AnVIL

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

Michael Schatz

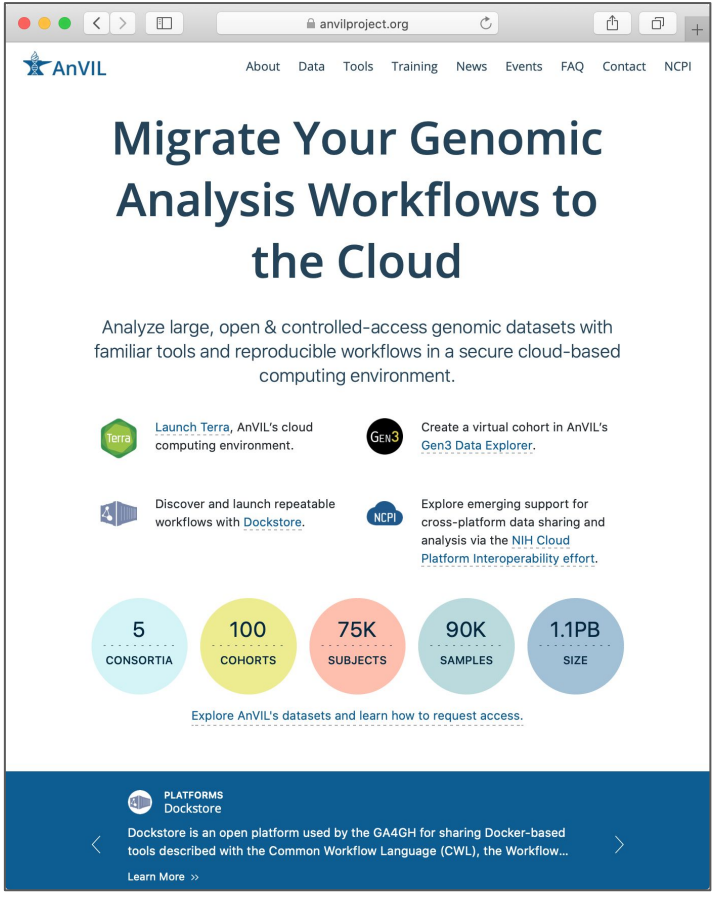
@mike_schatz

July 26, 2021
ISMB NIH/ODSS Workshop

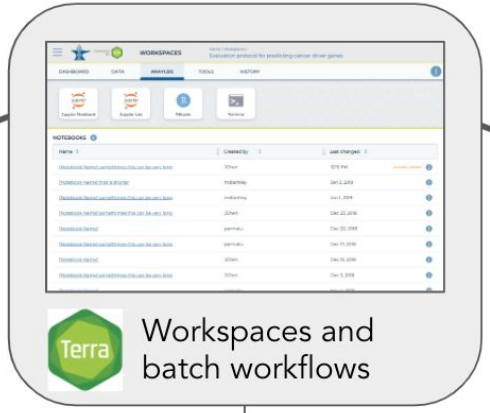
What is the AnVIL?

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



https://anvilproject.org

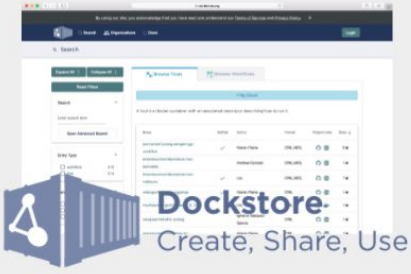


Terra

Workspaces and batch workflows

GEN3 Data Commons

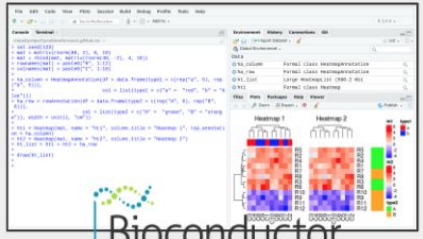
Data models, indexing, querying



Sharing containerized tools and workflows

Live code, equations, visualizations and narratives

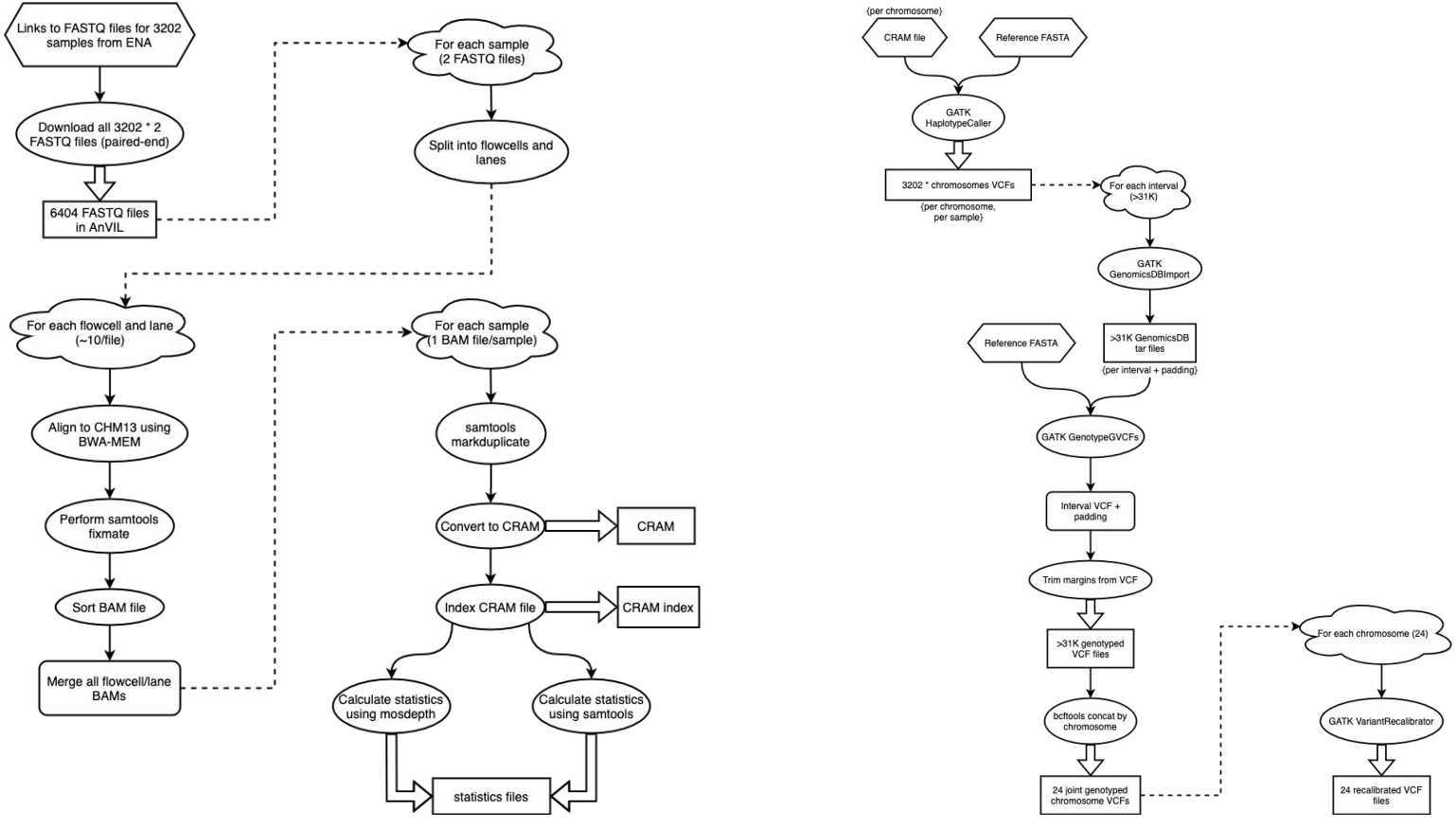
Accessible, reproducible, and transparent research



