Customized for **NCBI Staff in collaboration with the bioinformatics community.**

Federated Data Access

In Four Acts.

Scenes

Act 1: Virus Characterization and Discovery (x2!)
Act 2: Genome Graphs
Act 3: Annotation of Haplotypes (and Graphs)
Act 4: Indexing Data for Federated Discovery on any Platform, Anywhere in the World
Act 5: Epilogue: Metadata Matters

Disclosures

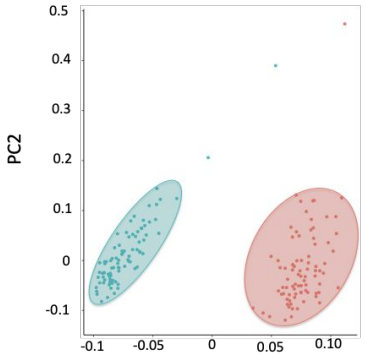
The idea and material presented here is not necessarily the view of NCBI, NLM, NIH or any other federal agency. Some of the underlying work was supported by the Intramural Research Program of the NLM and various other federal agencies.

Ben Busby is contracted to NCBI through Medical Science and Computing (MSC).

Ben Busby is also an advisor to:

- Johns Hopkins
- Ariel Precision Medicine
- Deloitte

through Mountain Genomics, a company headquartered in Pittsburgh, PA.



Overview

As the volume of publicly available genomic data expands, it is becoming increasingly clear the SRA and other public repositories play an evermore important role of being stewards, allowing researchers to leverage the statistical and discovery power represented in huge arrays of datasets. That said, these repositories must not simply become 'bags of data', but indexed repositories where data can be found, approximately assessed for quality, mixed with other data, and most importantly, used by researchers to ask fundamental biological and biomedical questions.

Nearly as important, we must provision reproducible workflows, such that investigators can go from finding data to answering questions as simply and quickly as possible. These workflows will range in complexity from basic blast searching, variant calling and annotation, and transcript counting to building genome graphs and discovering novel viruses and back-spliced RNA. While doing all this we must be cognisant that while building a simple GUI interface for these types of analysis is impractical and likely impossible, the onus is on us to teach fledgeling biological data scientists to apply computational tools to their existing biological questions.

Federation of Sequencing Data in the Cloud:

Four Illustrative Topic Areas

1. Virus Characterization and Discovery.

   Generation of a cloud-based indexing system to allow investigators to identify data sets of interest based on taxonomic, gene and protein domain profiles

2. Genome Graphs

   Generation of simple, usable systems to not compare an individual patient or organism to another -- or small group of -- individual, but to an entire community, dramatically compress data, and immediately find "toxic paths".

3. Annotation of Haplotypes (and Graphs)

   Annotation of haplotypes, instead of graphs, to allowing investigators to query complex disease

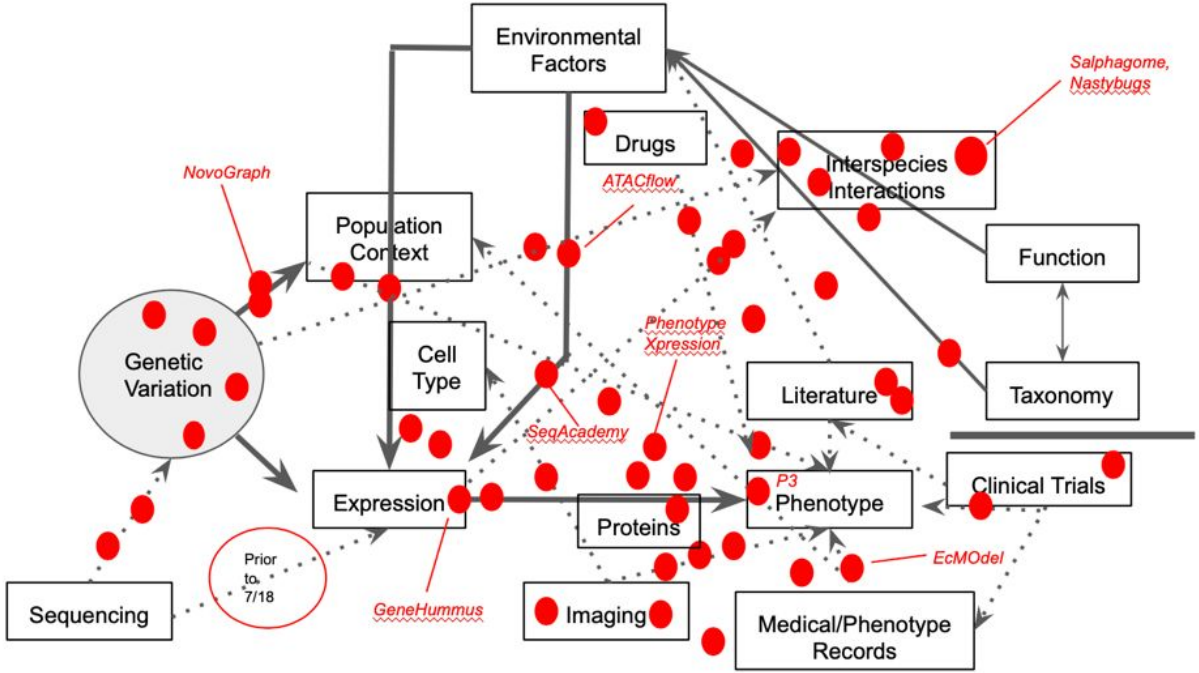
4. Indexing Data for Federated Discovery on any Platform, Anywhere in the World

   Flexible presentation (API) of viral protein domains, host-pathogen interactions and eventually graph loops as proof-of-principle data federation.

Previous Hackathons (commits 7/1/2018 - 1/1/19 only)

Human, Plant and Animal Genomes

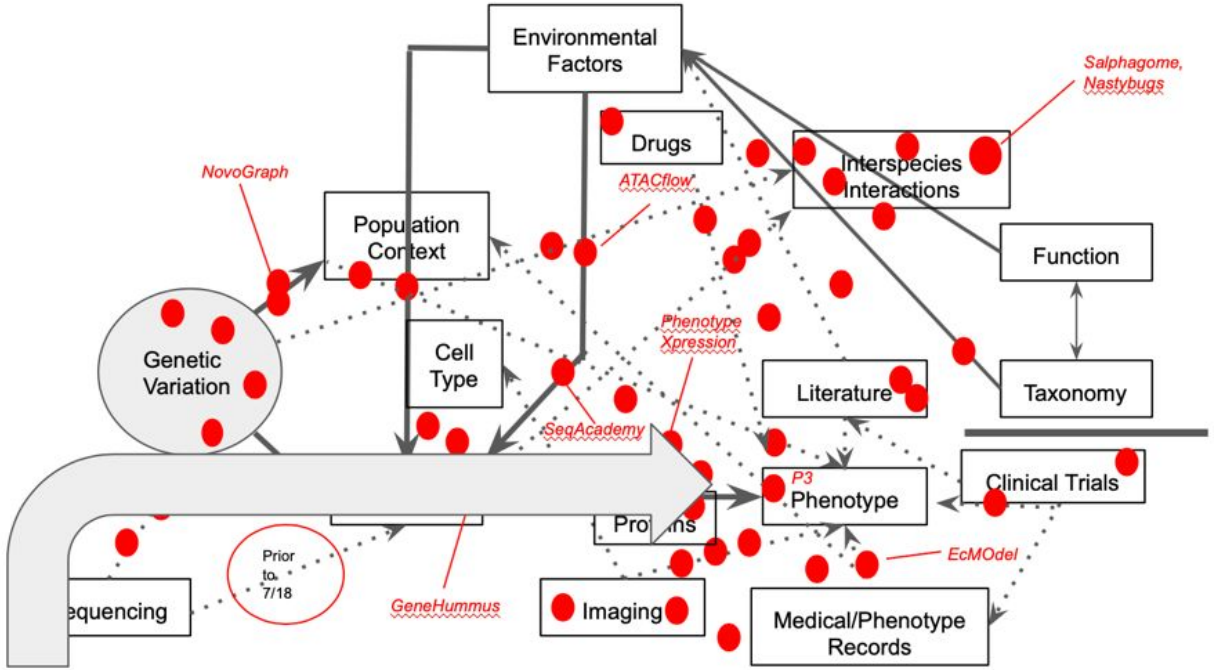
Bacterial, Viral, Fungus and Other Genomes

