage. See [Standard Machine Types](#) for more information and current pricing. A light blue box contains the text: 'For more information and a broader overview of cloud costs, see [Understanding Cloud Costs](#).' Below this is the 'Estimating your Cloud Costs' section, which says: 'To estimate your costs, use the [AnVIL Cost Estimator](#) Google Sheet to calculate costs for computing, storage, and network usage (egress) for your grant proposal.' The 'Preparing a Budget Justification' section states: 'To prepare a budget justification, you can use the template Google Doc [AnVIL Budget Justification](#) as a guide to creating a budget justification paragraph for your proposal by including the information highlighted in pink (mostly copying entries from your [AnVIL Cost Estimator](#) Google Sheet).' A note says: 'Be sure to check that the prices are up to date by using the links listed above or in the [AnVIL Cost Estimator](#).' The 'Budget Justification Example Text' section says: 'An example budget justification is given below. For a Google Doc version of the example, see [AnVIL Budget Justification Example](#).' An 'Example' section follows with the text: '**AnVIL Data Storage** - We anticipate collecting and storing genotype data on between 100,000 individuals representing a maximum of 10 TB of storage. These data will be stored on the [AnVIL](#) system for distribution and analysis through the [AnVIL](#) projects.'

The screenshot shows the 'AnVIL\_Cost\_Estimator' Google Sheet. The table displays costs for computing, storage, and network usage. The columns are: Costs/Hour, Number of hours, Costs/Month, and Costs/Year. The rows are categorized by resource type.

	A	B	C	D	E	F	G
1	<b>Costs for Computing</b>	<b>Costs/Hour</b>	<b>Number of hours</b>	<b>Costs/Month</b>	<b>Costs/Year</b>		
2	n1-standard-4 instance consisting of 4 vCPUs and 15 GB of RAM			\$87.09	\$1,165.08	(monthly rates selected)	
3	n1-standard-8 instance consisting of 8 vCPUs and 30 GB of RAM	\$0.379998	174	\$66.12	\$793.44	(hourly rates selected)	
4							
5							
6	<b>Costs for Storage</b>	<b>Costs/Month (1 GB)</b>	<b>Number of GB</b>	<b>Costs/Month</b>	<b>Costs/Year</b>		
7	Local SSD provisioned space	\$0.080	375	\$30.00	\$360.00		
8	Standard Storage, single region storage: lowa (us-central1)	\$0.02	4096	\$81.92	\$983.04		
9							
10							
11	<b>Costs for Network usage (egress)</b>	<b>Cost/GB</b>	<b>Number of GB</b>	<b>Costs/Month</b>	<b>Costs/Year</b>		
12	0-1 TB tier	\$0.12	1024	\$122.88	\$1,474.56		
13	1-10 TB tier	\$0.11	1024	\$112.64	\$1,351.68		
14	10+ TB tier	\$0.08	0	\$0.00	\$0.00		
15	Total @ 2 TB egress to Worldwide Destinations				\$2,826.24		
16							
17							
18	<b>Additional Information</b>						
19	Please label the numbers highlighted in pink.						
20	Pricing based on rates on 12/01/2020, please check for up-to-dateness by using the links listed below.						
21	Storage and network usage are calculated in binary gigabytes (GB): 1 TB is 1024 GBs.						
22							
23	Costs for Computing is driven by CPU and memory requirements.						
24	<a href="https://cloud.google.com/compute/all-pricing#n1_standard_machine_types">https://cloud.google.com/compute/all-pricing#n1_standard_machine_types</a>						
25							
26	Costs for Storage is driven by the amount of data and the length of time to store the data.						
27	<a href="https://cloud.google.com/compute/all-pricing#localssdpricing">https://cloud.google.com/compute/all-pricing#localssdpricing</a>						
28	<a href="https://cloud.google.com/storage/pricing#storage-pricing">https://cloud.google.com/storage/pricing#storage-pricing</a>						
29							
30	Costs for Egress is driven by the amount of data being transferred out of a Cloud resource.						
31	<a href="https://cloud.google.com/storage/pricing#network-egress">https://cloud.google.com/storage/pricing#network-egress</a>						
32							
33							
34							
35							
36							
37							
38							
39							
40							
41							
42							
43							
44							
45							
46							
47							
48							
49							
50							
51							
52							
53							
54							
55							
56							

# AnVIL Cloud Cost Budget Templates



The screenshot shows a Google Docs interface for a document titled "AnVIL\_Budget\_Justification". The document content is as follows:

**Budget Justification**

**AnVIL Data Storage:** Insert a short description of the project and the type and amount of data that will be collected.

**AnVIL Temporary Data Storage:** We anticipate that up to insert\_cell\_C8\_here GB of intermediate processed files, results, figures, and analysis products will be stored at any given time. Based on current Google Cloud Storage Pricing for single region storage (accessed MM/DD/YYYY) the cost for this storage will be insert\_cell\_D8\_here per month for a total of insert\_cell\_E8\_here per year in storage costs.

