Establishing Gen3 to enable better human genome data sharing in Australia

Welcome! The webinar will commence at 1pm AEDT /12pm AEST/ 11:30am ACDT/ 10am AWST

Australian BioCommons

Australian BioCommons

Actively supporting Australian life sciences research through bioinformatics and bioscience data infrastructure

biocommons.org.au ▶ AustralianBioCommons @AusBiocommons

Acknowledgement of Country

We acknowledge the Traditional Owners and their custodianship of the lands on which we meet today.

We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.

We recognise their valuable contributions to Australian and global society.

Housekeeping

Session is recorded

Autogenerated captions available

Questions via Q&A function

Establishing Gen3 to enable better human genome data sharing in Australia

SPEAKERS

Associate Professor Bernie Pope, Australian BioCommons / Melbourne Bioinformatics
Professor Oliver Hofmann, University of Melbourne Center for Cancer Research
Mr Kamile Taouk, Children's Cancer Institute
Dr Marie Wong-Erasmus, Children's Cancer Institute

zero CHILDHOOD CANCER

Melbourne Bioinformatics
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

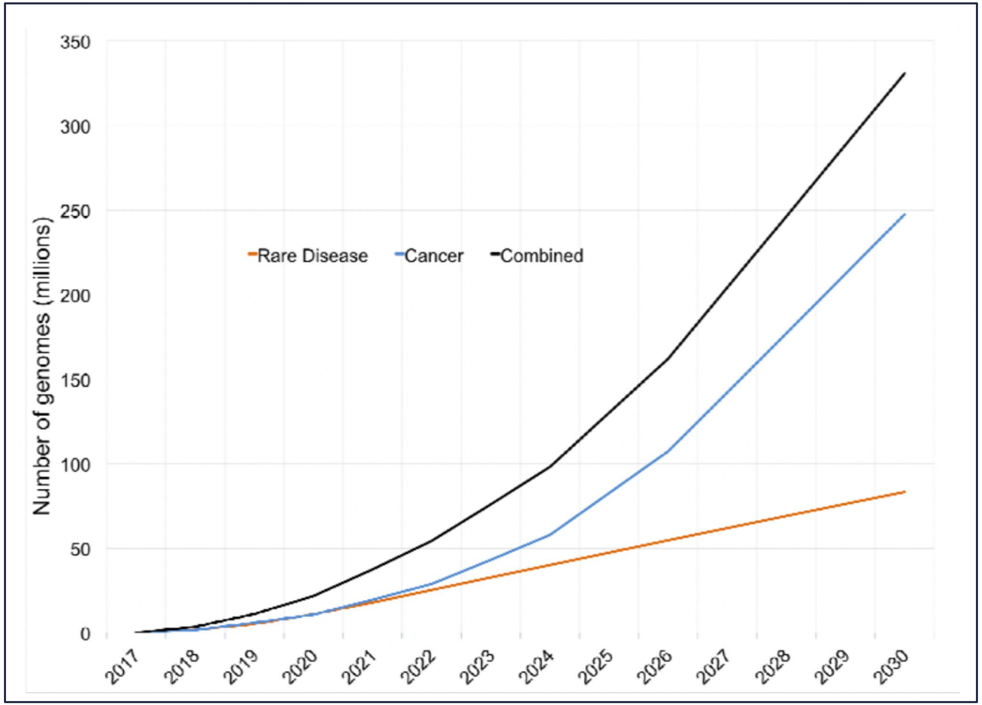
THE UNIVERSITY OF MELBOURNE

AUSTRALIAN ACCESS FEDERATION

GEN3 DATA COMMONS

Australian BioCommons

Predicted global growth of healthcare funded sequenced human genomes

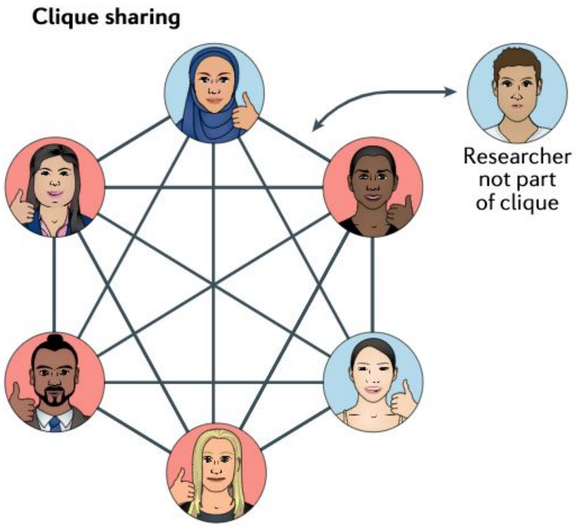


Global storage requirements in 2025 to be exabytes to low zettabytes.