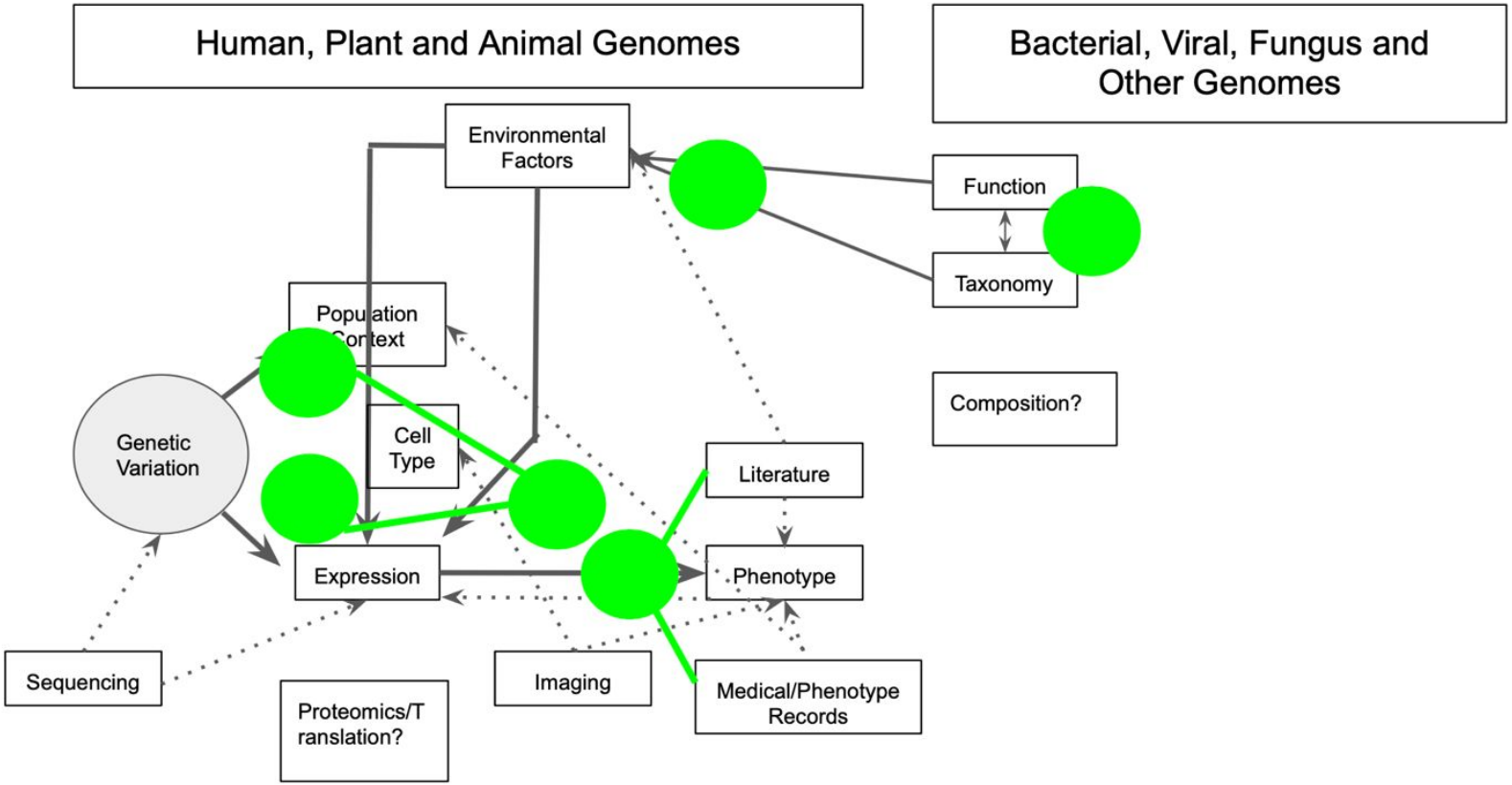


Previous Hackathons (commits 7/1/2018 - 1/1/19 only)

Human, Plant and Animal Genomes

Bacterial, Viral, Fungus and Other Genomes

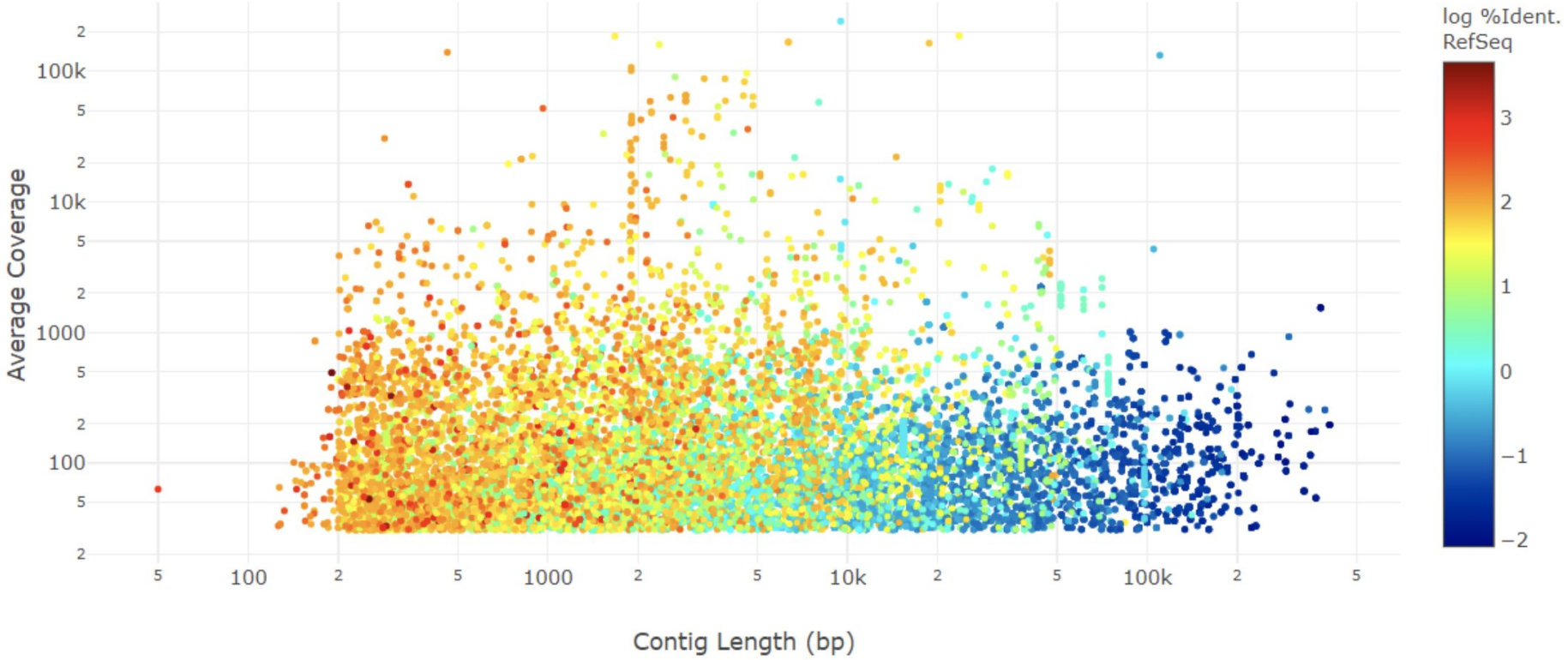


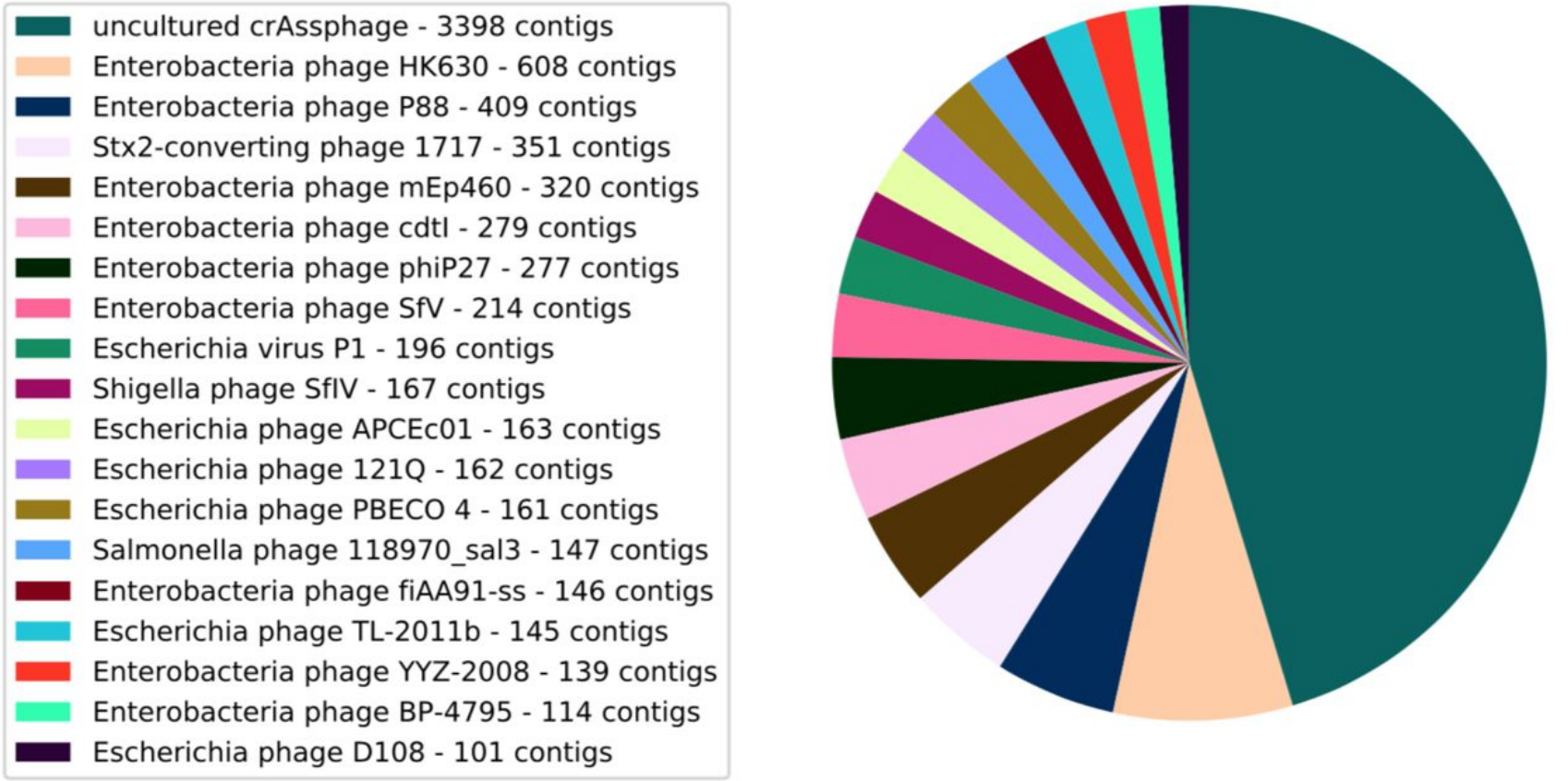

Human, Plant and Animal Genomes
Bacterial, Viral, Fungus and Other Genomes
Environmental Factors
Function
Taxonomy
Population Context
Genetic Variation
Cell Type
Composition?
Literature
Expression
Phenotype
Sequencing
Proteomics/Translation?
Imaging
Medical/Phenotype Records