Analysis and comprehension of genomic data in R

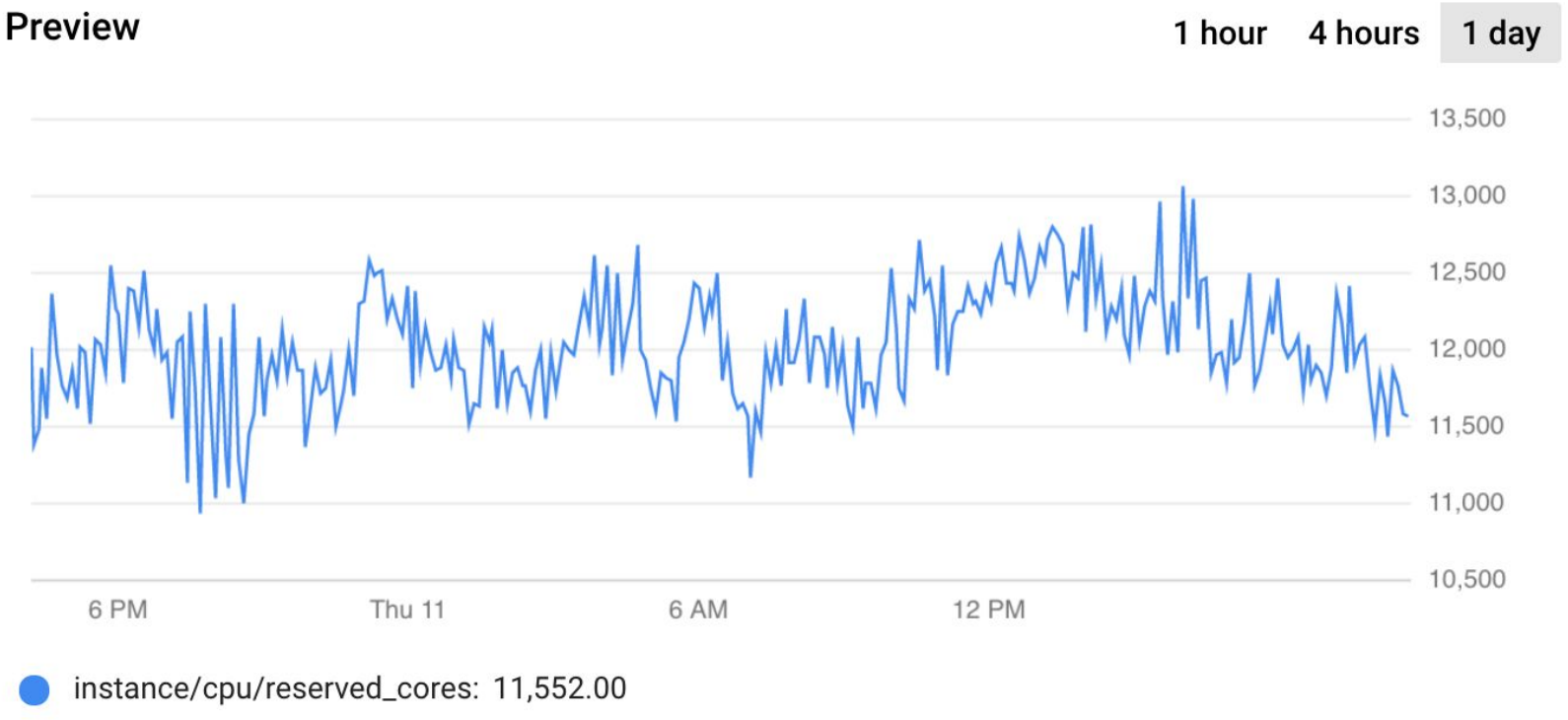
http://anvilproject.org

T2T Analysis in AnVIL

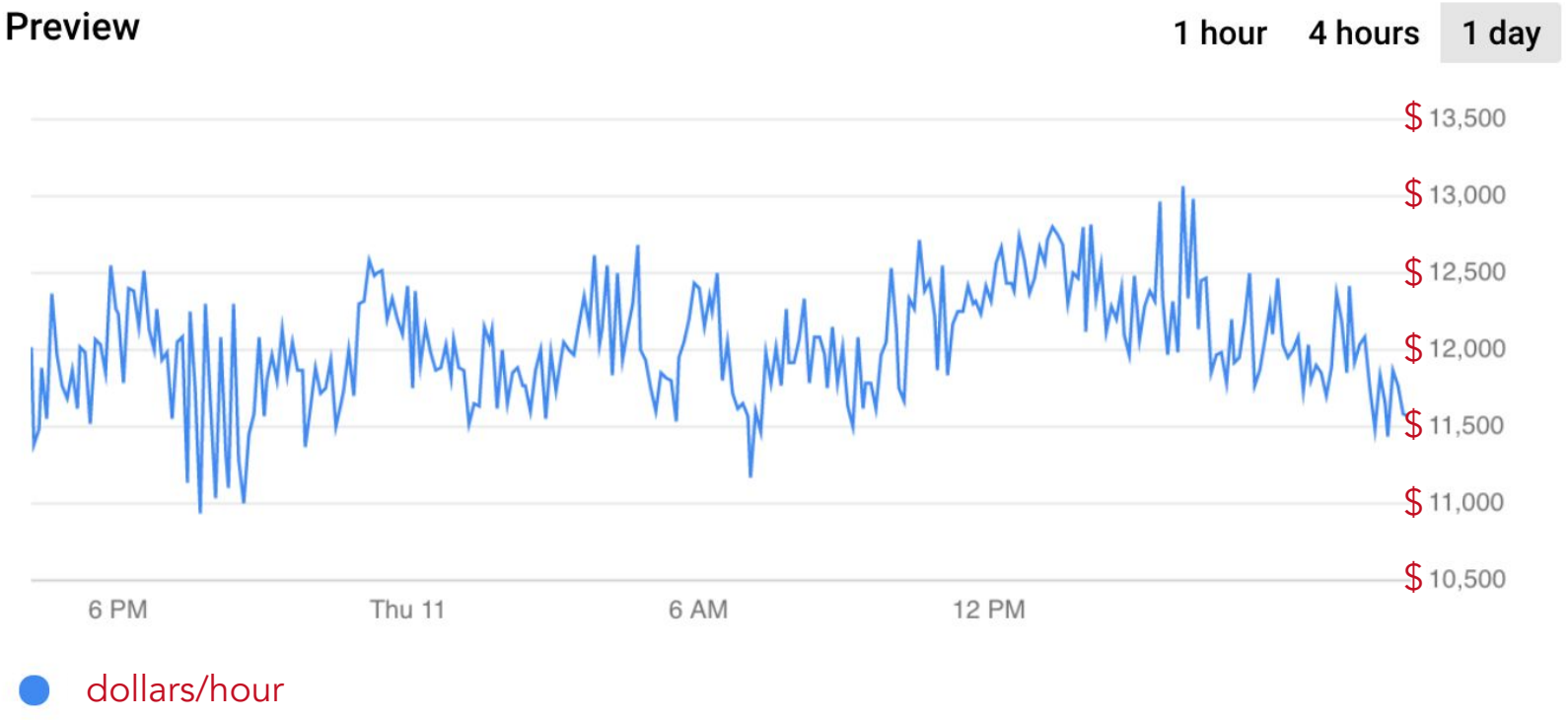


A complete reference genome improves analysis of human genetic variation
Aganezov, S*, Yan, SM*, Soto, DC*, Kirsche, M*, Zarate, S*, *et al.* (2021) *bioRxiv* doi: 10.1101/2021.07.12.452063

T2T Analysis on Google Cloud Platform



T2T Analysis on Google Cloud Platform



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 / Hourly

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 E2 custom vCPUs and memory.			

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 / Hourly

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 E2 custom vCPUs and memory.			

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 / Hourly

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 E2 custom vCPUs and memory.			

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 / Hourly

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 Custom vCPU and memory.			

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 / Hourly

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 Custom vCPU and memory.			

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.