Birney, E., Vamathevan, J., and Goodhand, P. (2017). Genomics in healthcare: GA4GH looks to 2022. bioRxiv

Siloed data

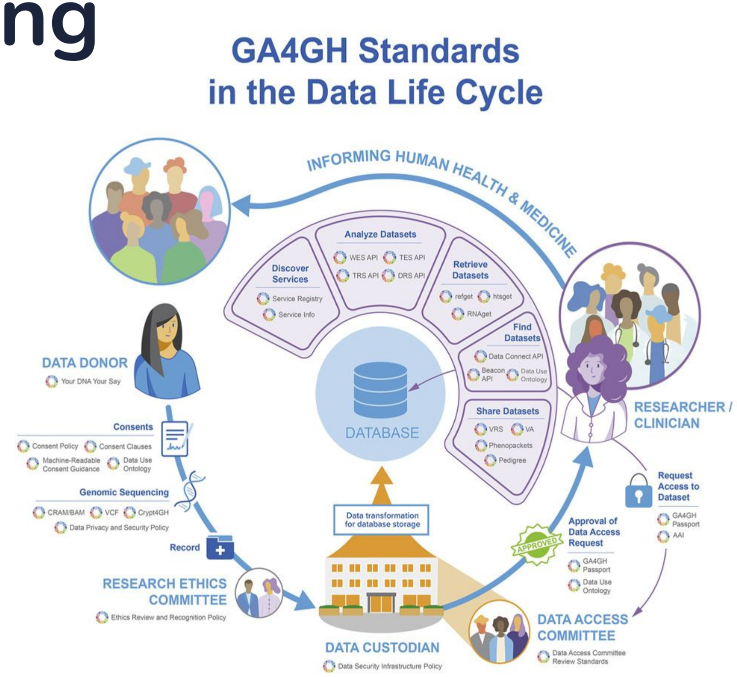
- Human genomics data has often been siloed.
- This limits reuse and reanalysis.
- Public benefit is increased when data is shared.
- Sharing is frequently necessary in human health, especially in rare disease and cancer.
- Large cohorts are needed for statistical power.
- National and international data sharing is highly beneficial but requires considerable collaboration and coordination.



Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X. & Greene, C. S. Responsible, practical genomic data sharing that accelerates research. Nat. Rev. Genet. 21, 615–629 (2020).

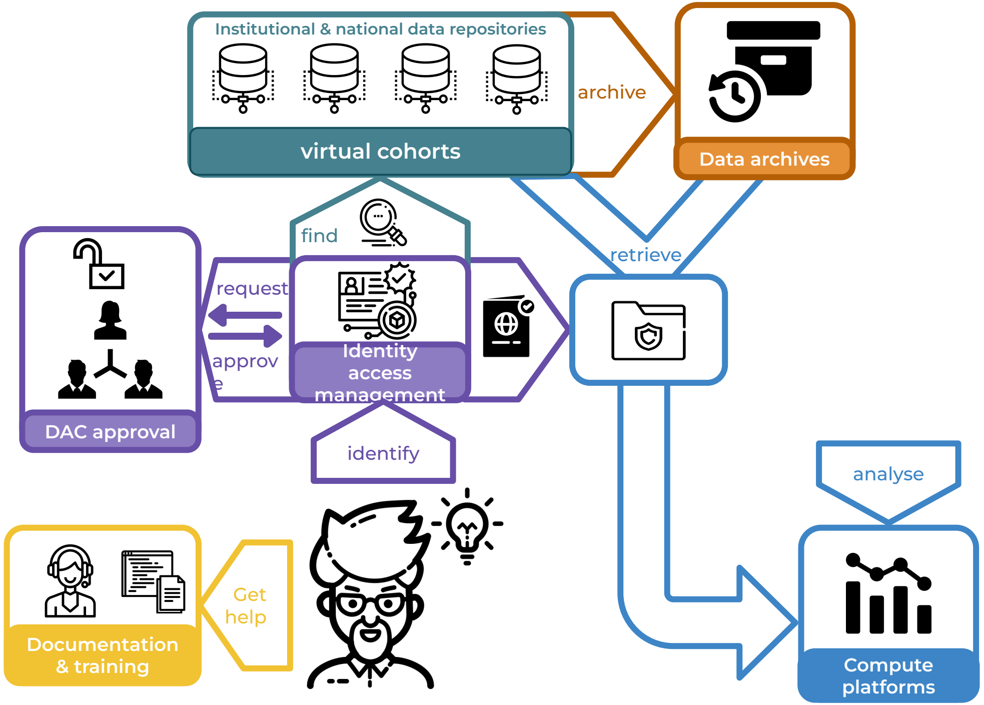
Towards global standards for data sharing