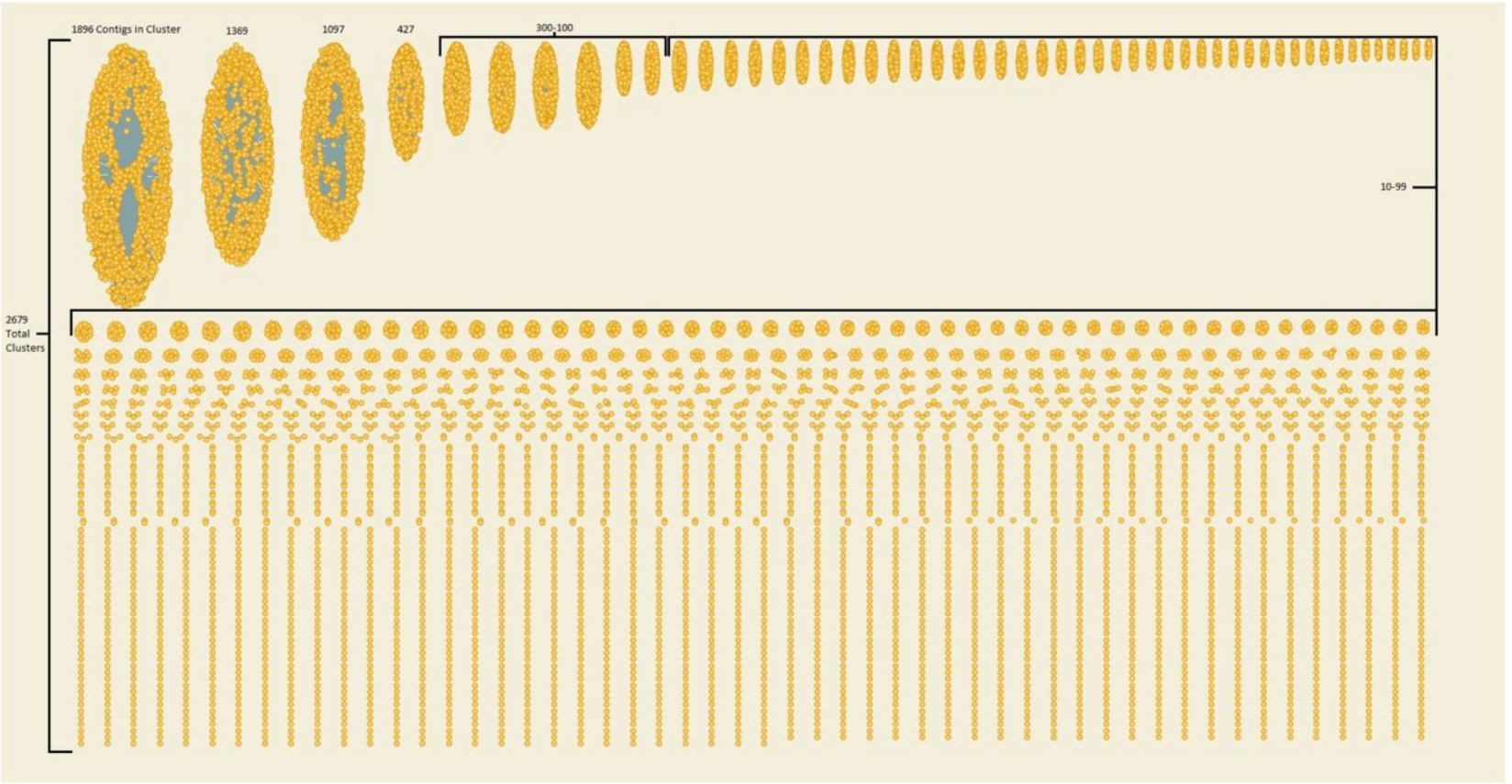
Virus Characterization and Discovery

Powered by:

We have extracted known viruses, heretofore unknown family members, sequences identifiable as viral by the protein domains that decorate their proteins, and the true genomic “dark matter” and built a prototype index for viral signatures across the vast metadata space within SRA (~90,000-150,000 datasets). This is an example of how a researcher can reach across a massive data repository with a very simple bioinformatic tool and extract dramatic results, except in this case, we have pre-built this index for researchers working on almost any kind of virus, including noroviruses, hemorragic fever-causing viruses, and bacteriophages, the most abundant organisms on earth.

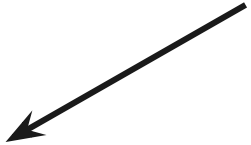

Average Coverage
2
100k
5
2
10k
5
2
1000
5
2
100
5
2
100
1000
10k
100k
Contig Length (bp)
log %Ident.
RefSeq
3
2
1
0
−1
−2

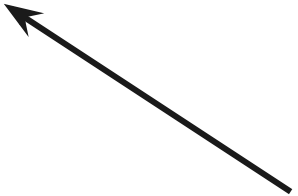

uncultured crAssphage - 3398 contigs
Enterobacteria phage HK630 - 608 contigs
Enterobacteria phage P88 - 409 contigs
Stx2-converting phage 1717 - 351 contigs
Enterobacteria phage mEp460 - 320 contigs
Enterobacteria phage cdtI - 279 contigs
Enterobacteria phage phiP27 - 277 contigs
Enterobacteria phage SfV - 214 contigs
Escherichia virus P1 - 196 contigs
Shigella phage SfIV - 167 contigs
Escherichia phage APCEc01 - 163 contigs
Escherichia phage 121Q - 162 contigs
Escherichia phage PBECO 4 - 161 contigs
Salmonella phage 118970_sal3 - 147 contigs
Enterobacteria phage fiAA91-ss - 146 contigs
Escherichia phage TL-2011b - 145 contigs
Enterobacteria phage YYZ-2008 - 139 contigs
Enterobacteria phage BP-4795 - 114 contigs
Escherichia phage D108 - 101 contigs