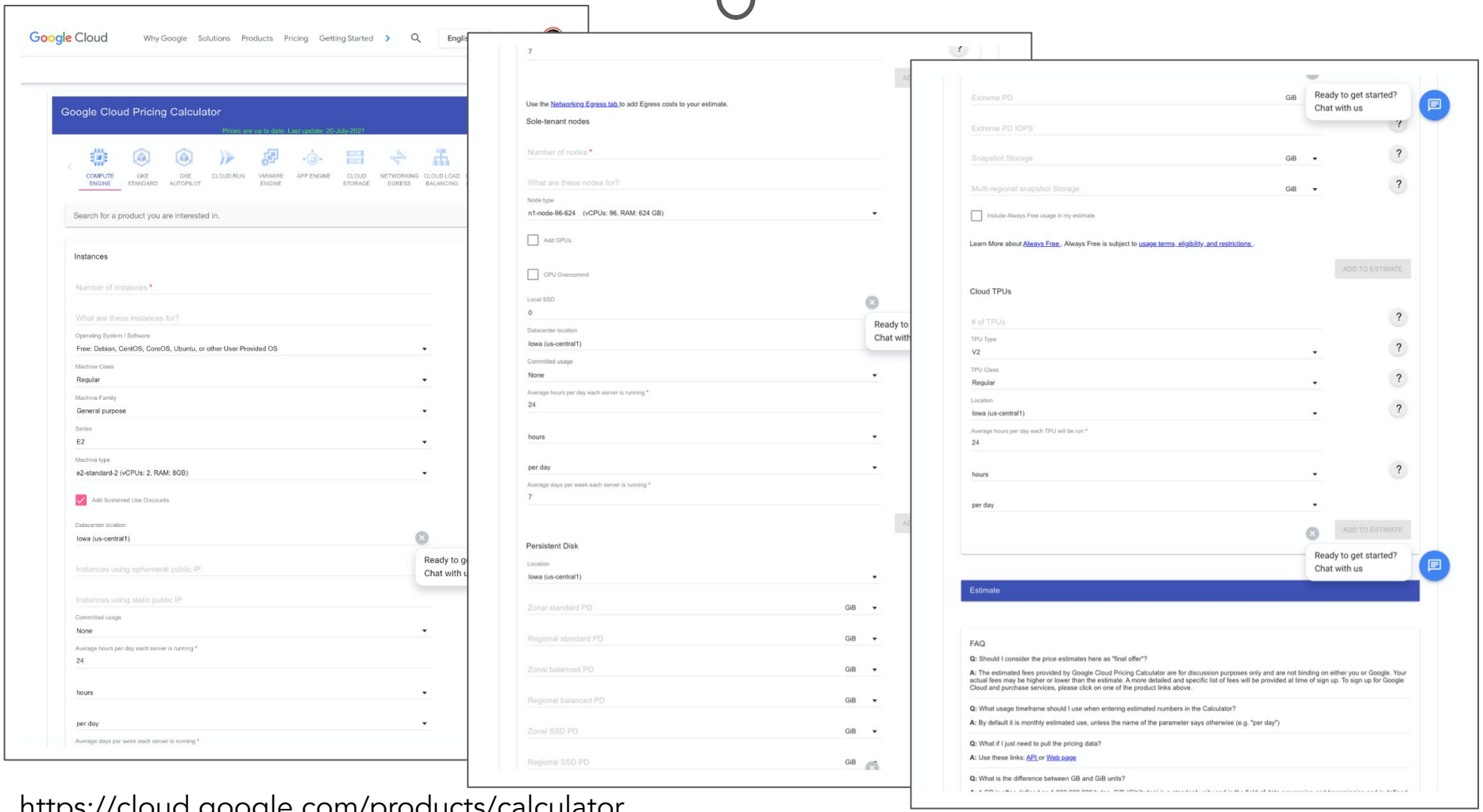
High-CPU machine types have one vCPU for every 1 GB of memory. High-CPU machine types are ideal for tasks that require moderate memory configurations for the needed vCPU count.

Iowa (us-central1) — Monthly / Hourly

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 Custom vCPUs and memory.			

https://cloud.google.com/compute/vm-instance-pricing

Cloud Costs are complicated



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

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

<u>Purchased car</u>

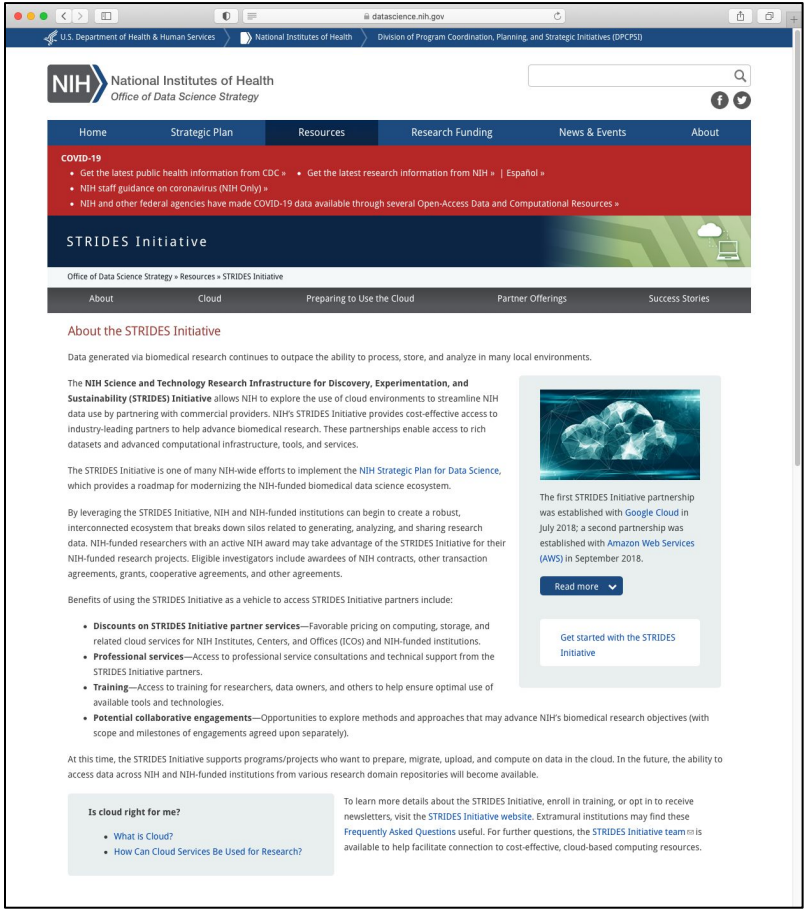
- You buy the car
- You fill up at a station
- You pay less per mile