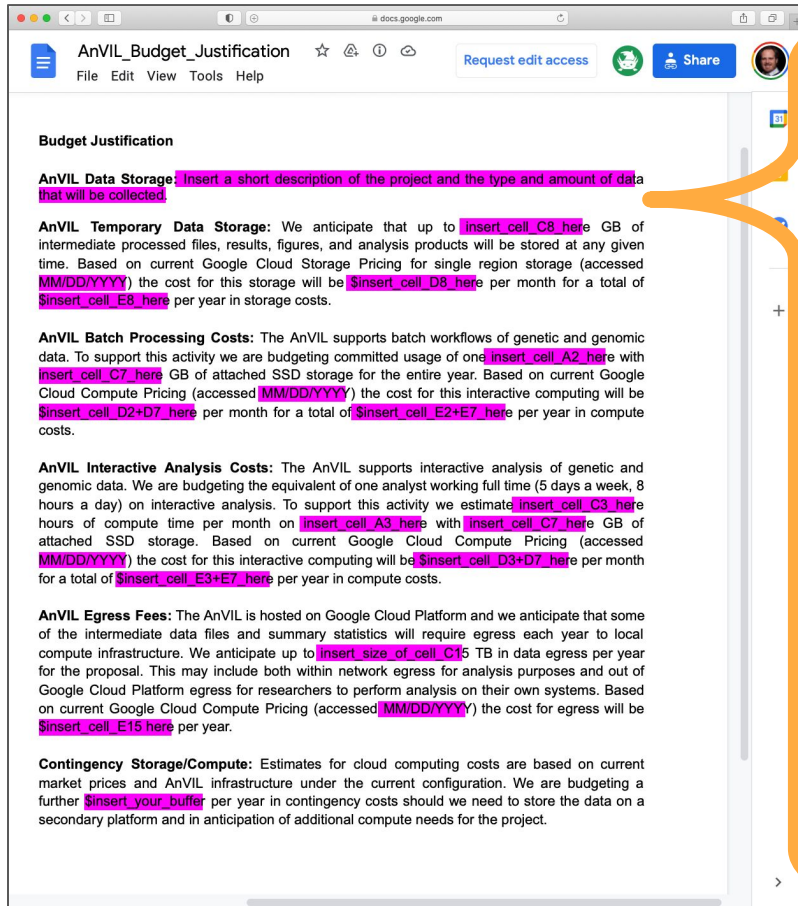
**AnVIL Batch Processing Costs:** The AnVIL supports batch workflows of genetic and genomic data. To support this activity we are budgeting committed usage of one insert\_cell\_A2\_here with insert\_cell\_C7\_here GB of attached SSD storage for the entire year. Based on current Google Cloud Compute Pricing (accessed MM/DD/YYYY) the cost for this interactive computing will be insert\_cell\_D2+D7\_here per month for a total of insert\_cell\_E2+E7\_here per year in compute costs.

**AnVIL Interactive Analysis Costs:** The AnVIL supports interactive analysis of genetic and genomic data. We are budgeting the equivalent of one analyst working full time (5 days a week, 8 hours a day) on interactive analysis. To support this activity we estimate insert\_cell\_C3\_here hours of compute time per month on insert\_cell\_A3\_here with insert\_cell\_C7\_here GB of attached SSD storage. Based on current Google Cloud Compute Pricing (accessed MM/DD/YYYY) the cost for this interactive computing will be insert\_cell\_D3+D7\_here per month for a total of insert\_cell\_E3+E7\_here per year in compute costs.

**AnVIL Egress Fees:** The AnVIL is hosted on Google Cloud Platform and we anticipate that some of the intermediate data files and summary statistics will require egress each year to local compute infrastructure. We anticipate up to insert\_size\_of\_cell\_C1 5 TB in data egress per year for the proposal. This may include both within network egress for analysis purposes and out of Google Cloud Platform egress for researchers to perform analysis on their own systems. Based on current Google Cloud Compute Pricing (accessed MM/DD/YYYY) the cost for egress will be insert\_cell\_E15\_here per year.

**Contingency Storage/Compute:** Estimates for cloud computing costs are based on current market prices and AnVIL infrastructure under the current configuration. We are budgeting a further insert\_your\_buffer per year in contingency costs should we need to store the data on a secondary platform and in anticipation of additional compute needs for the project.

# AnVIL Cloud Cost Budget Templates



The screenshot shows a Google Docs document titled "AnVIL\_Budget\_Justification". The document content is as follows:

**Budget Justification**

**AnVIL Data Storage:** Insert a short description of the project and the type and amount of data that will be collected.

**AnVIL Temporary Data Storage:** We anticipate that up to `insert_cell_C8 here` GB of intermediate processed files, results, figures, and analysis products will be stored at any given time. Based on current Google Cloud Storage Pricing for single region storage (accessed `MM/DD/YYYY`) the cost for this storage will be `insert_cell_D8 here` per month for a total of `insert_cell_E8 here` per year in storage costs.

**AnVIL Batch Processing Costs:** The AnVIL supports batch workflows of genetic and genomic data. To support this activity we are budgeting committed usage of one `insert_cell_A2 here` with `insert_cell_C7 here` GB of attached SSD storage for the entire year. Based on current Google Cloud Compute Pricing (accessed `MM/DD/YYYY`) the cost for this interactive computing will be `insert_cell_D2+D7 here` per month for a total of `insert_cell_E2+E7 here` per year in compute costs.

**AnVIL Interactive Analysis Costs:** The AnVIL supports interactive analysis of genetic and genomic data. We are budgeting the equivalent of one analyst working full time (5 days a week, 8 hours a day) on interactive analysis. To support this activity we estimate `insert_cell_C3 here` hours of compute time per month on `insert_cell_A3 here` with `insert_cell_C7 here` GB of attached SSD storage. Based on current Google Cloud Compute Pricing (accessed `MM/DD/YYYY`) the cost for this interactive computing will be `insert_cell_D3+D7 here` per month for a total of `insert_cell_E3+E7 here` per year in compute costs.

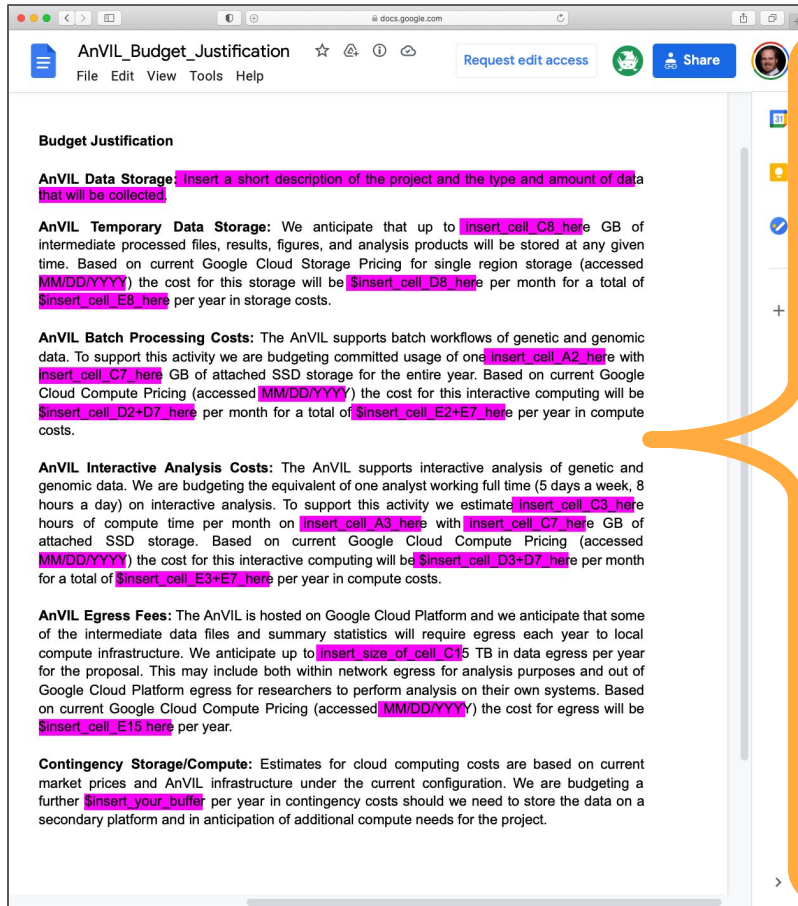
**AnVIL Egress Fees:** The AnVIL is hosted on Google Cloud Platform and we anticipate that some of the intermediate data files and summary statistics will require egress each year to local compute infrastructure. We anticipate up to `insert_size_of_cell_C1` 5 TB in data egress per year for the proposal. This may include both within network egress for analysis purposes and out of Google Cloud Platform egress for researchers to perform analysis on their own systems. Based on current Google Cloud Compute Pricing (accessed `MM/DD/YYYY`) the cost for egress will be `insert_cell_E15 here` per year.