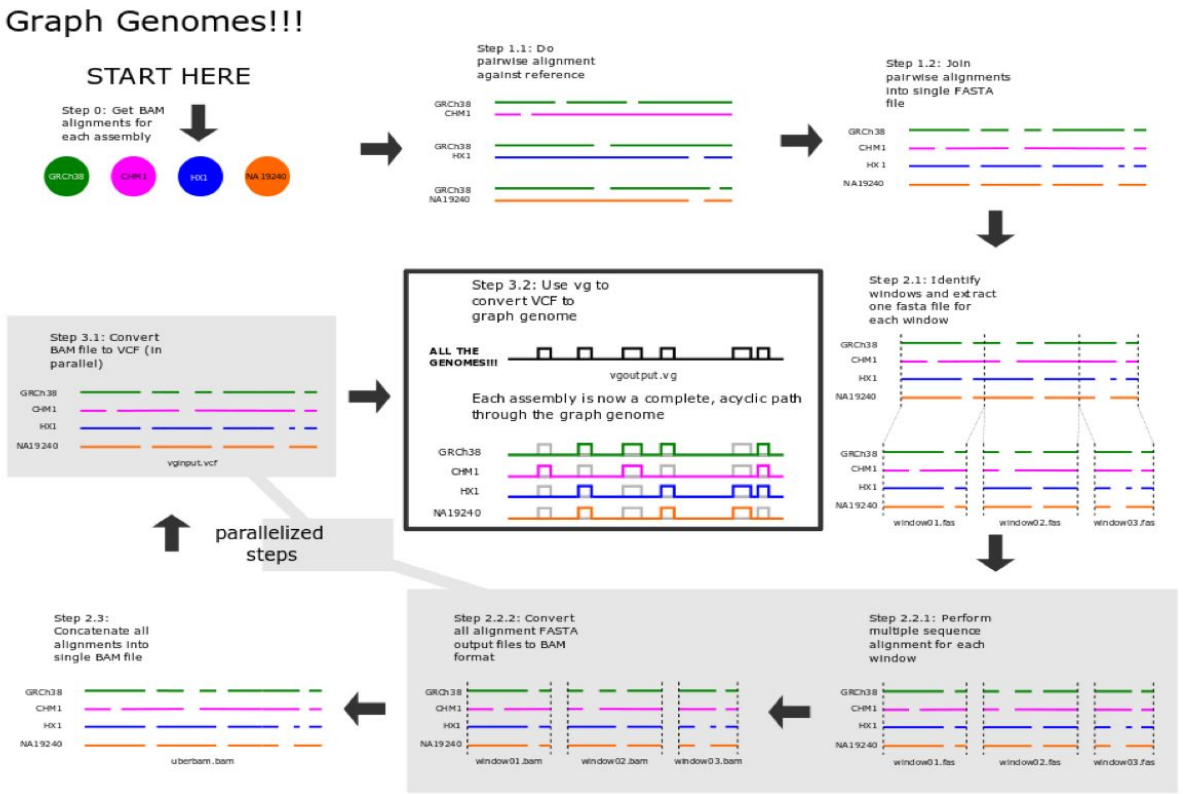

1896 Contigs in Cluster
1369
1097
427
300-100
10-99
2679
Total
Clusters

Genome Graphs

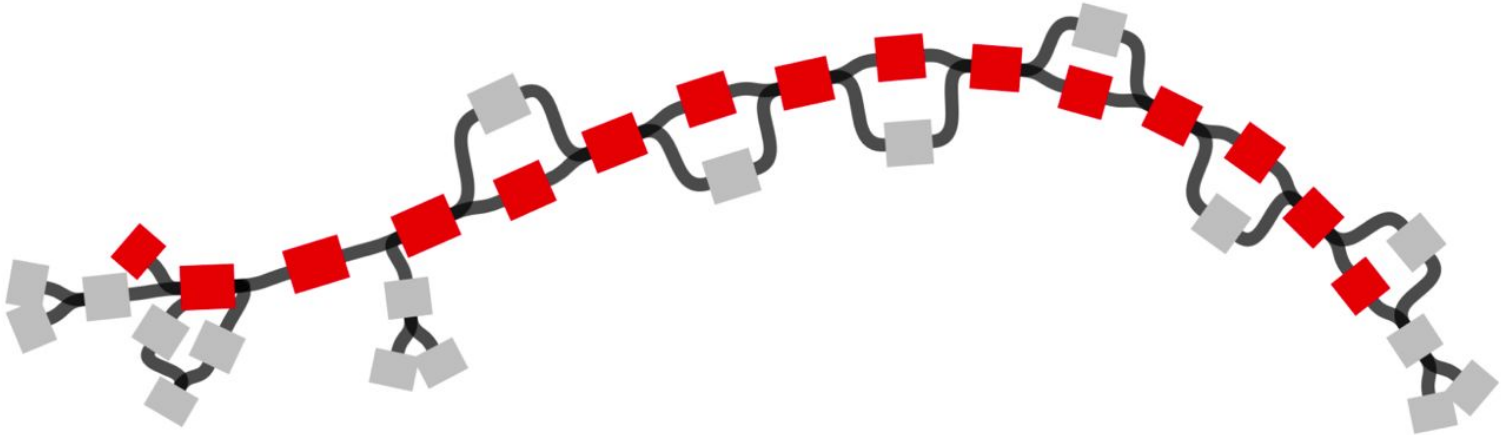
Initially leveraging standard workflows for annotation of primary datasets in cloud infrastructure mapped to linear genomes, we will likely transition to genome graphs over the next ten years. This will allow us to compress reads, do phasing automatically, do reference-guided assembly on complex genomes, and detect “toxic paths” automatically. Most importantly, this will allow us to see complex genotype - phenotype interactions easily.







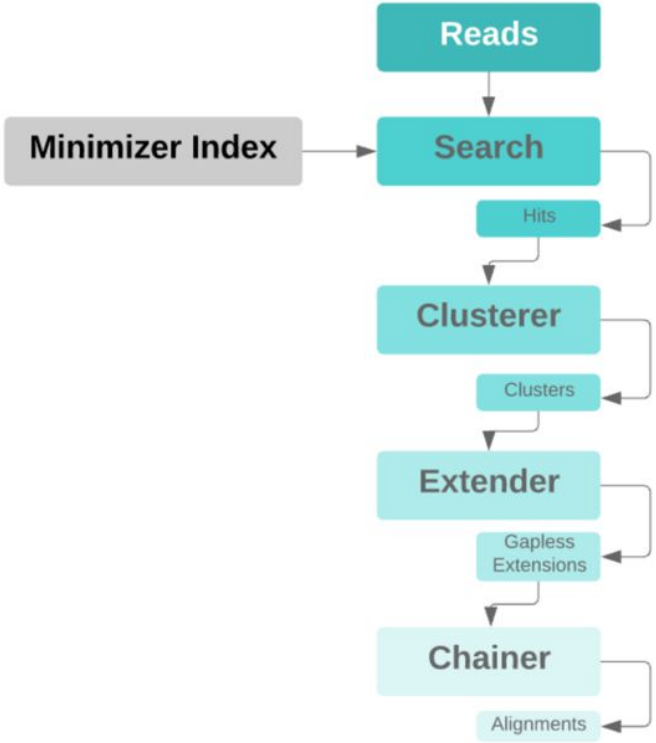
Evan Biederstedt, Alex Dilthey
many others



The path of GRCh38 through a graph! For details check the DS folder!

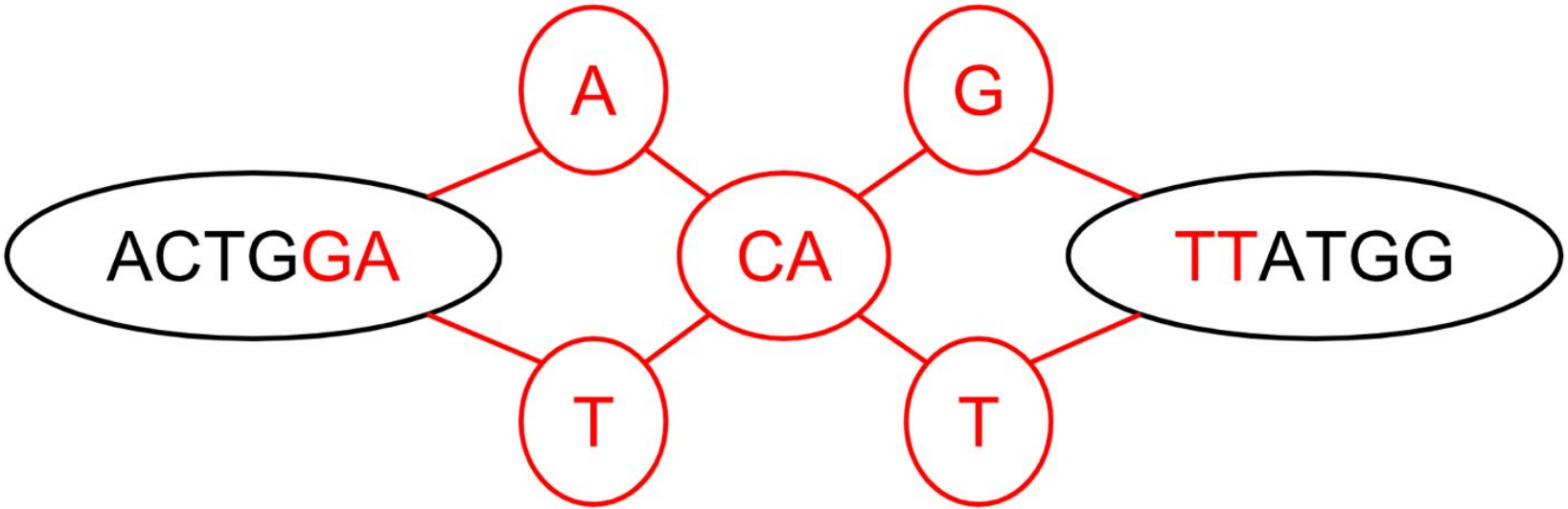
https://github.com/NCBI-Hackathons/TheHumanPangenome/tree/master/DS

Erik Garrison, many others



A Faster, Better Short-Read Mapper with Hit Chaining. Now with more corn! See the Giraffe folder for details!

https://github.com/NCBI-Hackathons/TheHumanPangenome/tree/master/Giraffe

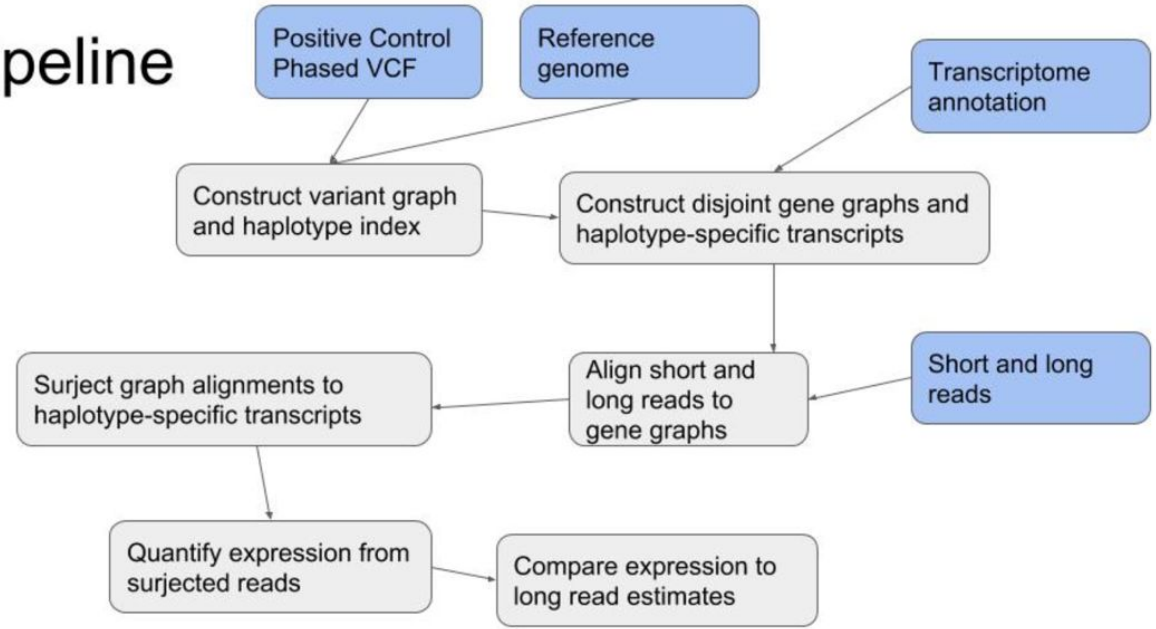


Most Genome Annotations can be Imported from gff3 (to ggff!) quite easily! See the annotation folder for details!