- The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a human rights framework.
- Australian Genomics is a driver project of GA4GH.
- A key outcome is the specification for standard APIs for data sharing technology.
- Recognition that the data life cycle in human genomics is complex and data storage and analysis are parts of a bigger ecosystem.



Rehm, H. L. *et al.* GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genomics* 1, (2021).

Infrastructure ecosystem



Example solutions:

- IAM: CILogon, GA4GH passports
- Data commons: Gen3
- DAC approval: REMS
- Analytics: national infrastructure, institutional infrastructure, commercial cloud
- Integrated: Broad Terra + DUOS

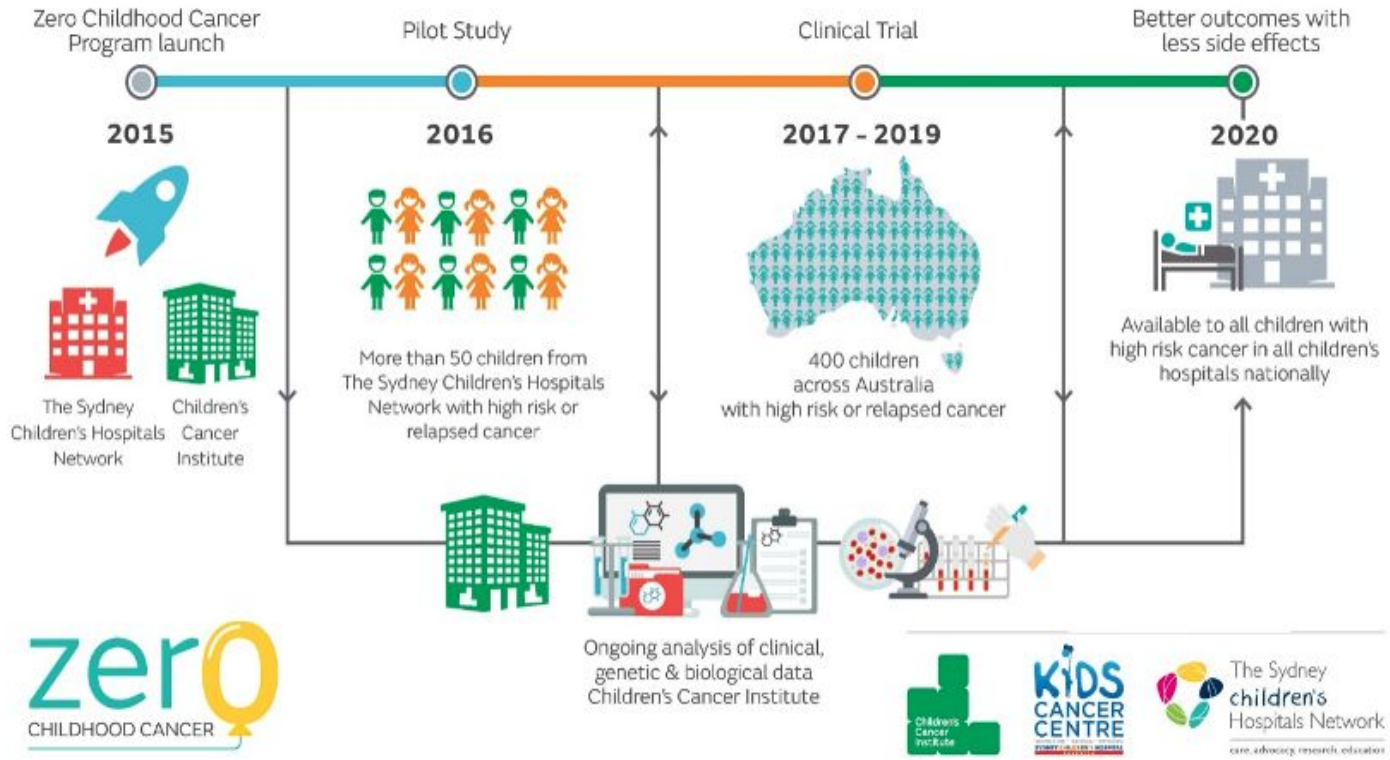
Marion Shadbolt

Establishing Gen3 to enable better Human Genome Data sharing in Australia

- Gen3 was identified as a leading candidate for building a human genomics data commons.
- In Q3 2021 we established a pilot project to assess the use of Gen3 as the foundation for a human genomics data commons.
- That project is now complete, and today we provide an overview of the motivations, process, and findings.

Zero Childhood Cancer - Australia's national paediatric cancer precision medicine program

ZERO2: by 2023
all children in Australia



Zero Childhood Cancer Program launch

2015

The Sydney Children's Hospitals Network

Children's Cancer Institute

Pilot Study

2016

More than 50 children from The Sydney Children's Hospitals Network with high risk or relapsed cancer

Clinical Trial

2017 - 2019

400 children across Australia with high risk or relapsed cancer

Better outcomes with less side effects

2020

Available to all children with high risk cancer in all children's hospitals nationally

Ongoing analysis of clinical, genetic & biological data Children's Cancer Institute

zero
CHILDHOOD CANCER
Children's Cancer Institute
KIDS CANCER CENTRE
The Sydney children's Hospitals Network
care, advocacy, research, education