**Contingency Storage/Compute:** Estimates for cloud computing costs are based on current market prices and AnVIL infrastructure under the current configuration. We are budgeting a further `insert_your_buffer` per year in contingency costs should we need to store the data on a secondary platform and in anticipation of additional compute needs for the project.

## Storage Principles

- Keep all essential input and output files to ensure work is reproducible
- Purge intermediate files after successful runs to limit long term data footprint. e.g. T2T: 100Tb -> 5Pb -> 100Tb
- Collect metadata early and often; prefer existing ontologies and standards over developing custom formats
- Prefer compressed formats for long term storage, e.g. BAM or bgzip over fastq/SAM or vcf/txt/fa; Some lossy formats may be acceptable, e.g. CRAM over BAM
- Cloud platforms may be able to provide "free storage" for certain shared datasets

# AnVIL Cloud Cost Budget Templates



**Budget Justification**

**AnVIL Data Storage:** Insert a short description of the project and the type and amount of data that will be collected.

**AnVIL Temporary Data Storage:** We anticipate that up to **insert\_cell\_C8 here** GB of intermediate processed files, results, figures, and analysis products will be stored at any given time. Based on current Google Cloud Storage Pricing for single region storage (accessed **MM/DD/YYYY**) the cost for this storage will be **insert\_cell\_D8 here** per month for a total of **insert\_cell\_E8 here** per year in storage costs.

**AnVIL Batch Processing Costs:** The AnVIL supports batch workflows of genetic and genomic data. To support this activity we are budgeting committed usage of one **insert\_cell\_A2 here** with **insert\_cell\_C7 here** GB of attached SSD storage for the entire year. Based on current Google Cloud Compute Pricing (accessed **MM/DD/YYYY**) the cost for this interactive computing will be **insert\_cell\_D2+D7 here** per month for a total of **insert\_cell\_E2+E7 here** per year in compute costs.

**AnVIL Interactive Analysis Costs:** The AnVIL supports interactive analysis of genetic and genomic data. We are budgeting the equivalent of one analyst working full time (5 days a week, 8 hours a day) on interactive analysis. To support this activity we estimate **insert\_cell\_C3 here** hours of compute time per month on **insert\_cell\_A3 here** with **insert\_cell\_C7 here** GB of attached SSD storage. Based on current Google Cloud Compute Pricing (accessed **MM/DD/YYYY**) the cost for this interactive computing will be **insert\_cell\_D3+D7 here** per month for a total of **insert\_cell\_E3+E7 here** per year in compute costs.

**AnVIL Egress Fees:** The AnVIL is hosted on Google Cloud Platform and we anticipate that some of the intermediate data files and summary statistics will require egress each year to local compute infrastructure. We anticipate up to **insert\_size\_of\_cell\_C1** 5 TB in data egress per year for the proposal. This may include both within network egress for analysis purposes and out of Google Cloud Platform egress for researchers to perform analysis on their own systems. Based on current Google Cloud Compute Pricing (accessed **MM/DD/YYYY**) the cost for egress will be **insert\_cell\_E15 here** per year.

**Contingency Storage/Compute:** Estimates for cloud computing costs are based on current market prices and AnVIL infrastructure under the current configuration. We are budgeting a further **insert\_your\_buffer** per year in contingency costs should we need to store the data on a secondary platform and in anticipation of additional compute needs for the project.

## Computing Principles

- Interactive analyses (e.g. RStudio or Jupyter notebooks) tend to be very inexpensive (<\$1/hr) and will feel very familiar to desktop counterparts
- Batch analyses (e.g. WDL/CWL/Galaxy Workflows) vary enormously in computing costs from <<\$1 to >>\$10k
- Your spend rate is primarily determined by the #virtual machines (cores x RAM x GPUs x disk) running in parallel plus the amount of cloud storage used
- Your spend rate is ultimately limited by your quotas. GCP has separate quotas for VMs, cores, RAM, GPUs, IP Addresses, etc.
  - Increase quotas to accelerate analysis
  - Decrease quotas to throttle spend
- Benchmark, benchmark, benchmark....



# Galaxy Usage (usegalaxy.org)

**Galaxy** Workflow Visualize Shared Data Help User

Using 0%

Tools  Upload Data

- Get Data
- Collection Operations
- GENERAL TEXT TOOLS
  - Text Manipulation
  - Filter and Sort
  - Join, Subtract and Group
  - Datamash
- GENOMIC FILE MANIPULATION
  - FASTA/FASTQ
  - FASTQ Quality Control
  - SAM/BAM
  - BED
  - VCF/BCF
  - Nanopore
  - Convert Formats
  - Lift-Over
- COMMON GENOMICS TOOLS
  - Interactive tools
  - Operate on Genomic Intervals
  - Fetch Sequences/Alignments
- GENOMICS ANALYSIS
  - Assembly
  - Annotation
  - Mapping
  - Variant Calling
  - ChIP-seq

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

**James P. Taylor Foundation for Open Science.**

“The most important job of senior faculty is to mentor junior faculty and students.” — @jtxx

**Announcing the James P. Taylor (JTX) Foundation for Open Science**

[Learn More](#)

Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal at [covid19.galaxyproject.org](https://covid19.galaxyproject.org)