https://github.com/NCBI-Hackathons/TheHumanPangenome/tree/master/annotation

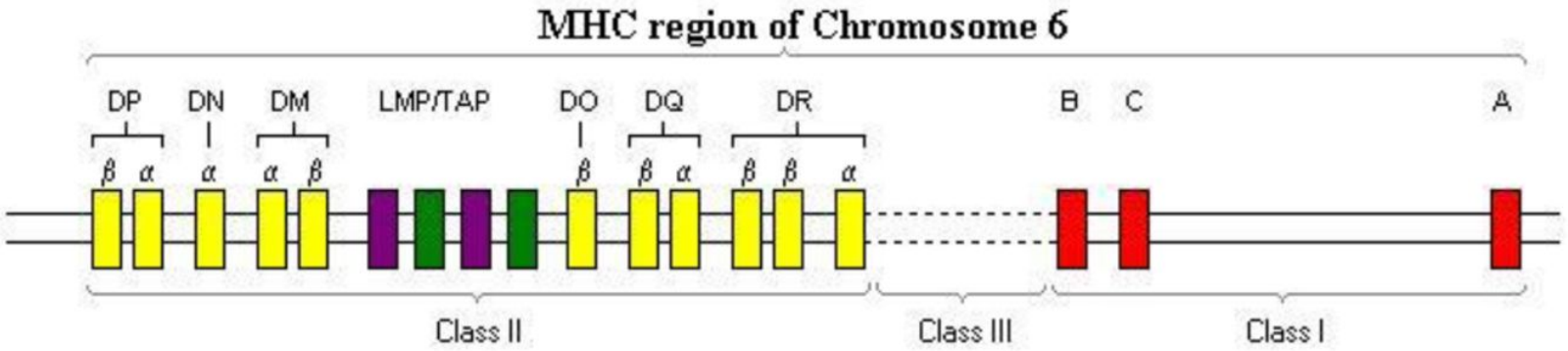
Jordan Izenga, many others

Pipeline



Allele-Specific Expression can be calculated in a lightweight fashion! Grab the workflow in the RNA folder!

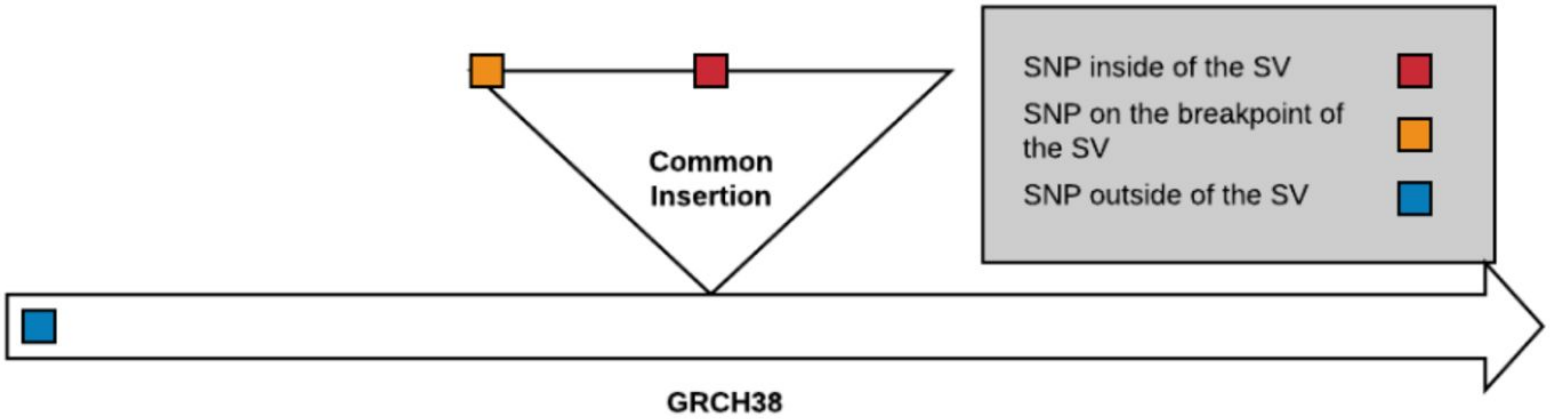
https://github.com/NCBI-Hackathons/TheHumanPangenome/tree/master/RNA



We Mapped the MHC Region of HG002 to a Graph with Long Reads, Long Reads, 10X and love. Check out the MHC folder!

https://github.com/NCBI-Hackathons/TheHumanPangenome/tree/master/MHC

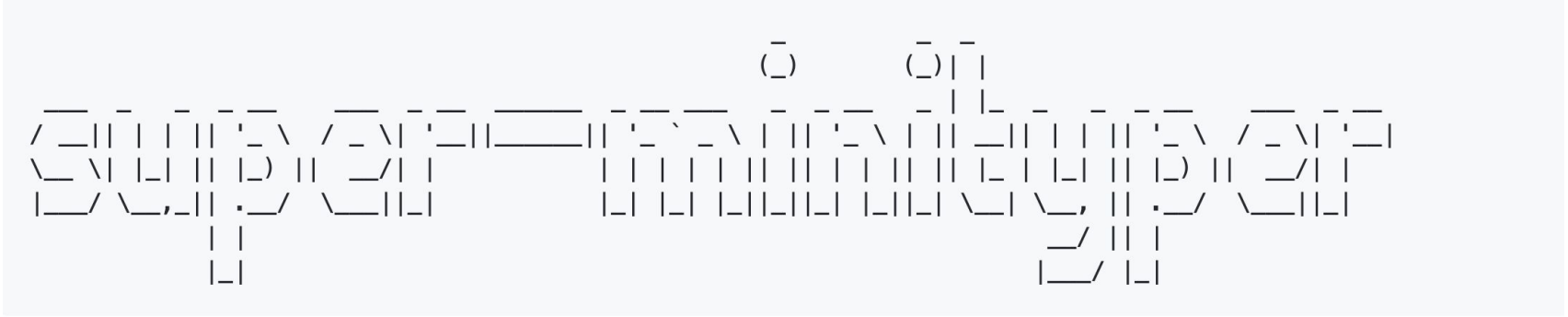
Justin Zook, many others



Calculate Mutations In, Outside of, and at Breakpoints of Structural Variants! See the SV folder for Details!

https://github.com/NCBI-Hackathons/TheHumanPangenome/tree/master/SV/HG002

Fritz Sedlazeck, many others

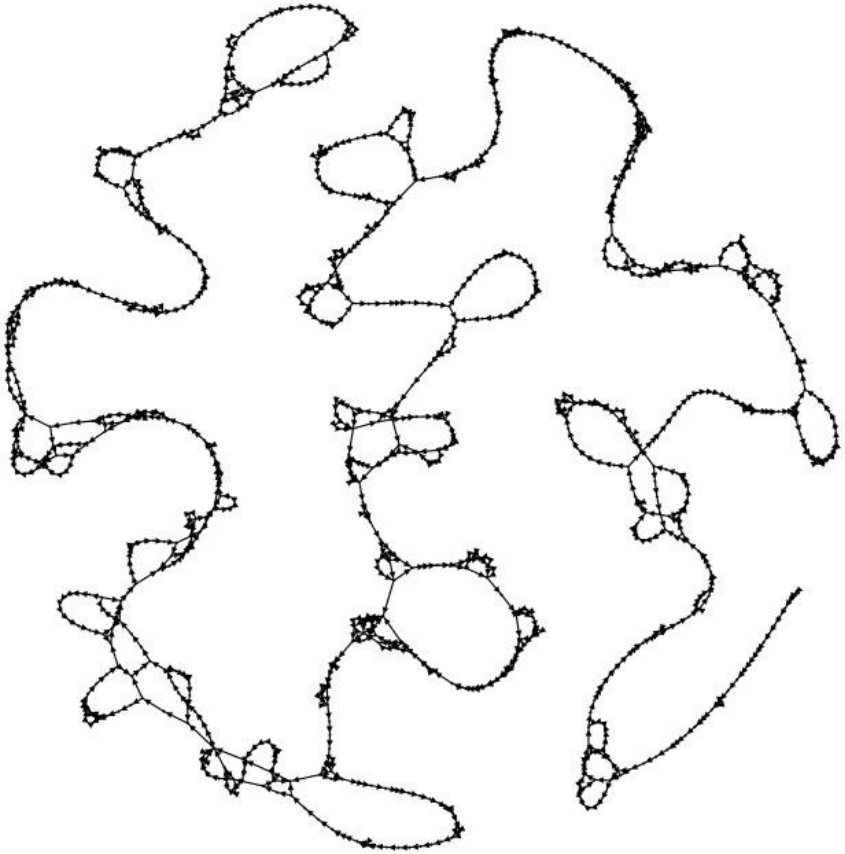


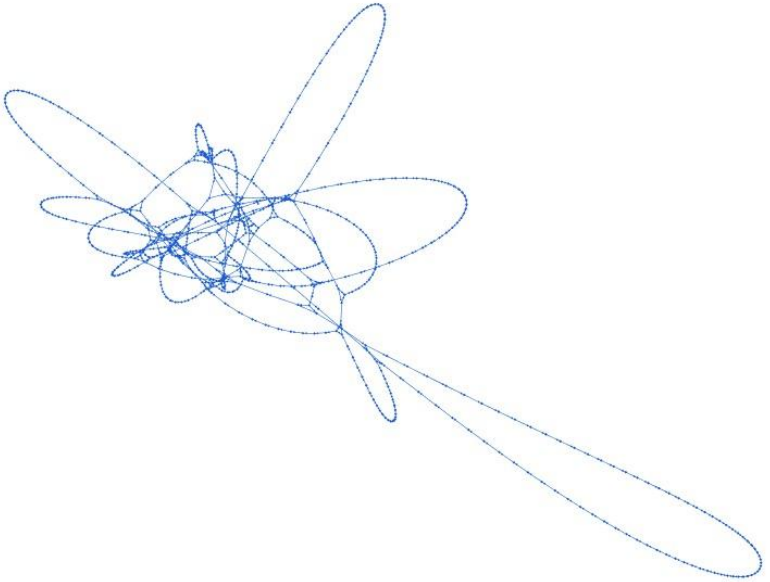
`super-minityper` is a set of cloud-based workflows for constructing SV graphs and mapping reads to them.