<u>ZipCar</u>

- 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

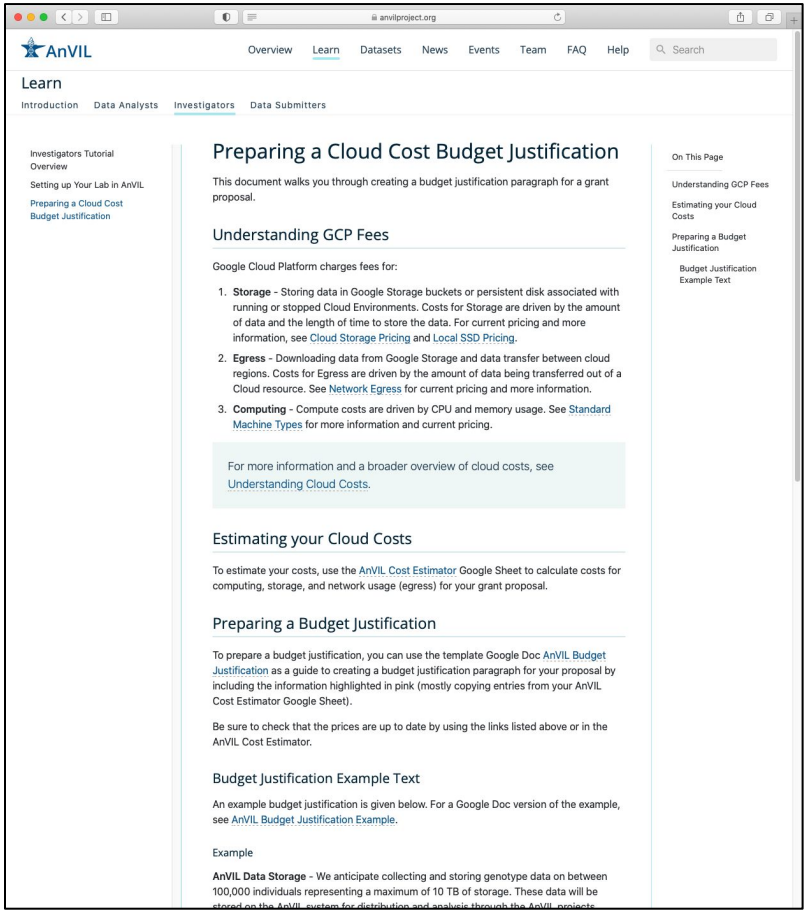


https://datascience.nih.gov/strides

STRIDES Benefits

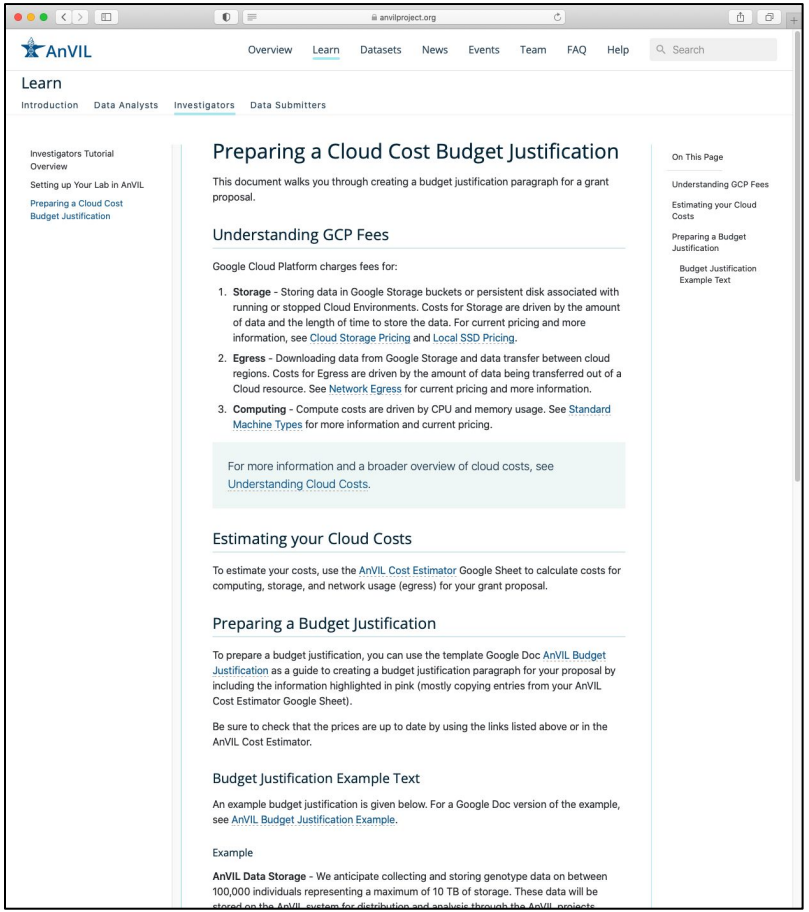
- *Discounts* (typically 10%-25%) on computing, storage, and related cloud services for NIH Institutes, Centers, and Offices (ICOs) and NIH-funded institutions & investigators.
- *Professional services* — Access to professional service consultations and technical support from the STRIDES Initiative partners.
- *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

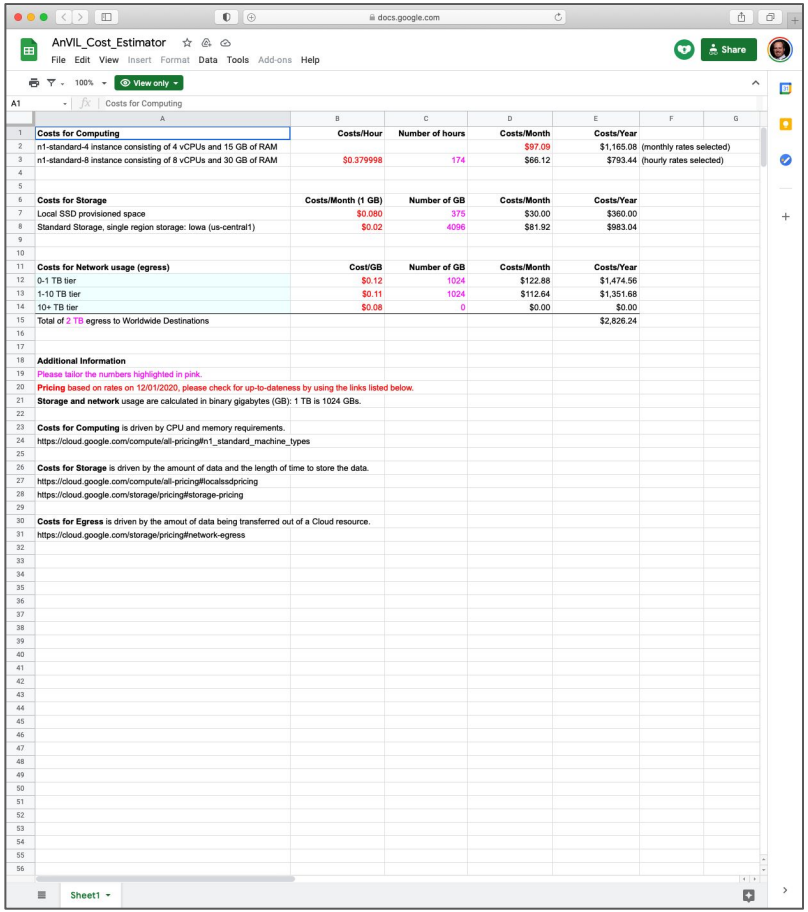
AnVIL Cloud Cost Budget Templates



https://anvilproject.org/learn/investigators/budget-templates

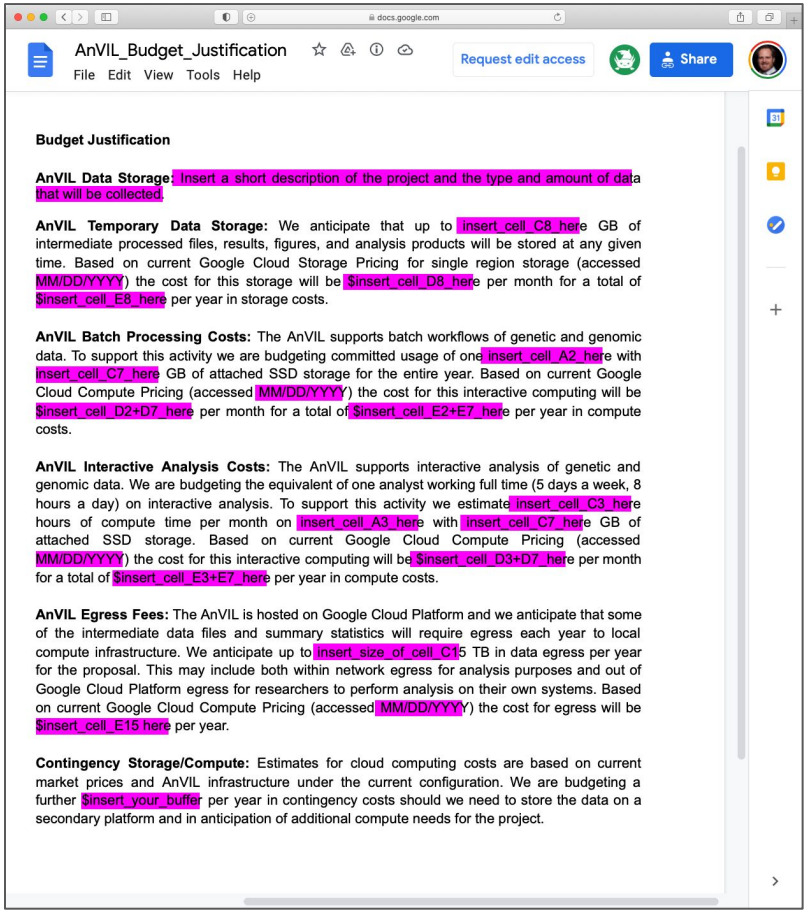
AnVIL Cloud Cost Budget Templates





https://anvilproject.org/learn/investigators/budget-templates

AnVIL Cloud Cost Budget Templates



AnVIL_Budget_Justification

File Edit View Tools Help

Request edit access | Share

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.

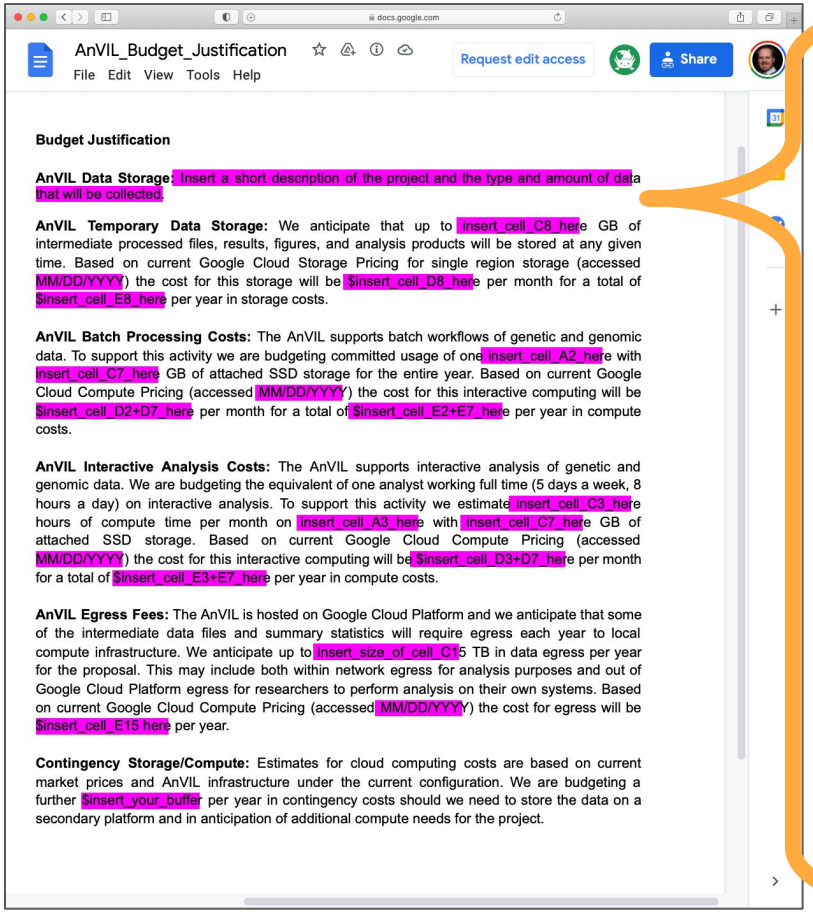
AnVIL Interactive Analysis Costs: The AnVIL supports interactive analysis of genetic and genomic data. We are budgeting the equivalent of one analyst working full time (5 days a week, 8 hours a day) on interactive analysis. To support this activity we estimate insert_cell_C3_here hours of compute time 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_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 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.

https://anvilproject.org/learn/investigators/budget-templates

AnVIL Cloud Cost Budget Templates



AnVIL_Budget_Justification

File Edit View Tools Help

Request edit access | Share

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_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.