**History** search datasets

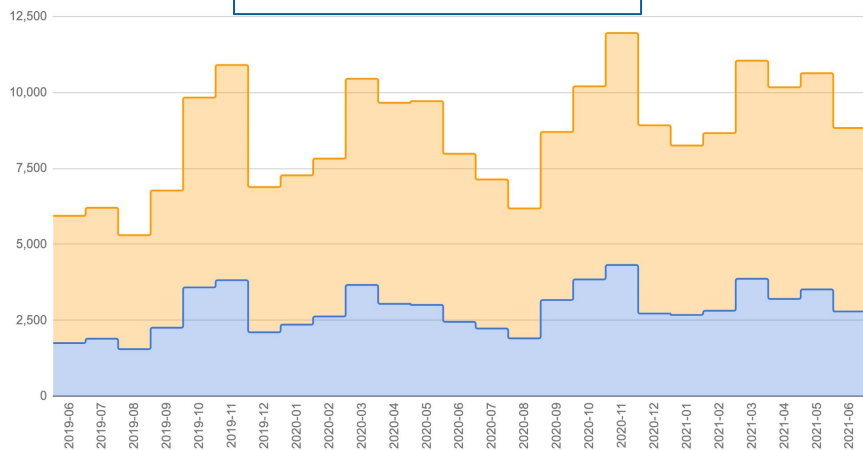
**vadsti-2021**  
23 shown, 9 deleted  
30.01 MB

- 30: transeq on data 29
- 29: bedtools GetFastaBed on data 13 and data 28
- 28: Text reformatting on data 27
- 27: Advanced Cut on data 26
- 26: DNAdiff on data 13 and data 5: qdiff
- 25: DNAdiff on data 13 and data 5: rdiff
- 24: DNAdiff on data 13 and data 5: snps
- 23: DNAdiff on data 13 and data 5: mcoords
- 22: DNAdiff on data 13 and data 5: 1coords
- 21: DNAdiff on data 13 and data 5: mdelta
- 20: DNAdiff on data 13 and data 5: 1delta
- 19: DNAdiff on data 13 and data 5: delta
- 18: DNAdiff on data 13 and data 5: report
- 15: SPAdes on data 4, data 3, and others: log

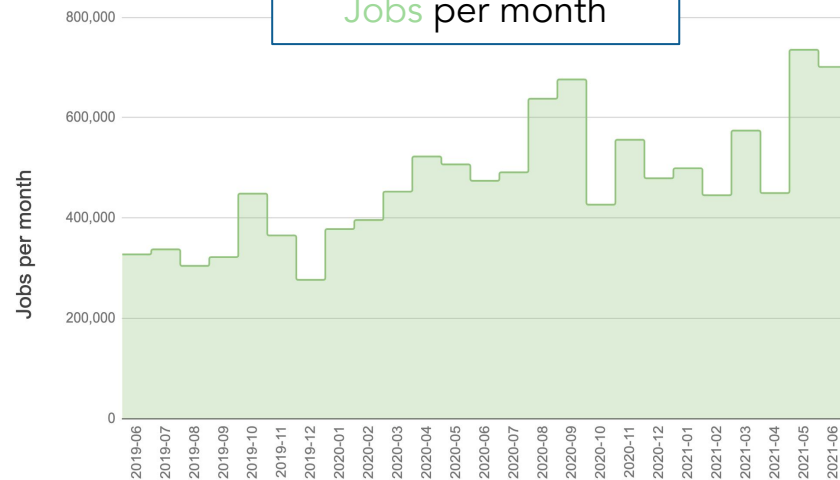
# Galaxy Usage (usegalaxy.org)

The screenshot shows the Galaxy web interface. At the top, there is a navigation bar with 'Galaxy' logo and menu items: 'Workflow', 'Visualize', 'Shared Data', 'Help', 'User', and a grid icon. On the right, it says 'Using 0%'. Below the navigation bar, there is a 'Tools' section with a search bar and an 'Upload Data' button. A central banner features the text 'James P. Taylor Foundation for Open Science'. On the right, there is a 'History' section with a search bar and a list of datasets, including 'vadsti-2021' and '30: transeq on data 29'.

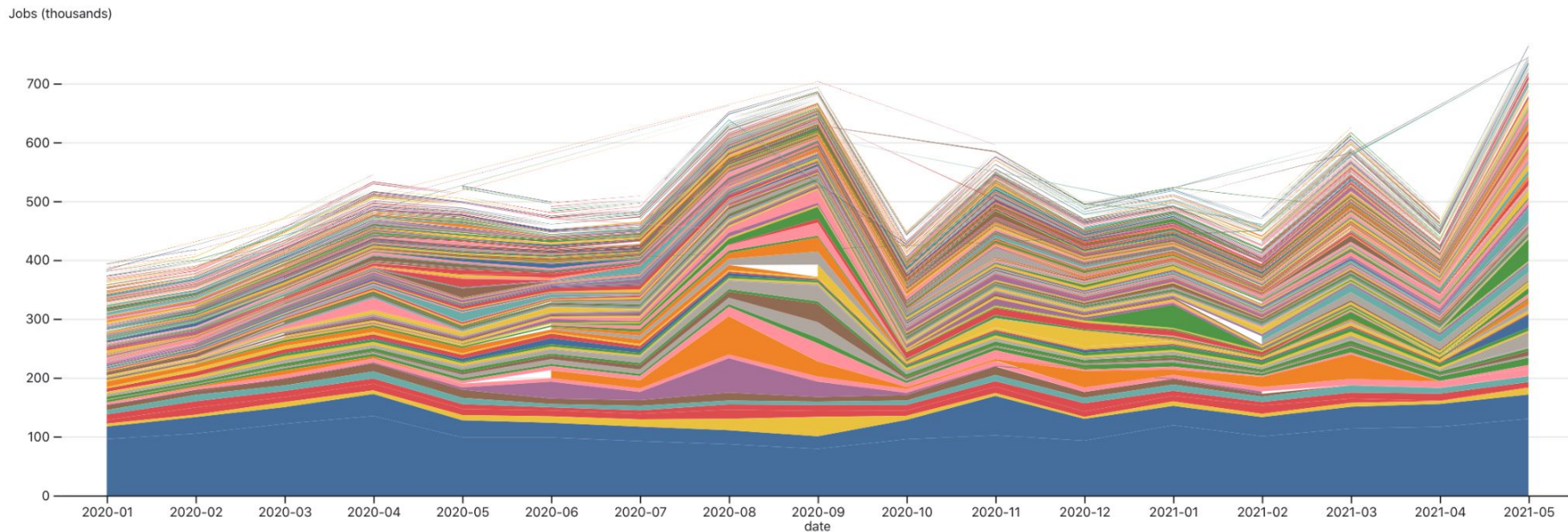
New and Active users



Jobs per month



# Phase I: Jobs per month, by tool



Upload (galaxy internal)