Structural variants frustrate read mapping because aligners often choose not to map read portions which map very distantly. Graphs allow incorporating known variants, including large ones, and then mapping directly to these. While this has been shown to reduce reference bias and improve read mappings when a sample contains variants in the graph, constructing graph genomes and operating on them has historically been difficult and time-consuming.

We present a set of cloud-based workflows — composed mostly of preexisting and optimized tools — to construct graphs containing structural variants and map reads to them. These workflows allow users to take arbitrary SV calls, construct a graph, and map reads to these graphs. This workflow prioritizes ease-of-use and speed, ingesting common input formats and returning results in minutes on commodity cloud VMs.

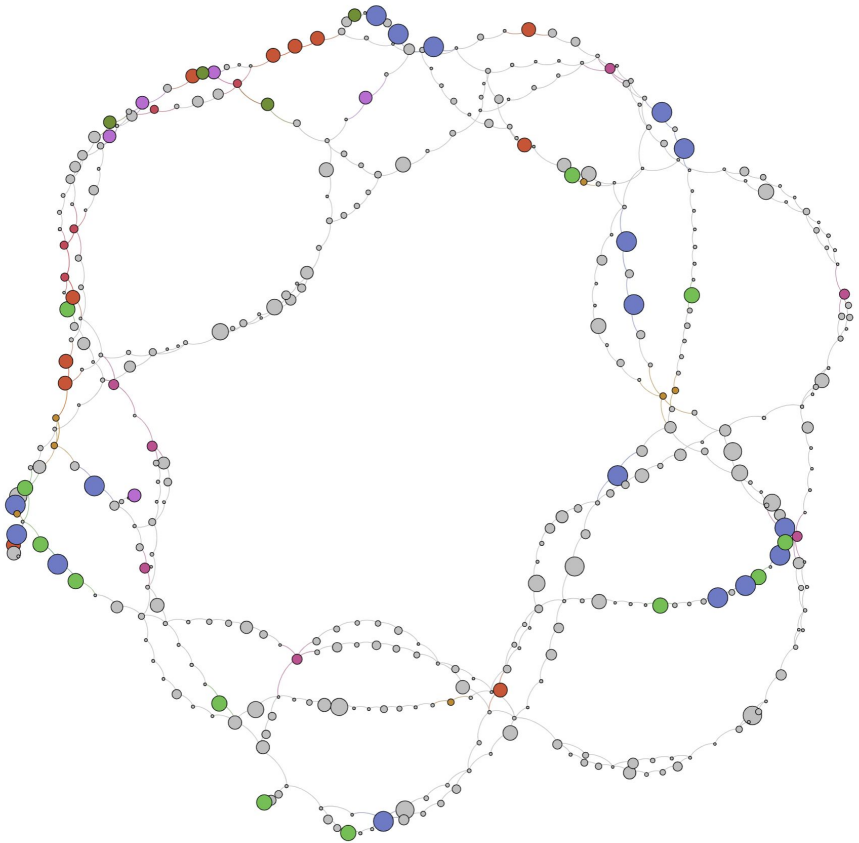
Eric Dawson, Fernanda Forterre, many others





https://github.com/NCBI-Codeathons/SWIGG

Jason Chin, Alex Gener, many others



swigg_shiny

Shiny app for visualizing genome graphs from SWIGG

In development, current priorities:

1. Interactive plot

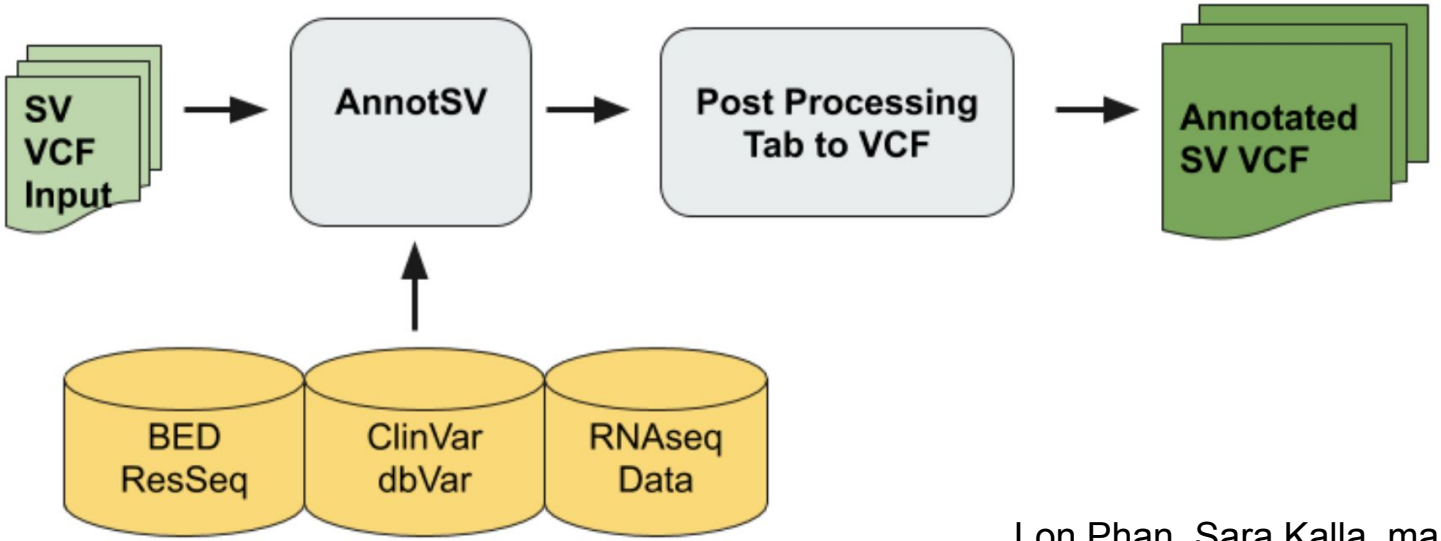
```
- Goal: View attributes of a node or edge when you hover over it in the plot

- Plan: Since plotly currently doesn't work with ggraph, try converting code from ggraph to igraph
```