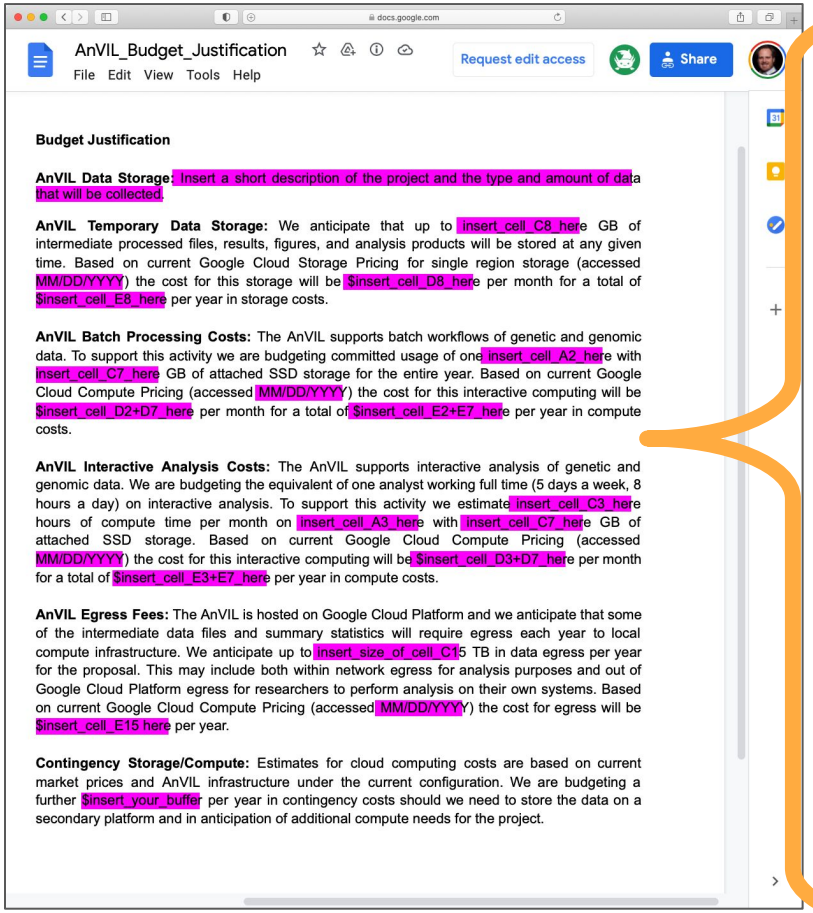
Contingency Storage/Compute: Estimates for cloud computing costs are based on current market prices and AnVIL infrastructure under the current configuration. We are budgeting a further $insert_your_buffer per year in contingency costs should we need to store the data on a secondary platform and in anticipation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

https://anvilproject.org/learn/investigators/budget-templates

AnVIL Cloud Cost Budget Templates



AnVIL_Budget_Justification

File Edit View Tools Help

Request edit access | Share

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_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.

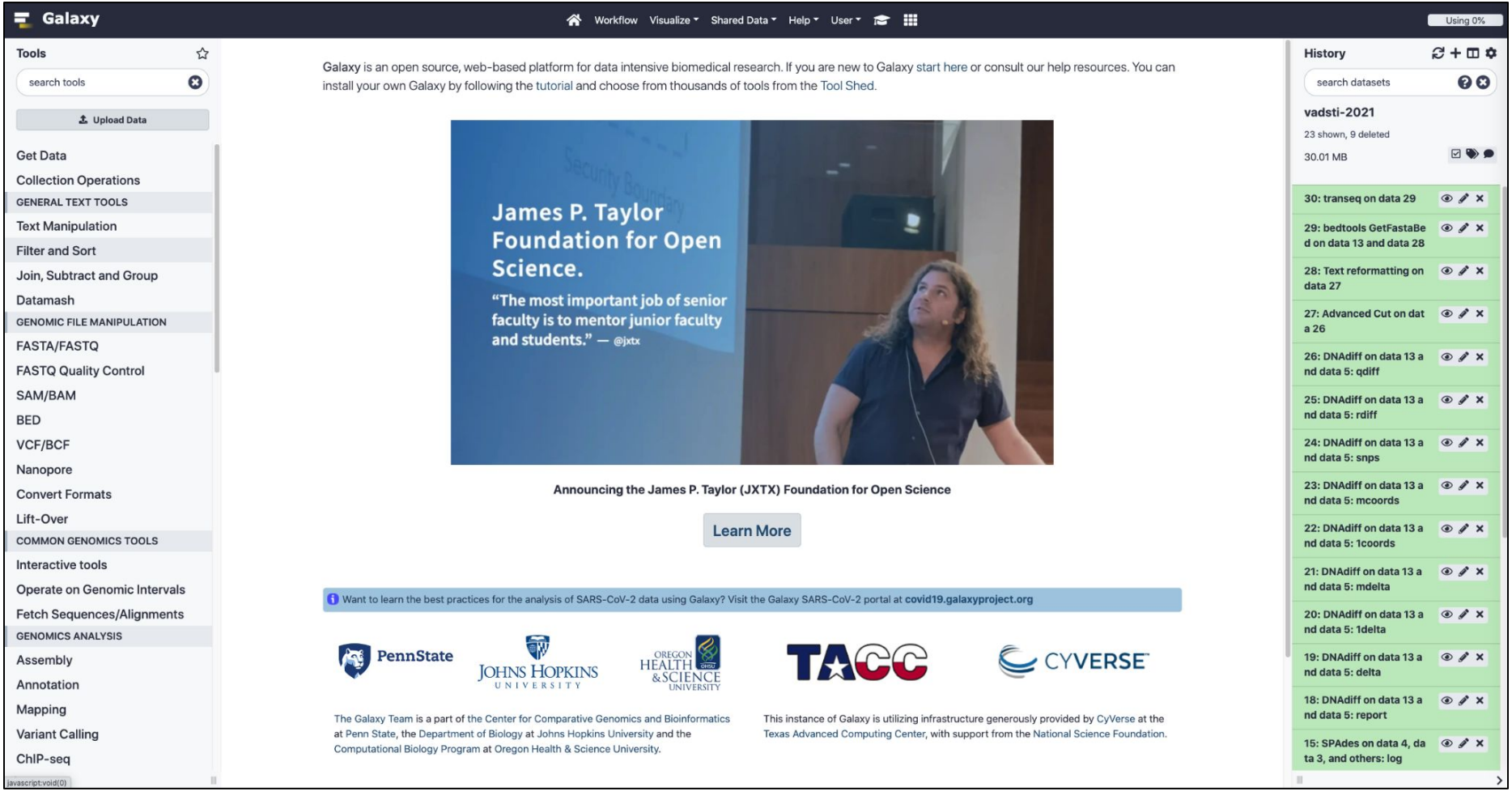
Contingency Storage/Compute: Estimates for cloud computing costs are based on current market prices and AnVIL infrastructure under the current configuration. We are budgeting a further $insert_your_buffer per year in contingency costs should we need to store the data on a secondary platform and in anticipation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