fastqc

bwa\_mem

bowtie2

hisat2

fasterq\_dump

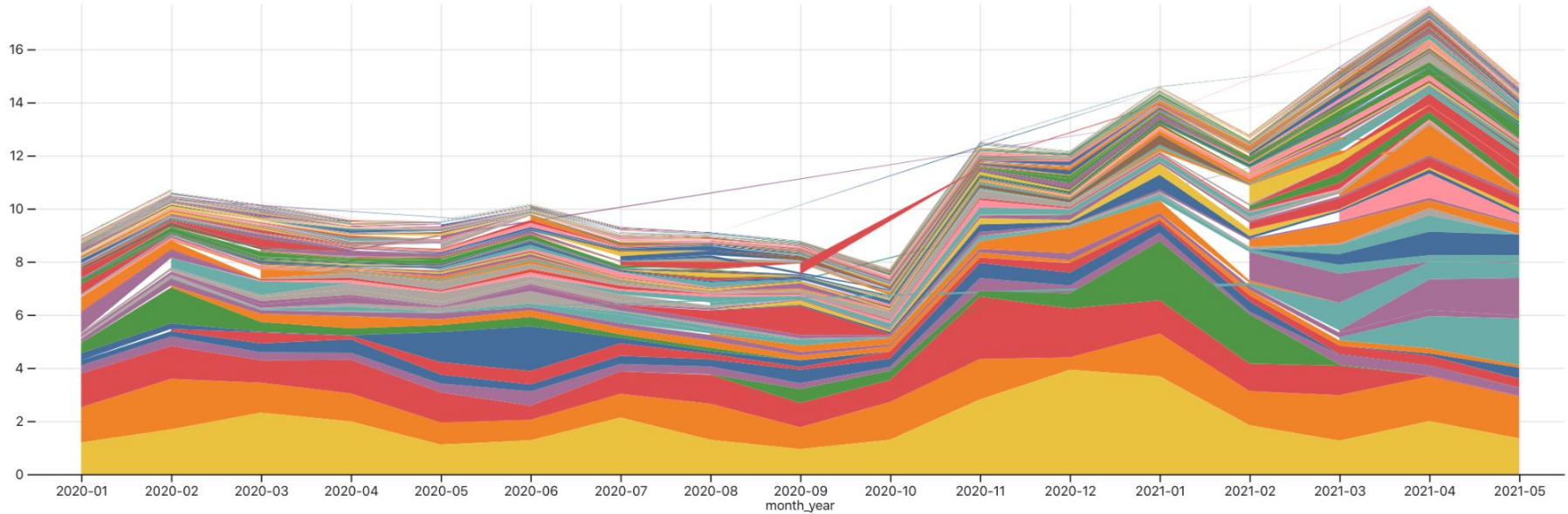
fastp

picard markDuplicates

abricate

# Phase I: Total CPU time per month, by tool

total\_cpu\_time (thousand trillion)



tophat2

bwa\_mem

bowtie2

rna\_star

bwa

hisat2

freebayes

htseq\_compare

lofreq\_call

# Phase II: Benchmarking Popular Tools

**Tools** | **Benchmarking RNA-seq Cloud Costs**

search tools

**Inputs**

- FASTA Dataset (output (input))

**Get Data**

**Collection Operations**

**GENERAL TEXT TOOLS**

- Text Manipulation
- Filter and Sort
- Join, Subtract and Group

**GENOMIC FILE MANIPULATION**

- FASTQ Quality Control
- SAM/BAM
- BED
- VCF/BCF
- Nanopore
- Convert Formats
- Lift-Over

**MISCELLANEOUS TOOLS**

- Virology

**COMMON GENOMICS TOOLS**

- Operate on Genomic Intervals
- Fetch Sequences/Alignments

**GENOMICS ANALYSIS**

- Assembly
- Mapping
- Variant Calling
- RNA-seq
- Multiple Alignments
- Phenotype Association
- deepTools
- RSeQC

**STATISTICS AND VISUALIZATION**

- Statistics
- Graph/Display Data

**Workflow Steps:**

- Map with BWA-MEM**
  - Select fastq dataset
  - Map with BWA-MEM on input dataset(s) (mapped reads in BAM format) (bam)
- Bowtie2**
  - FASTA/Q file
  - Bowtie2 on input dataset(s): alignments (bam, qname\_input\_sorted.bam, sam)
- Kallisto quant**
  - Reads in FASTQ format
  - Kallisto quant on input dataset(s): Abundances (HDF5) (h5)
  - Kallisto quant on input dataset(s): Abundances (tabular) (tabular)
- Salmon quant**
  - FASTQ/FASTA file
  - File containing a mapping of transcripts to genes
  - Salmon quant on input dataset(s) (Quantification) (tabular)
  - Salmon quant on input dataset(s) (Gene Quantification) (tabular)
- RNA STAR**
  - RNA-Seq FASTQ/FASTA file
  - RNA STAR on input dataset(s): log (txt)
  - RNA STAR on input dataset(s): splice junctions.bed (interval)
  - RNA STAR on input dataset(s): mapped.bam (bam)
- StringTie**
  - Input mapped reads
  - StringTie on input dataset(s): Assembled transcripts (gtf)

**Map with BWA-MEM**  
- map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.17.1)

**Label**

Add a step label.

**Step Annotation**

Add an annotation or notes to this step. Annotations are available when a workflow is viewed.

**Will you select a reference genome from your history or use a built-in index?**

Use a built-in genome index

Built-ins were indexed using default options. See 'Indexes' section of help below

**Using reference genome**

Human (Homo sapiens) (b3...