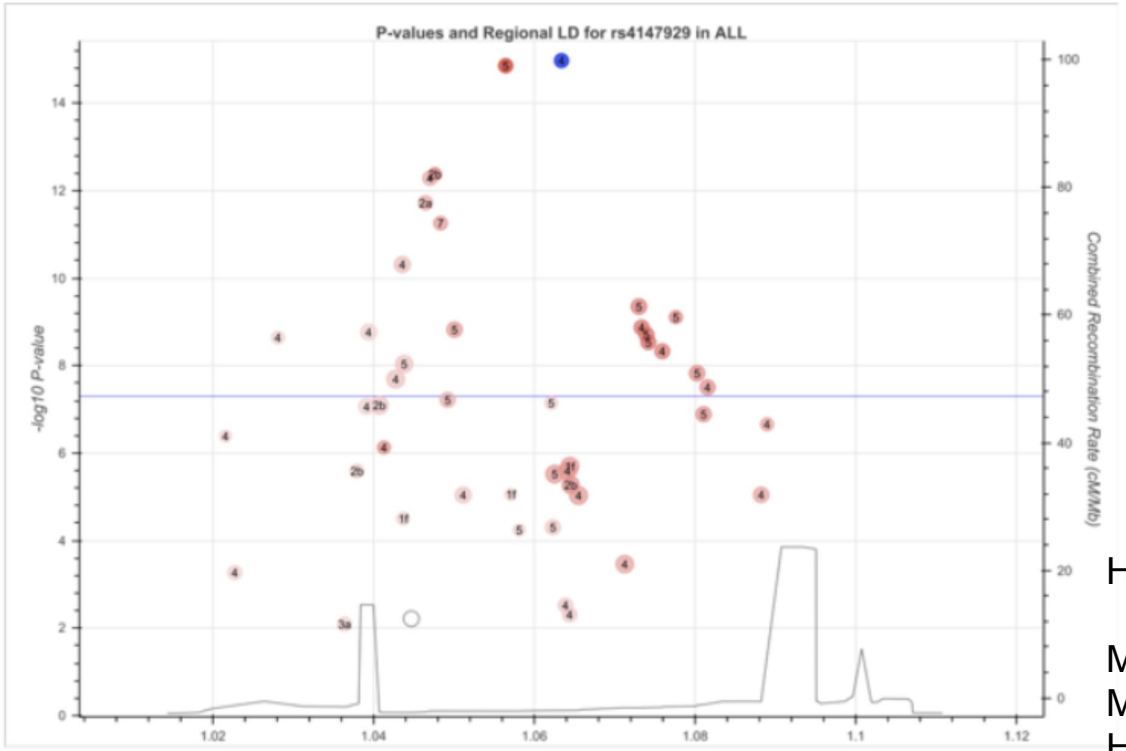
github.com/ncbi-codeathons/virus_graphs

Mike Tisza, Alexis Norris, many others

Medical Annotations of Graphs and Haplotypes

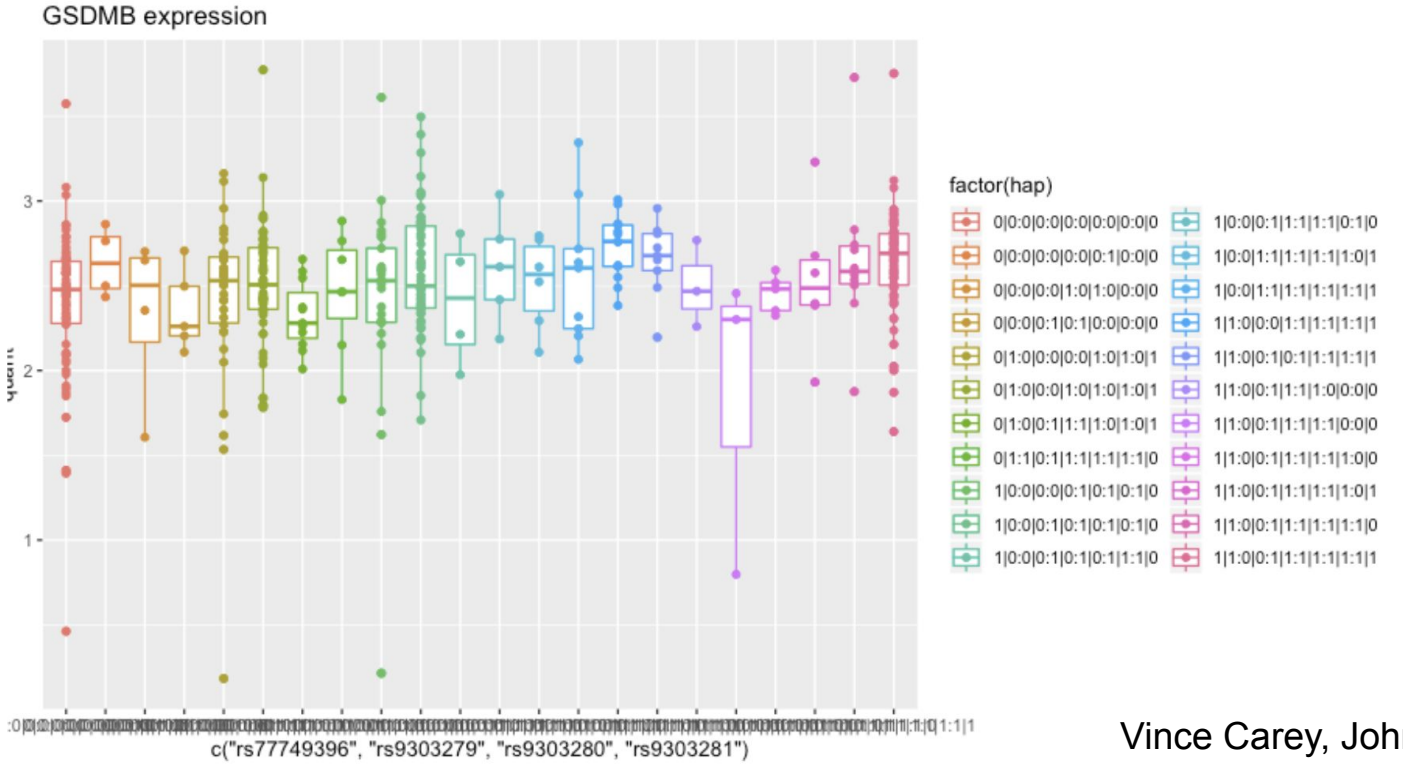


Lon Phan, Sara Kalla, many others



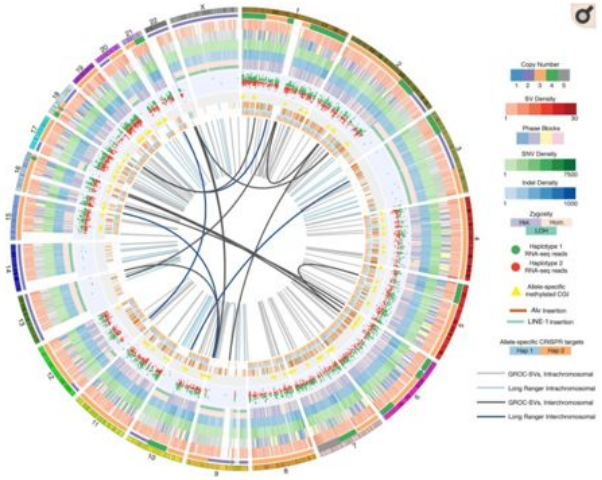
Haplocravat2

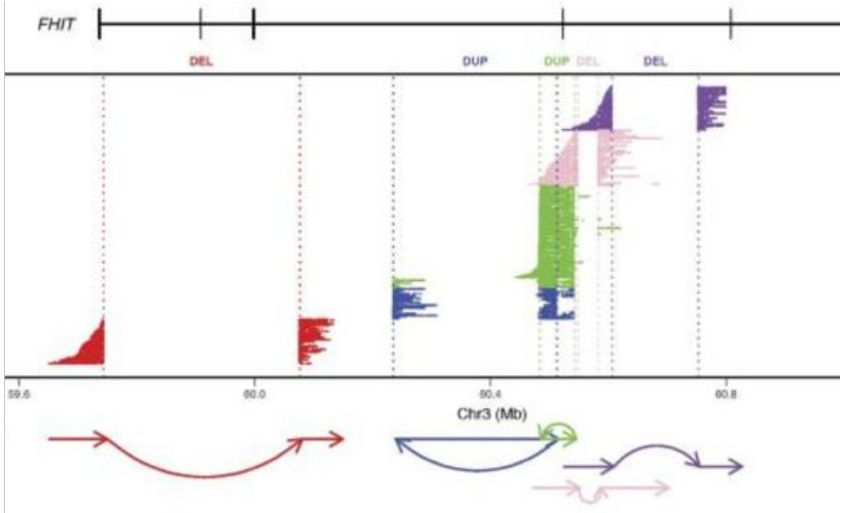
Mike Smallegan, Kyle Moad, Matt Hynes-Grace, Dina Mikdadi, many others