WGS

RNAseq

Methylation

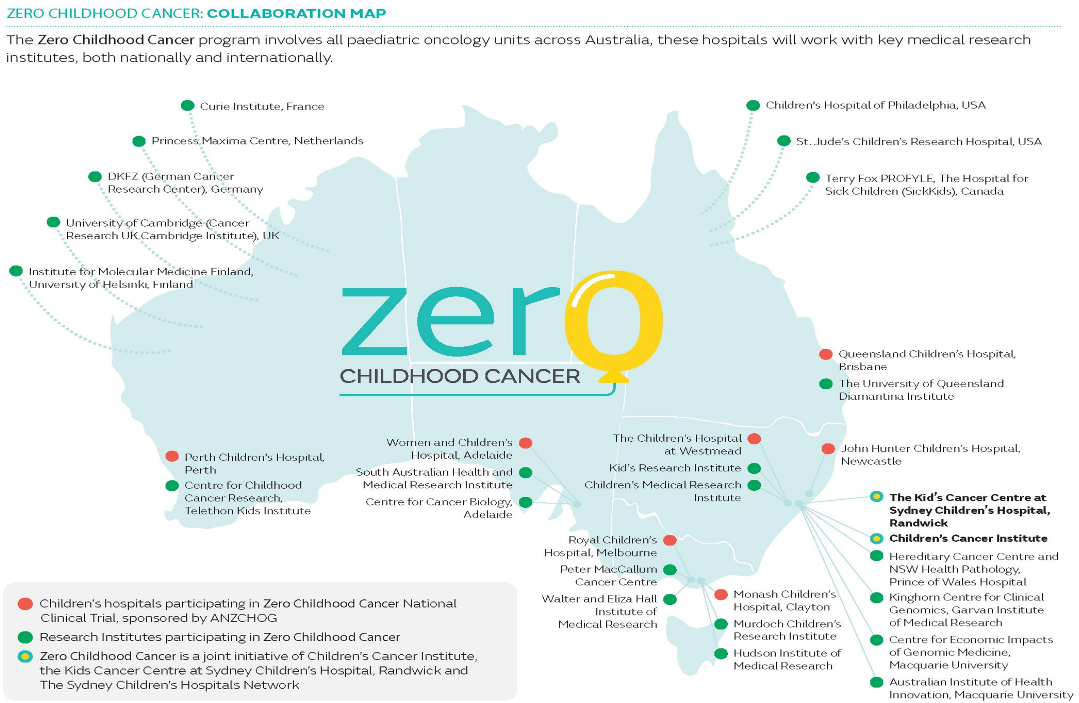
HTS

PDX

clinical

> 650 high-risk patients recruited

Scale up and share



ZERO CHILDHOOD CANCER: **COLLABORATION MAP**

The **Zero Childhood Cancer** program involves all paediatric oncology units across Australia, these hospitals will work with key medical research institutes, both nationally and internationally.

Curie Institute, France
Princess Maxima Centre, Netherlands
DKFZ (German Cancer Research Center), Germany
University of Cambridge (Cancer Research UK Cambridge Institute), UK
Institute for Molecular Medicine Finland, University of Helsinki, Finland
Children's Hospital of Philadelphia, USA
St. Jude's Children's Research Hospital, USA
Terry Fox PROFYLE, The Hospital for Sick Children (SickKids), Canada
zero CHILDHOOD CANCER
Queensland Children's Hospital, Brisbane
The University of Queensland Diamantina Institute
Perth Children's Hospital, Perth
Centre for Childhood Cancer Research, Telethon Kids Institute
Women and Children's Hospital, Adelaide
South Australian Health and Medical Research Institute
Centre for Cancer Biology, Adelaide
The Children's Hospital at Westmead
Kid's Research Institute
Children's Medical Research Institute
John Hunter Children's Hospital, Newcastle
The Kid's Cancer Centre at Sydney Children's Hospital, Randwick
Children's Cancer Institute
Hereditary Cancer Centre and NSW Health Pathology, Prince of Wales Hospital
Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research
Centre for Economic Impacts of Genomic Medicine, Macquarie University
Australian Institute of Health Innovation, Macquarie University
Royal Children's Hospital, Melbourne
Peter MacCallum Cancer Centre
Walter and Eliza Hall Institute of Medical Research
Monash Children's Hospital, Clayton
Murdoch Children's Research Institute
Hudson Institute of Medical Research

- 🟥 Children's hospitals participating in Zero Childhood Cancer National Clinical Trial, sponsored by ANZCHOG
- 🟩 Research Institutes participating in Zero Childhood Cancer
- ● Zero Childhood Cancer is a joint initiative of Children's Cancer Institute, the Kids Cancer Centre at Sydney Children's Hospital, Randwick and The Sydney Children's Hospitals Network