Select genome from the list

**Single or Paired-end reads**

Single

Select between paired and single end data

**Select fastq dataset**

Data input 'fastq\_input1' (fastqsanger, fastqsanger.gz or fasta)

Specify dataset with single reads

**Set read groups information?**

Do not set

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

**Select analysis mode**

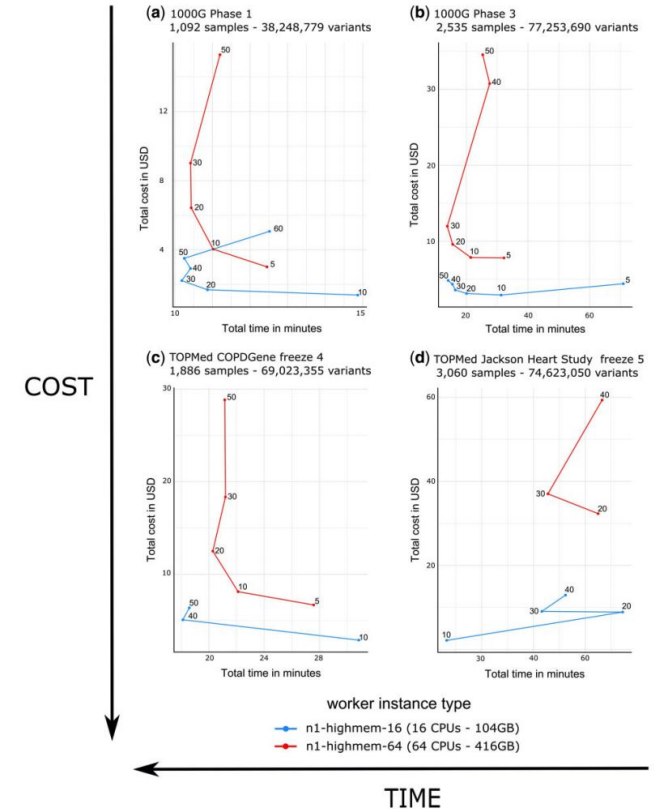
# Phase III: Modeling performance

## Expected Results

- Empirically measure the runtime, costs, and other performance metrics for several popular tools
- Also measure the performance when scaling for large numbers of samples / large amounts of data
- Particularly important to identify non-linear performance (e.g. 10-fold more data is more than 10-fold more expensive)

## Deployment Strategy

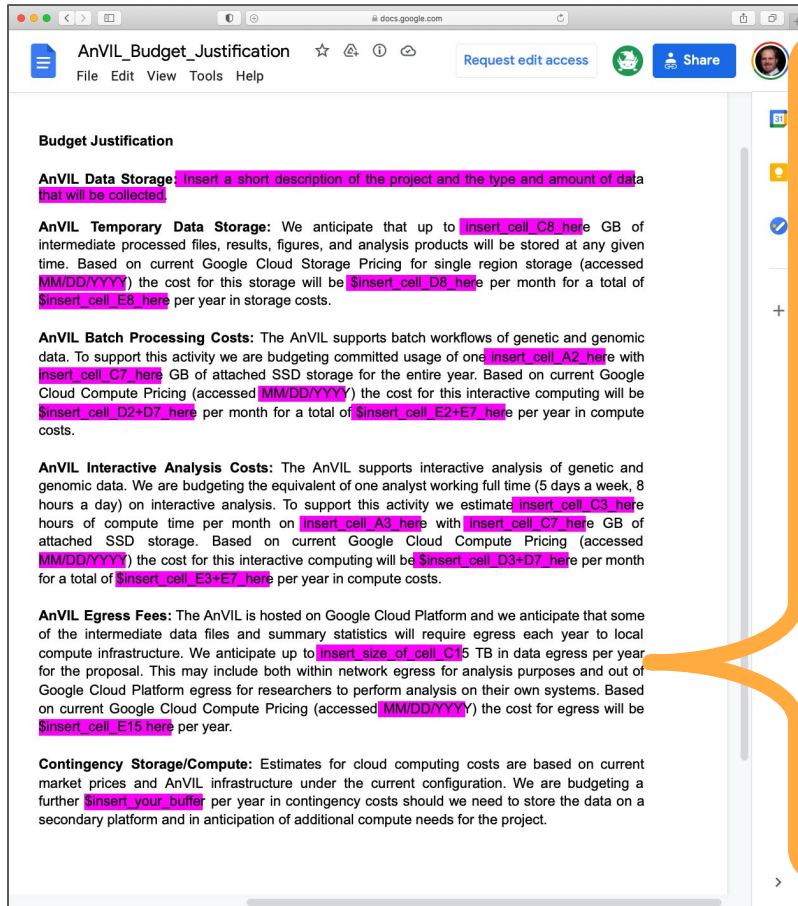
- Generate a tool-based lookup table with a range of inputs available
- Implement an API service to query for results
- Expand the range of tools covered by developing a predictive model



Scalability and cost-effectiveness analysis of whole genome-wide association studies on Google Cloud Platform and Amazon Web Services.

Krissaane et al. (2020) Journal of the American Medical Informatics Association. doi:10.1093/jamia/ocaa068

# AnVIL Cloud Cost Budget Templates



The screenshot shows a Google Docs document titled "AnVIL\_Budget\_Justification". The document content is as follows:

**Budget Justification**

**AnVIL Data Storage:** Insert a short description of the project and the type and amount of data that will be collected.

**AnVIL Temporary Data Storage:** We anticipate that up to **insert\_cell\_C8 here** GB of intermediate processed files, results, figures, and analysis products will be stored at any given time. Based on current Google Cloud Storage Pricing for single region storage (accessed **MM/DD/YYYY**) the cost for this storage will be **insert\_cell\_D8 here** per month for a total of **insert\_cell\_E8 here** per year in storage costs.

**AnVIL Batch Processing Costs:** The AnVIL supports batch workflows of genetic and genomic data. To support this activity we are budgeting committed usage of one **insert\_cell\_A2 here** with **insert\_cell\_C7 here** GB of attached SSD storage for the entire year. Based on current Google Cloud Compute Pricing (accessed **MM/DD/YYYY**) the cost for this interactive computing will be **insert\_cell\_D2+D7 here** per month for a total of **insert\_cell\_E2+E7 here** per year in compute costs.