Computing Principles

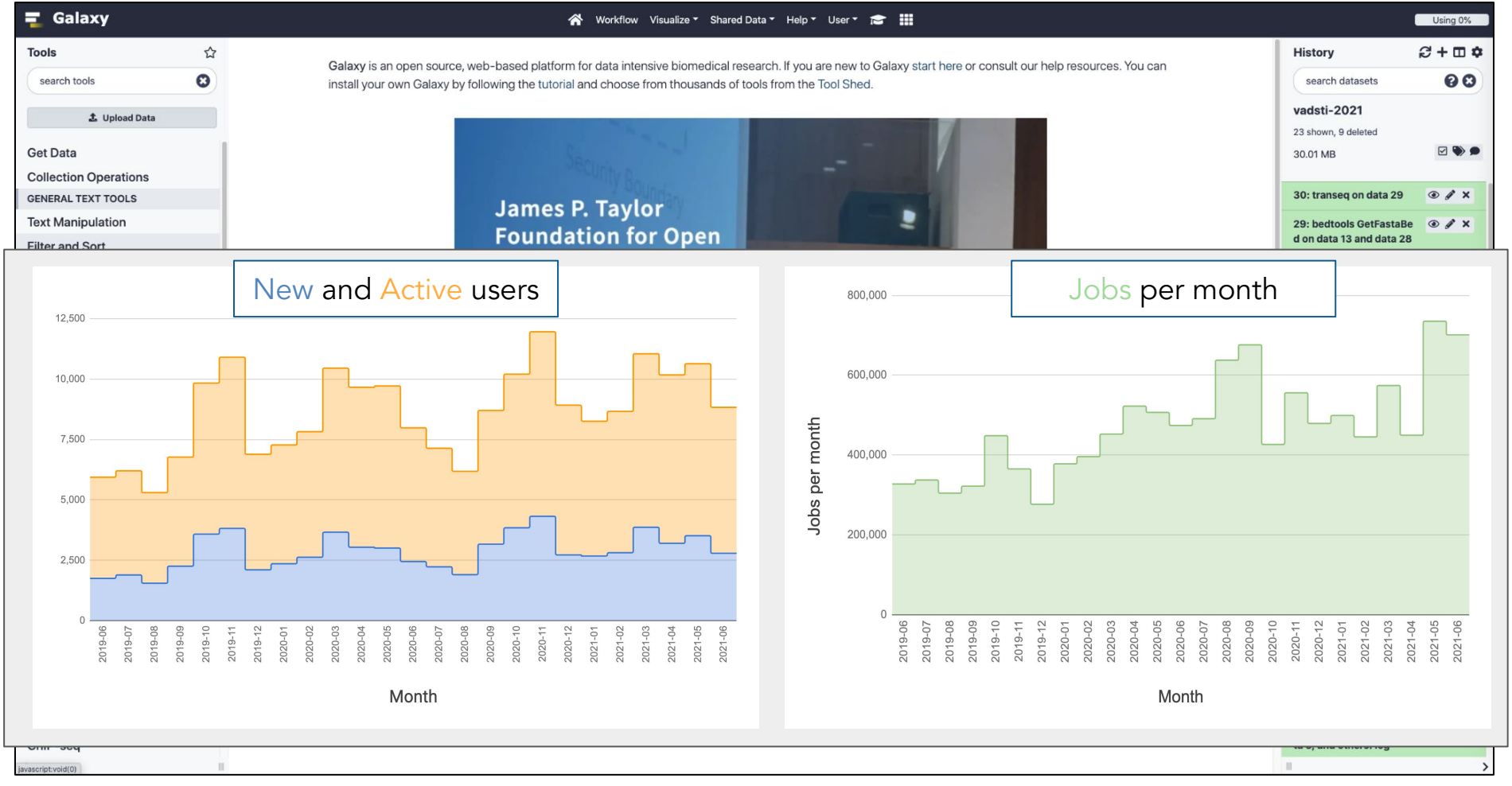
- Interactive analyses (e.g. RStudio or Jupyter notebooks) tend to be very inexpensive (<$1/hr) and will feel very familiar to desktop counterparts
- Batch analyses (e.g. WDL/CWL/Galaxy Workflows) vary enormously in computing costs from <<$1 to >>$10k
- Your spend rate is primarily determined by the #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….

https://anvilproject.org/learn/investigators/budget-templates

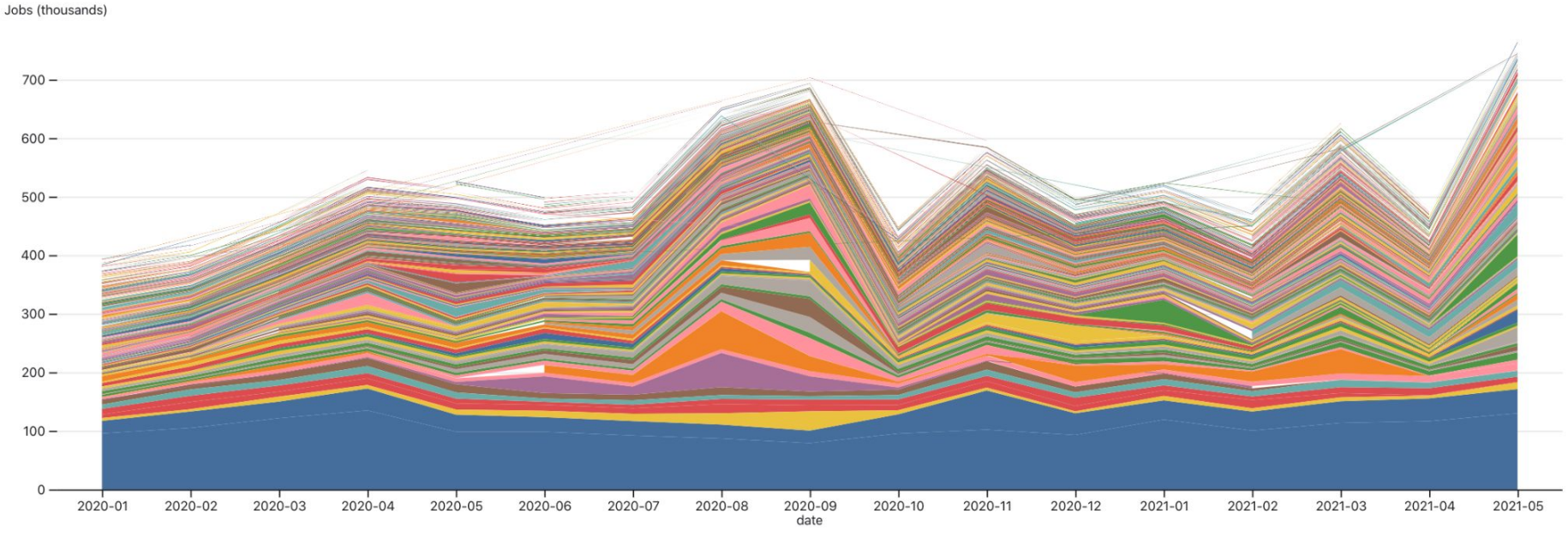
Galaxy Usage (usegalaxy.org)



Galaxy Usage (usegalaxy.org)



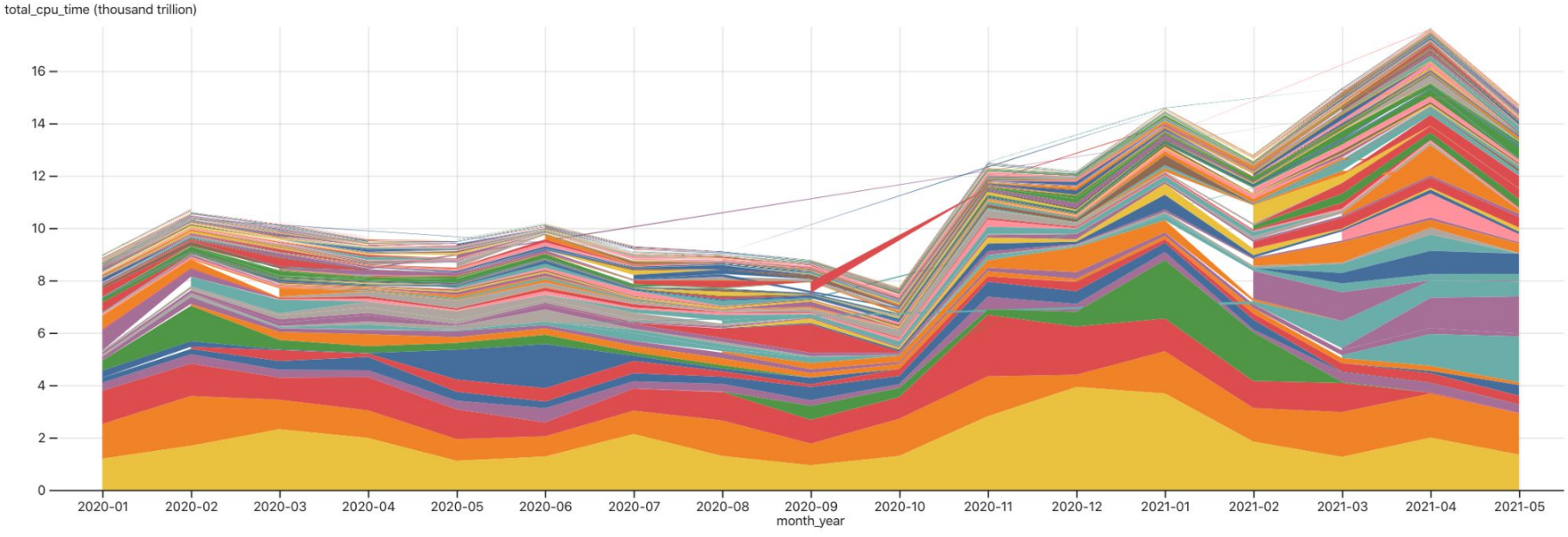
Phase I: Jobs per month, by tool



- Upload (galaxy internal)
- fastqc
- bwa_mem
- bowtie2
- hisat2
- fasterq_dump
- fastp
- picard markDuplicates
- abricate