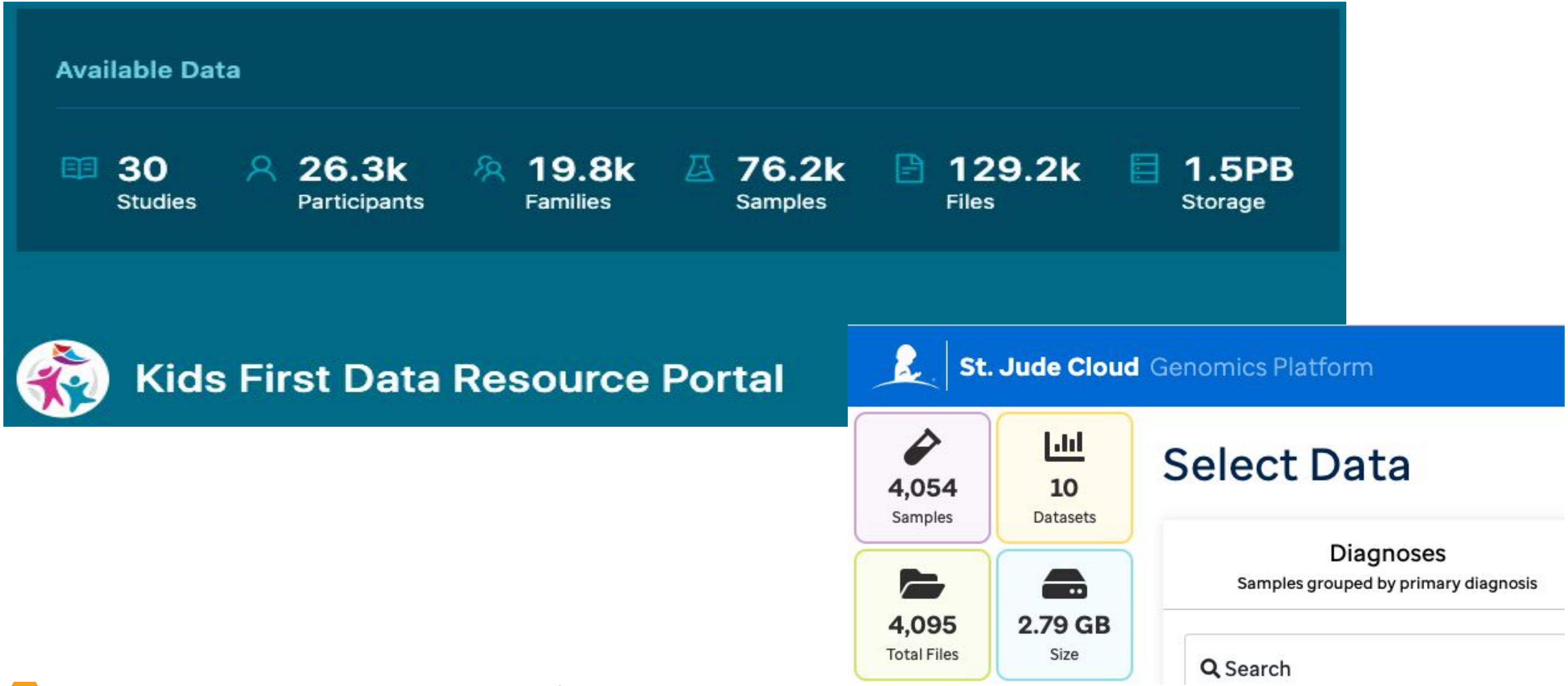
With only **~1000 new cases** of high-risk paediatric cancer per year, it is imperative to **aggregate** Australian data with global data to develop strategies to **effectively treat high-risk childhood cancer**

We need to share, analyse, integrate data more easily

image credit: https://www.zerochildhoodcancer.org.au/about/research---clinical-partners