**AnVIL Interactive Analysis Costs:** The AnVIL supports interactive analysis of genetic and genomic data. We are budgeting the equivalent of one analyst working full time (5 days a week, 8 hours a day) on interactive analysis. To support this activity we estimate **insert\_cell\_C3 here** hours of compute time per month on **insert\_cell\_A3 here** with **insert\_cell\_C7 here** GB of attached SSD storage. Based on current Google Cloud Compute Pricing (accessed **MM/DD/YYYY**) the cost for this interactive computing will be **insert\_cell\_D3+D7 here** per month for a total of **insert\_cell\_E3+E7 here** per year in compute costs.

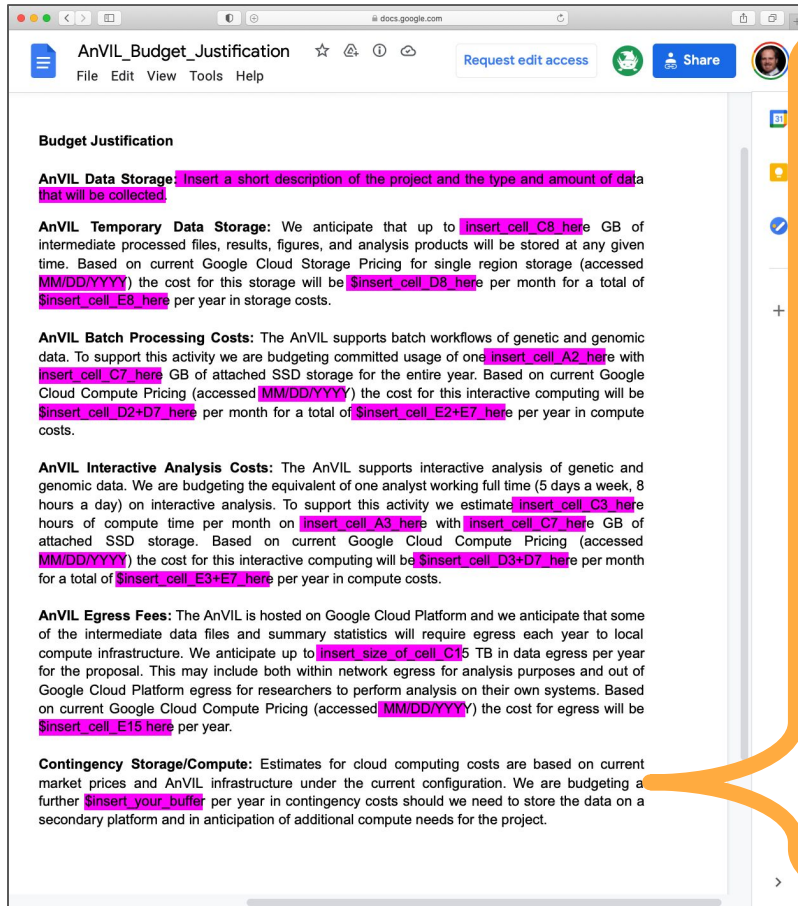
**AnVIL Egress Fees:** The AnVIL is hosted on Google Cloud Platform and we anticipate that some of the intermediate data files and summary statistics will require egress each year to local compute infrastructure. We anticipate up to **insert\_size\_of\_cell\_C1** 5 TB in data egress per year for the proposal. This may include both within network egress for analysis purposes and out of Google Cloud Platform egress for researchers to perform analysis on their own systems. Based on current Google Cloud Compute Pricing (accessed **MM/DD/YYYY**) the cost for egress will be **insert\_cell\_E15 here** per year.

**Contingency Storage/Compute:** Estimates for cloud computing costs are based on current market prices and AnVIL infrastructure under the current configuration. We are budgeting a further **insert\_your\_buffer** per year in contingency costs should we need to store the data on a secondary platform and in anticipation of additional compute needs for the project.

## Egress Principles

- Whenever possible, avoid egress fees by computing in the cloud as much as possible
- Prefer egress of summary/distilled files rather than raw data, e.g. egress vcf + samtools stats instead of CRAM
- \*\*\*When allowed\*\*\*, avoid egressing multiple times and share data with collaborators via other means, e.g. SFTP, Globus, etc
- Within AnVIL we are actively evaluating alternate approaches, e.g. GTEx is mirrored via an academic cloud to avoid egress fees

# AnVIL Cloud Cost Budget Templates



**Budget Justification**

**AnVIL Data Storage:** Insert a short description of the project and the type and amount of data that will be collected.

**AnVIL Temporary Data Storage:** We anticipate that up to insert\_cell\_C8 here GB of intermediate processed files, results, figures, and analysis products will be stored at any given time. Based on current Google Cloud Storage Pricing for single region storage (accessed MM/DD/YYYY) the cost for this storage will be insert\_cell\_D3 here per month for a total of insert\_cell\_E8 here per year in storage costs.

**AnVIL Batch Processing Costs:** The AnVIL supports batch workflows of genetic and genomic data. To support this activity we are budgeting committed usage of one insert\_cell\_A2 here with insert\_cell\_C7 here GB of attached SSD storage for the entire year. Based on current Google Cloud Compute Pricing (accessed MM/DD/YYYY) the cost for this interactive computing will be insert\_cell\_D2+D7 here per month for a total of insert\_cell\_E2+E7 here per year in compute costs.

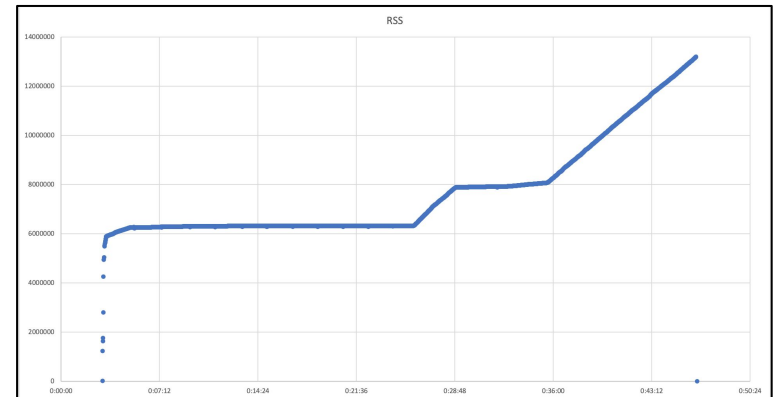
**AnVIL Interactive Analysis Costs:** The AnVIL supports interactive analysis of genetic and genomic data. We are budgeting the equivalent of one analyst working full time (5 days a week, 8 hours a day) on interactive analysis. To support this activity we estimate insert\_cell\_C3 here hours of compute time per month on insert\_cell\_A3 here with insert\_cell\_C7 here GB of attached SSD storage. Based on current Google Cloud Compute Pricing (accessed MM/DD/YYYY) the cost for this interactive computing will be insert\_cell\_D3+D7 here per month for a total of insert\_cell\_E3+E7 here per year in compute costs.