https://github.com/galaxyproject/usage-metering

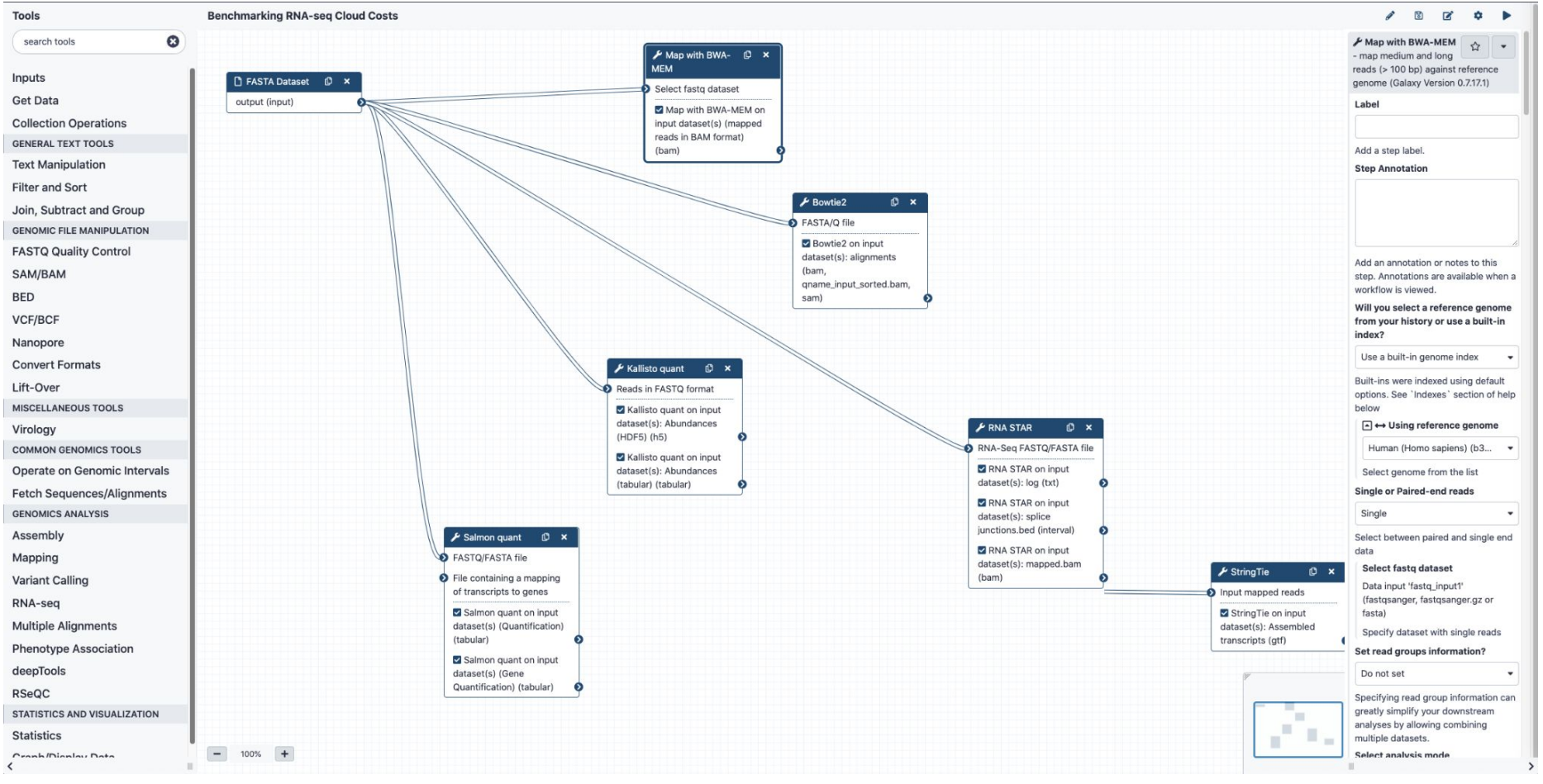
Phase I: Total CPU time per month, by tool



- tophat2
- bwa_mem
- bowtie2
- rna_star
- bwa
- hisat2
- freebayes
- htseq_compare
- lofreq_call

https://github.com/galaxyproject/usage-metering

Phase II: Benchmarking Popular Tools



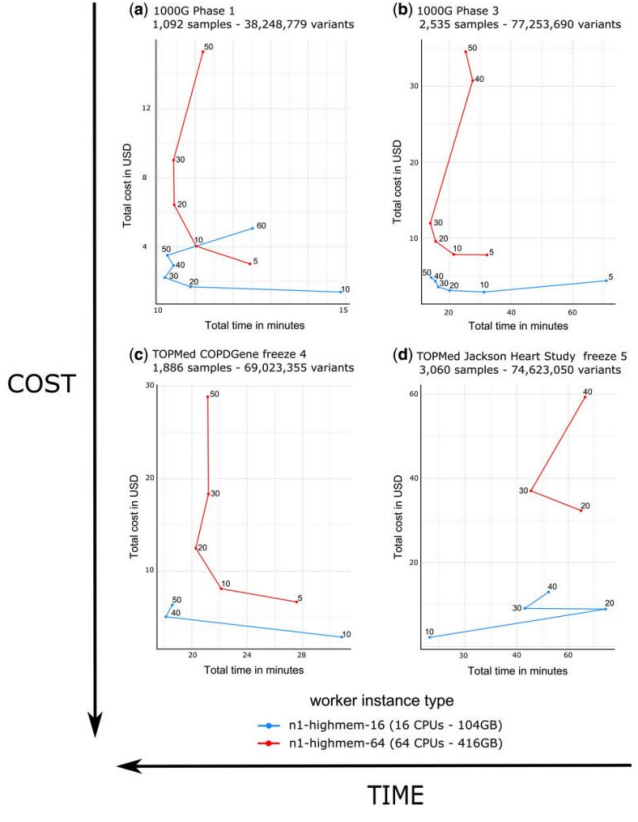
Phase III: Modeling performance

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 expansive)

Deployment Strategy

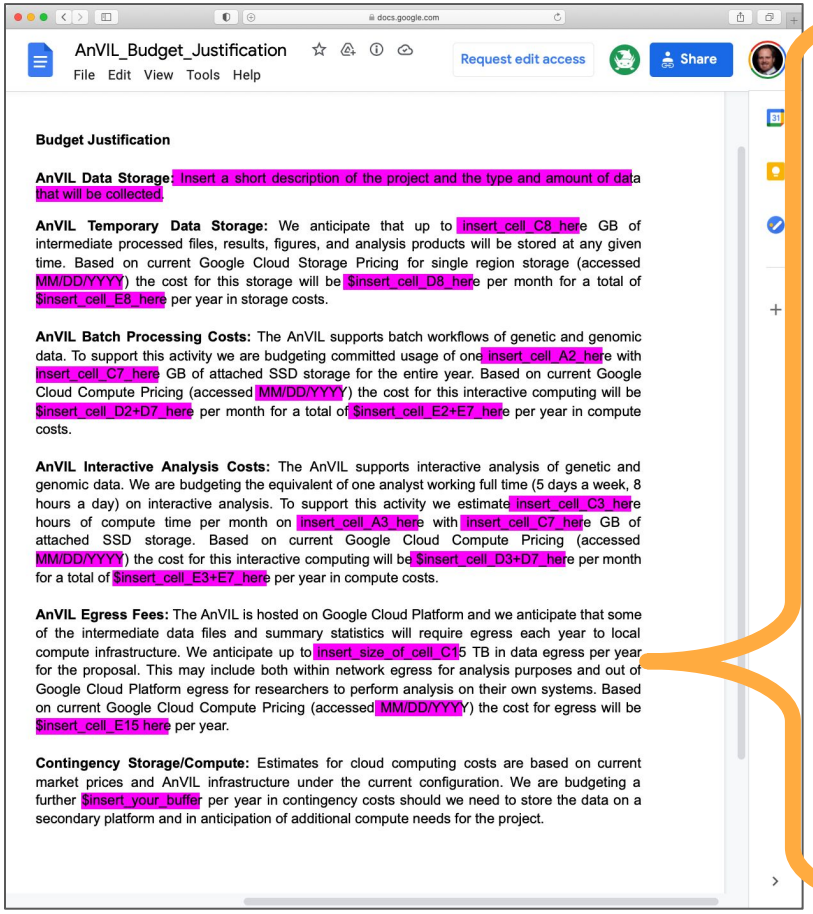
- Generate a tool-based lookup table with a range of inputs available
- Implement an API service to query for results
- Expand the range of tools covered by developing a predictive model



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



AnVIL_Budget_Justification

File Edit View Tools Help

Request edit access | Share

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_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.