Australian BioCommons

Paediatric data sources



Available Data

30	26.3k	19.8k	76.2k	129.2k	1.5PB
Studies	Participants	Families	Samples	Files	Storage

Kids First Data Resource Portal

St. Jude Cloud Genomics Platform

4,054	10
Samples	Datasets
4,095	2.79 GB
Total Files	Size

Select Data

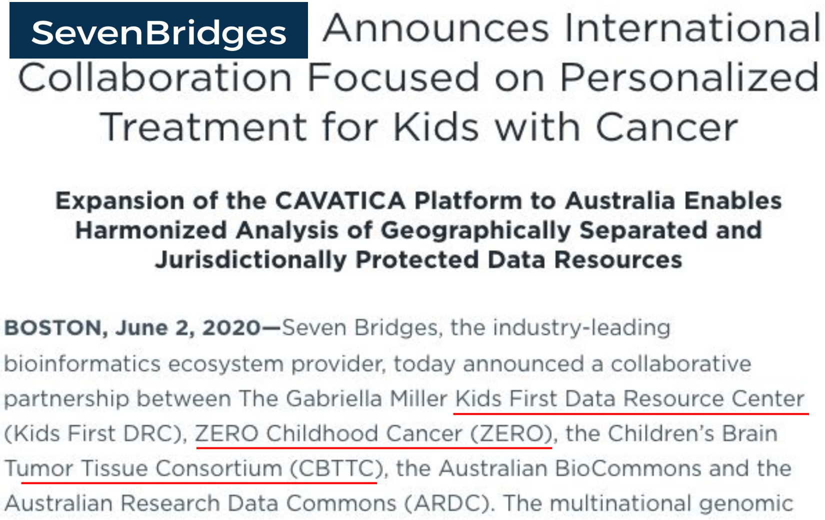
Diagnoses

Samples grouped by primary diagnosis

CCI - where's our data?