**AnVIL Egress Fees:** The AnVIL is hosted on Google Cloud Platform and we anticipate that some of the intermediate data files and summary statistics will require egress each year to local compute infrastructure. We anticipate up to insert\_size\_of\_cell\_C15 TB in data egress per year for the proposal. This may include both within network egress for analysis purposes and out of Google Cloud Platform egress for researchers to perform analysis on their own systems. Based on current Google Cloud Compute Pricing (accessed MM/DD/YYYY) the cost for egress will be insert\_cell\_E15 here per year.

**Contingency Storage/Compute:** Estimates for cloud computing costs are based on current market prices and AnVIL infrastructure under the current configuration. We are budgeting a further insert\_your\_buffer here per year in contingency costs should we need to store the data on a secondary platform and in anticipation of additional compute needs for the project.

## Contingency Principles

- Research is hard; Consider reserving 10% (or more) of your computing budget for unexpected errors
- The more “exotic” your analyses, the more you should reserve
- Start small, scale up slowly to find issues as early as possible





# Summary & Future work

---

## *Computing in the cloud offers tremendous advantages for scalability and efficiency*

- Previously, your available RAM / Disk / Cores were the biggest considerations for computing, but now cost is the single largest factor
- Apply for STRIDES discounts, purge intermediate files, compress the rest
- Benchmark, benchmark, & benchmark; pick optimized instance types; use quotas to throttle spend and accelerate compute

## *Once cost bottlenecks are identified, improve performance through:*

- Decreasing RAM requirements using more advanced data structures (e.g. Burrows-Wheeler transform (Langmead et al. 2009), Bloom filters (Chikhi and Rizk 2013), or Sequence Bloom Tree (Solomon and Kingsford 2016))
- Decreasing computing time by leveraging parallel & vectorized computing instructions (e.g. AVX512 vectorization (Darby et al. 2020)) or advanced search strategies (e.g. learned index structures (Kirsche et al. 2020; Kraska et al. 2017))
- Decreasing storage requirements by using compressed data formats (e.g. CRAM (Hsi-Yang Fritz et al. 2011)), using optimized IO routines (e.g. fixed length records instead of variable length records (Langmead et al. 2019)), and removing intermediate data.

# Acknowledgements

---

## Galaxy Community

Enis Afgan

Keith Suderman

Dannon Baker

Sergey Golitsynskiy

Bridget Carr

Victor Wen

Peiyuan Xu



National Human Genome  
Research Institute

## AnVIL Team

Jeff Leek

Fred Tan

Stephen Mosher

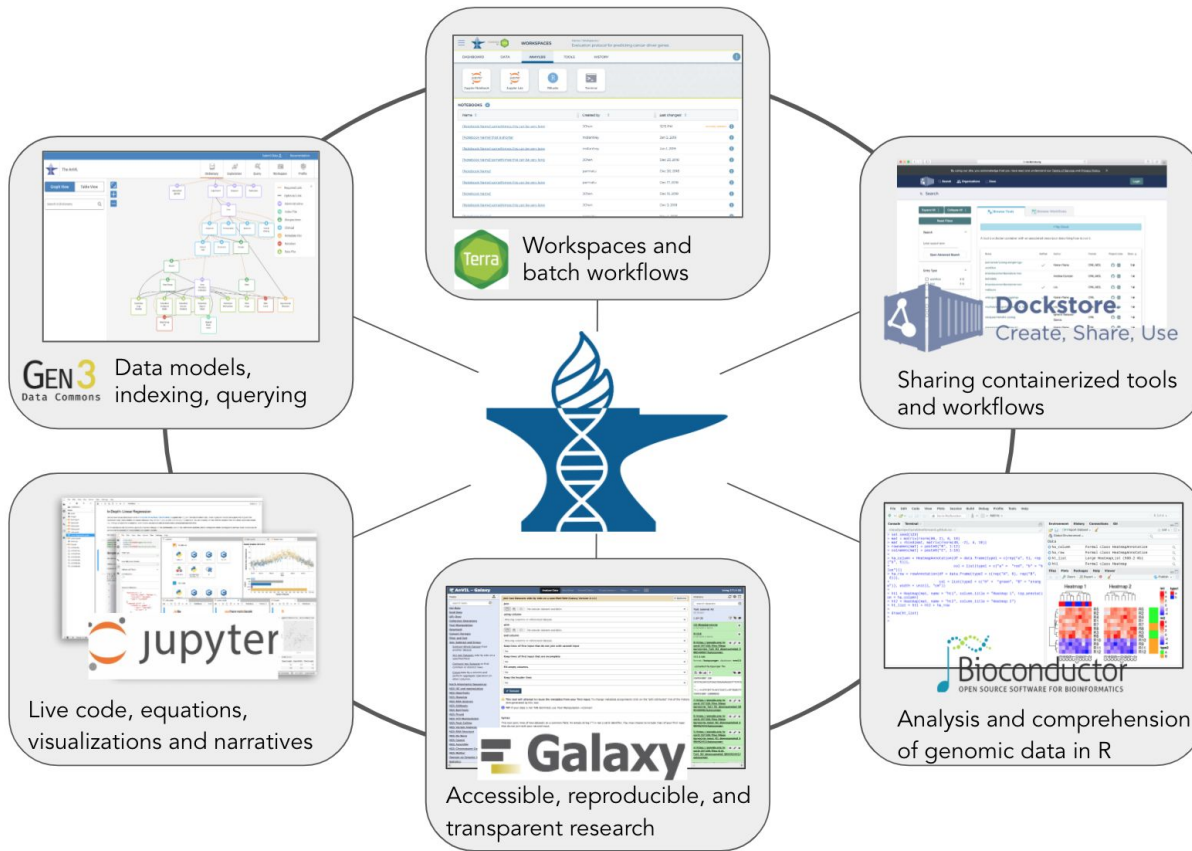
Sarah Wheelan

Ava Hoffman

David Rogers



National Institutes of Health  
*Office of Data Science Strategy*



# Thank you!

 @useAnVIL