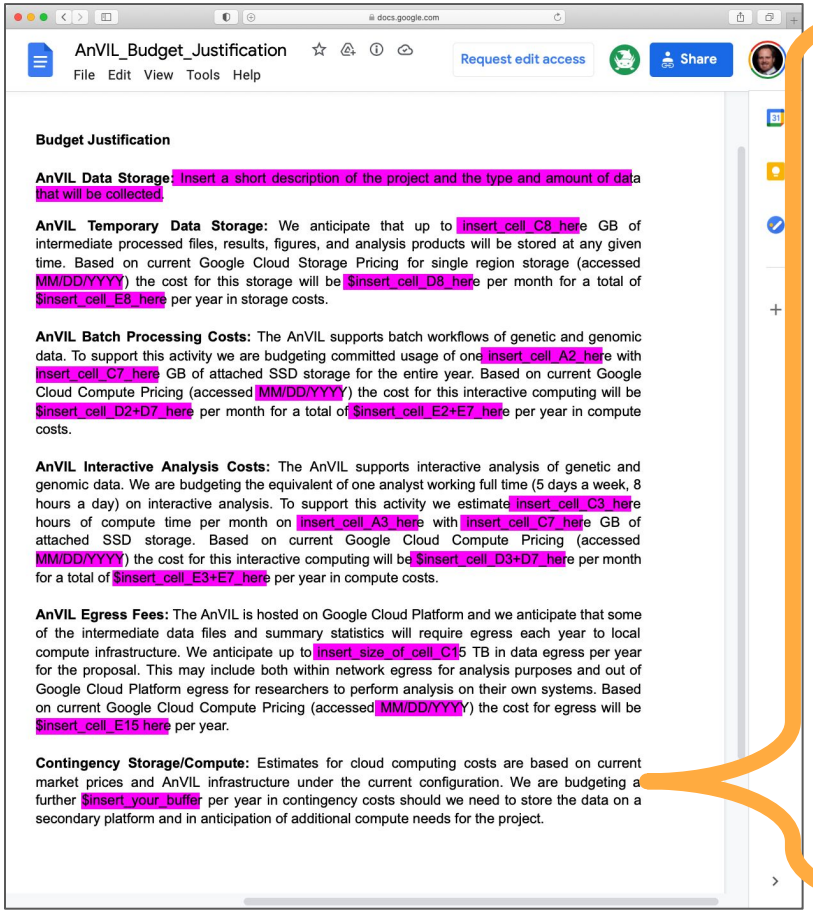
Contingency Storage/Compute: Estimates for cloud computing costs are based on current market prices and AnVIL infrastructure under the current configuration. We are budgeting a further $insert_your_buffer per year in contingency costs should we need to store the data on a secondary platform and in anticipation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

https://anvilproject.org/learn/investigators/budget-templates

AnVIL Cloud Cost Budget Templates



AnVIL_Budget_Justification

File Edit View Tools Help

Request edit access | Share

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.

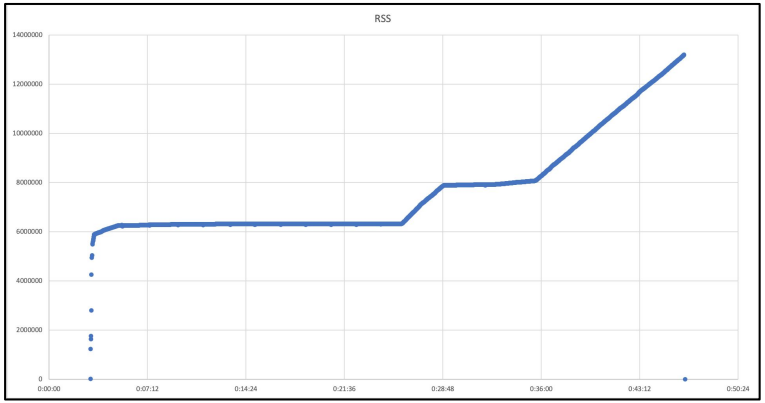
AnVIL Interactive Analysis Costs: The AnVIL supports interactive analysis of genetic and genomic data. We are budgeting the equivalent of one analyst working full time (5 days a week, 8 hours a day) on interactive analysis. To support this activity we estimate $insert_cell_C3_here hours of compute time 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_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 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



https://anvilproject.org/learn/investigators/budget-templates

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.

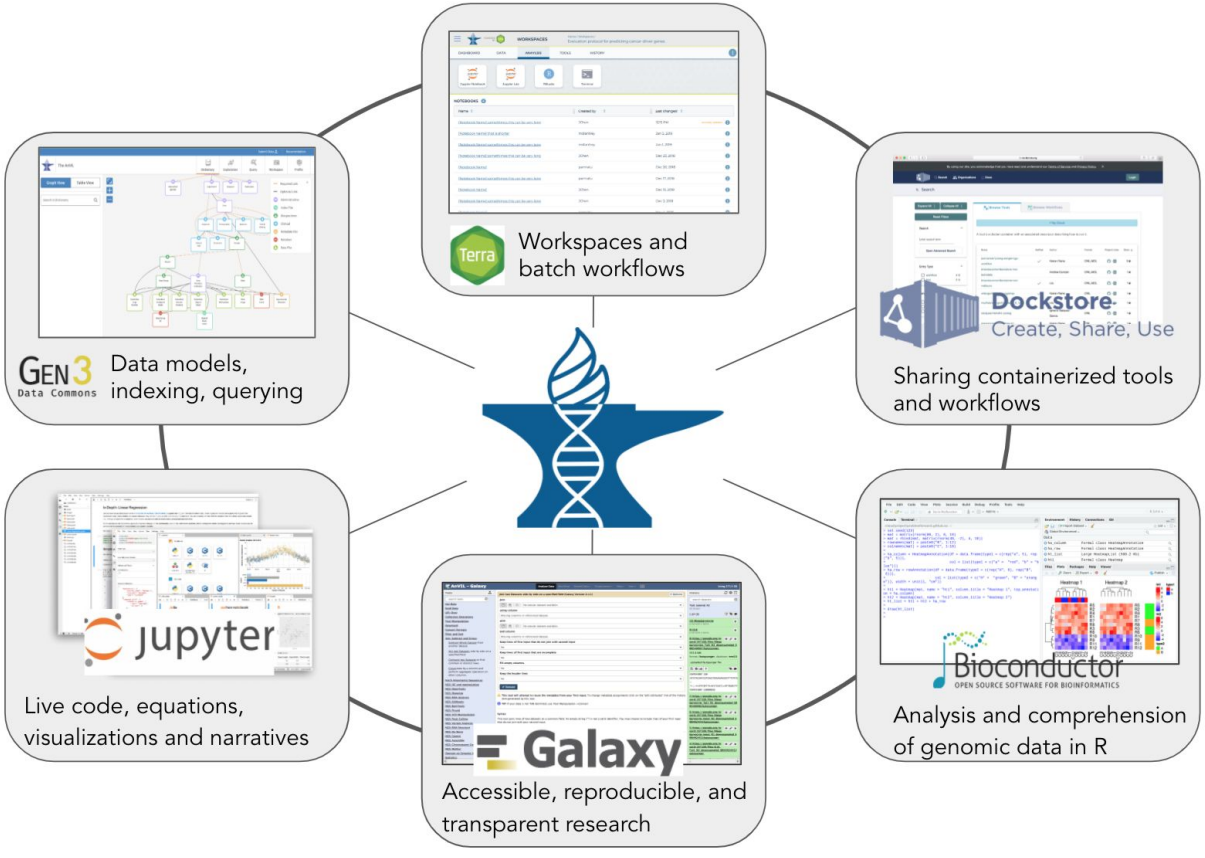
Acknowledgements

Galaxy Community
Enis Afgan
Keith Suderman
Dannon Baker
Sergey Golitsynskiy
Bridget Carr
Victor Wen
Peiyuan Xu

AnVIL Team
Jeff Leek
Fred Tan
Stephen Mosher
Sarah Wheelan
Ava Hoffman
David Rogers

National Human Genome Research Institute

National Institutes of Health
Office of Data Science Strategy



Thank you!

@useAnVIL