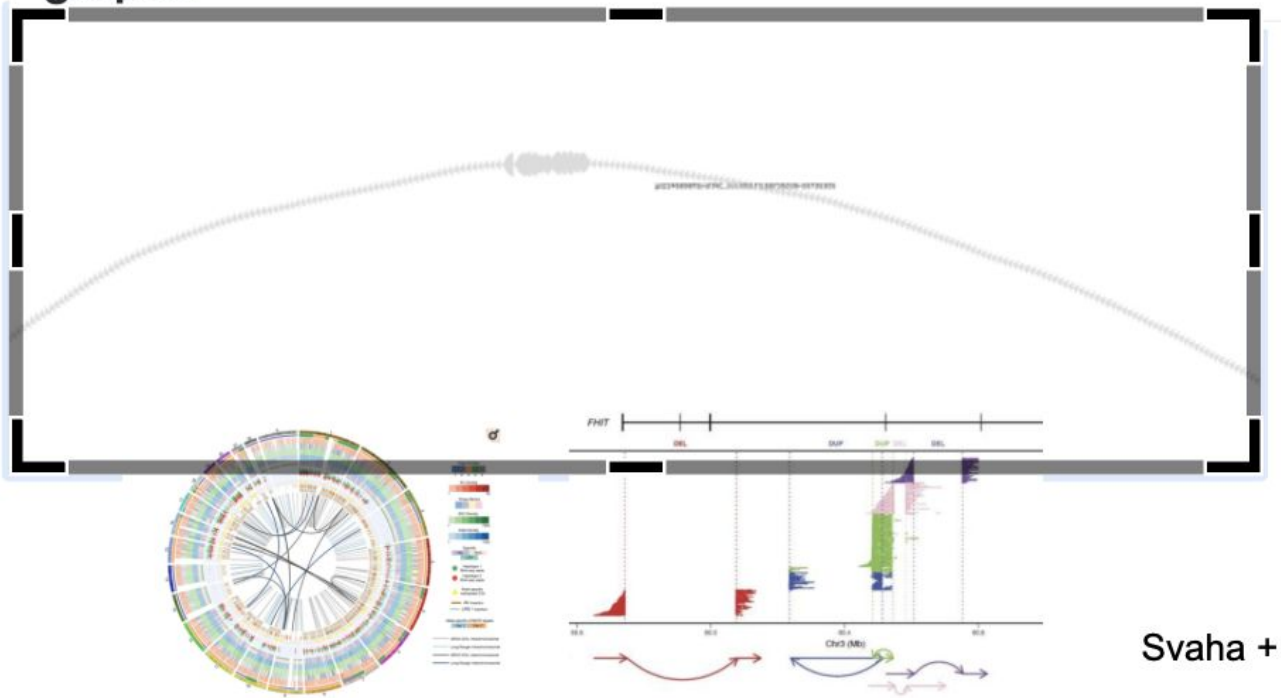
Vince Carey, John Didion, many others

Using p53 and K562 as examples to define toxic paths in graphs

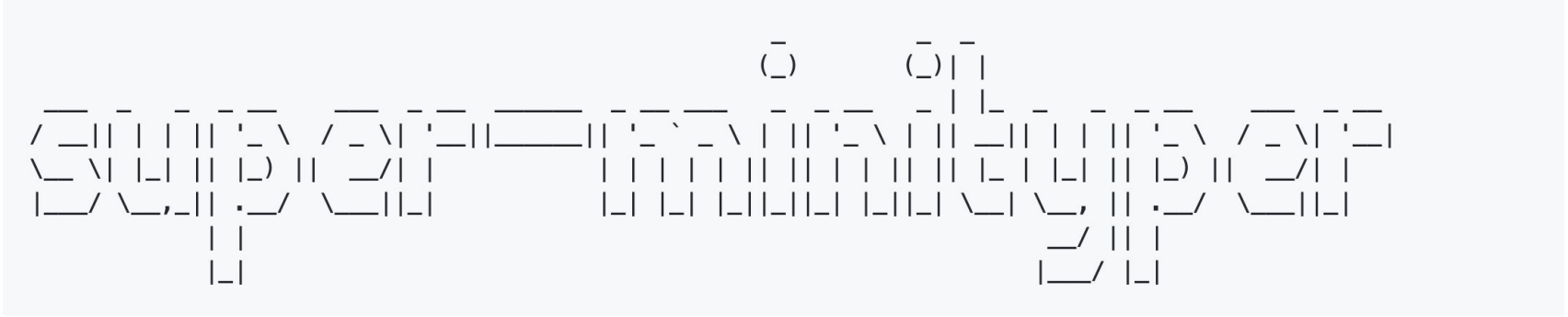




Using p53 and K562 as examples to define toxic paths in graphs



Svaha + Bandage



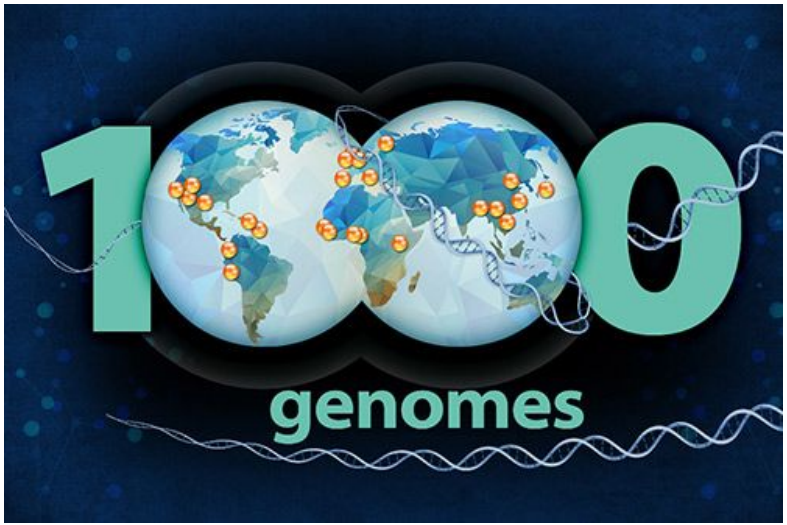
`super-minityper` is a set of cloud-based workflows for constructing SV graphs and mapping reads to them.

Structural variants frustrate read mapping because aligners often choose not to map read portions which map very distantly. Graphs allow incorporating known variants, including large ones, and then mapping directly to these. While this has been shown to reduce reference bias and improve read mappings when a sample contains variants in the graph, constructing graph genomes and operating on them has historically been difficult and time-consuming.

We present a set of cloud-based workflows — composed mostly of preexisting and optimized tools — to construct graphs containing structural variants and map reads to them. These workflows allow users to take arbitrary SV calls, construct a graph, and map reads to these graphs. This workflow prioritizes ease-of-use and speed, ingesting common input formats and returning results in minutes on commodity cloud VMs.

Eric Dawson, Fernanda Forterre, many others

A Brief Proposal for Graph-Based Sequence Compression and Phasing (personal opinion)



+

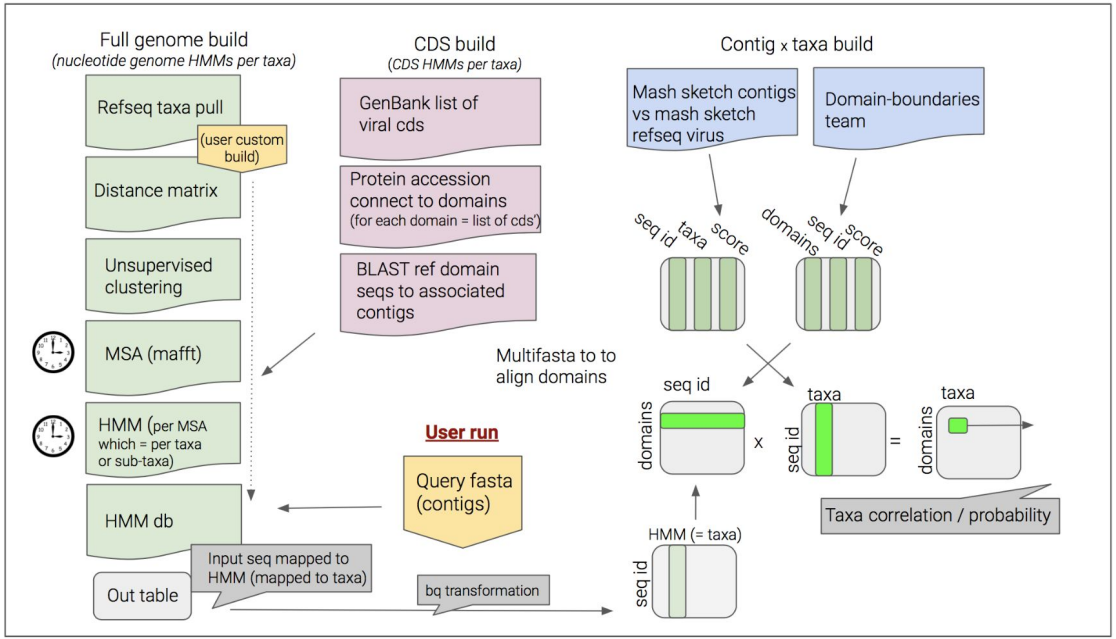
+

A Brief Proposal for Graph-Based Sequence Compression and Phasing (personal opinion)

=

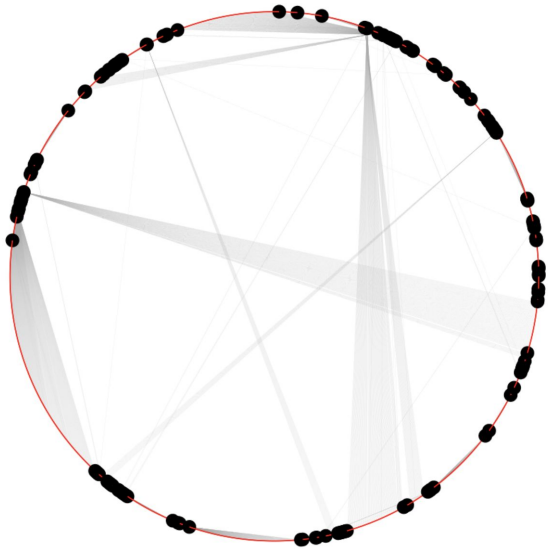
(no longer available on CD-ROM)

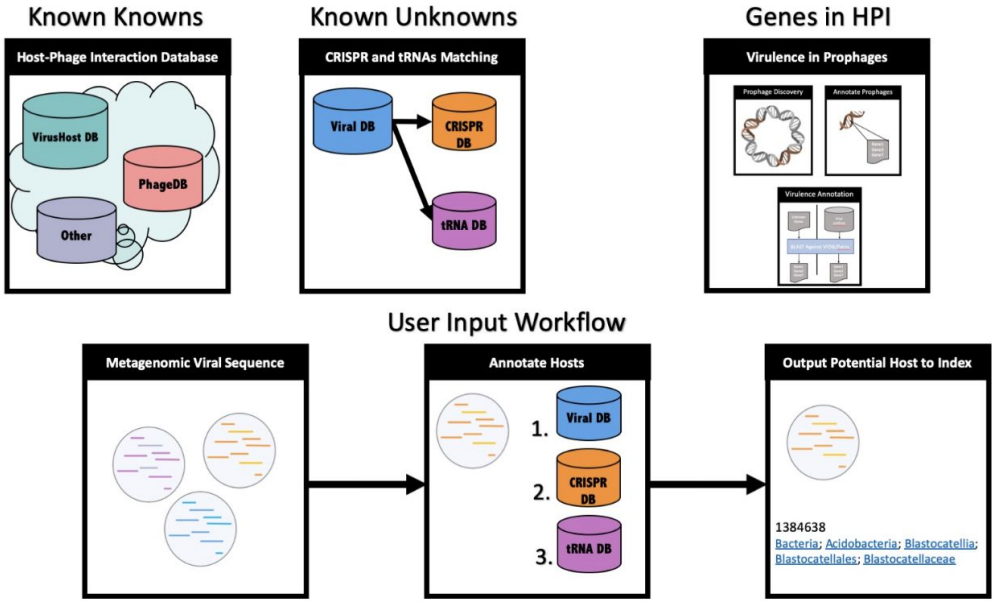
Indexing Data for Federated Discovery on any Platform, Anywhere in the World!



Indexing Data for Federated Discovery on any Platform, Anywhere in the World!

Hosts (black circles) Linked to Viruses (red circles)





Indexing Data for Federated Discovery on any Platform, anywhere in the World!

The_Virus_Index

A Federated Index of Virus Metadata and Hyperdata in Public Repositories

API

Status: Extensible DRAFT API

build passing

https://test.pypi.org/project/viral-index/

Requirements:

- `python3`
- A Google Cloud Platform (GCP) account. Please see GCP's getting started guide if you are new to GCP.

Christiam Camacho, Sej Modha, Alex Efremov, Joan Marti-Carreras

The Importance of Metadata

If biologists leverage the data indices or pipelines for the illustrative subject areas listed here, without metadata, they will simply be doing more indexing and cataloging. For maximum utility, the primary data must be accompanied by rich metadata that puts the samples in question in context, not just in terms of their technical origins, but also their biological origins; for example, their disease state, tissue type, and the precise geographic location, not of the sequencing, but where the individual organism lives.

https://ncbi-codeathons.github.io

https://biohackathons.github.io
(not an NCBI site)

https://www.github.com/ncbi-hackathons
(Archive)

https://www.github.com/ncbi-codeathons

https://ncbi-codeathons.github.io

https://biohackathons.github.io
(not an NCBI site)

https://www.github.com/ncbi-hackathons
(Archive)

https://www.github.com/ncbi-codeathons