NetApp

StorageGRID: smart, fast, and future proof object storage



SevenBridges Announces International Collaboration Focused on Personalized Treatment for Kids with Cancer

Expansion of the CAVATICA Platform to Australia Enables Harmonized Analysis of Geographically Separated and Jurisdictionally Protected Data Resources

BOSTON, June 2, 2020—Seven Bridges, the industry-leading bioinformatics ecosystem provider, today announced a collaborative partnership between The Gabriella Miller Kids First Data Resource Center (Kids First DRC), ZERO Childhood Cancer (ZERO), the Children's Brain Tumor Tissue Consortium (CBTTC), the Australian BioCommons and the Australian Research Data Commons (ARDC). The multinational genomic

CAVATICA

Data commons?

GEN3 DATA COMMONS

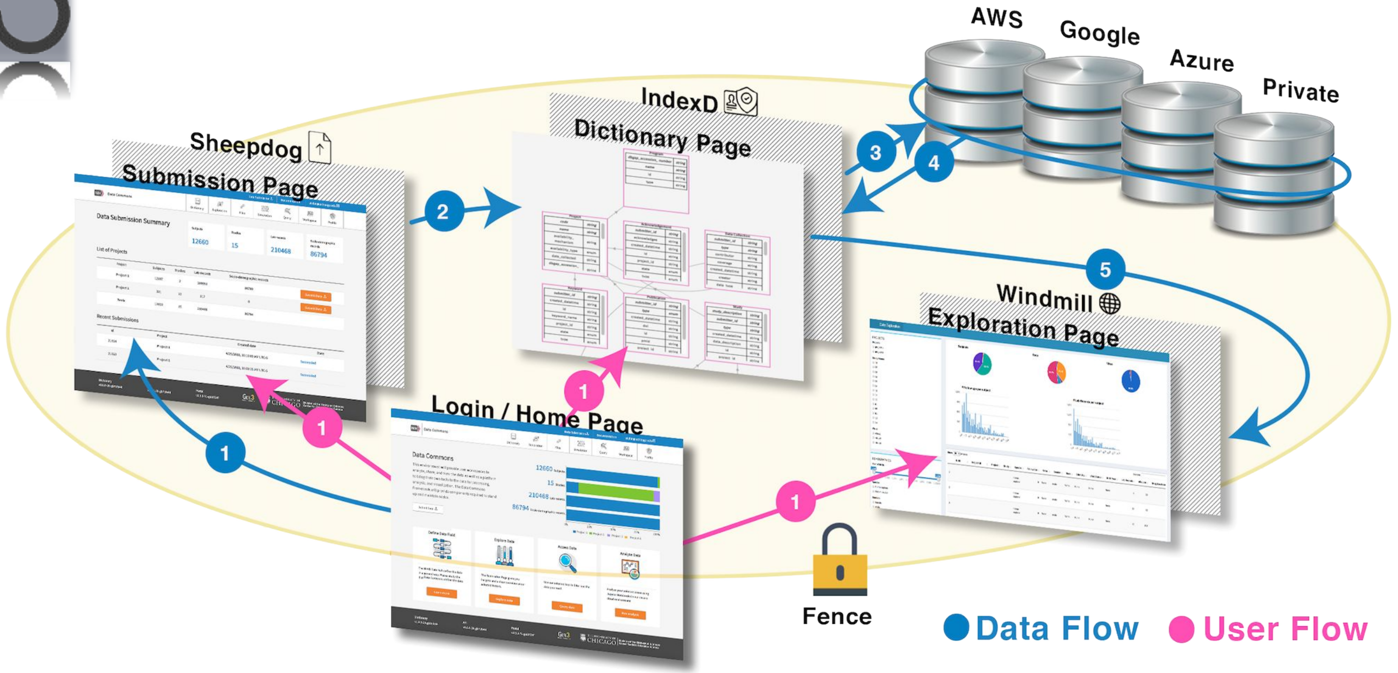
Australian BioCommons

How to share?

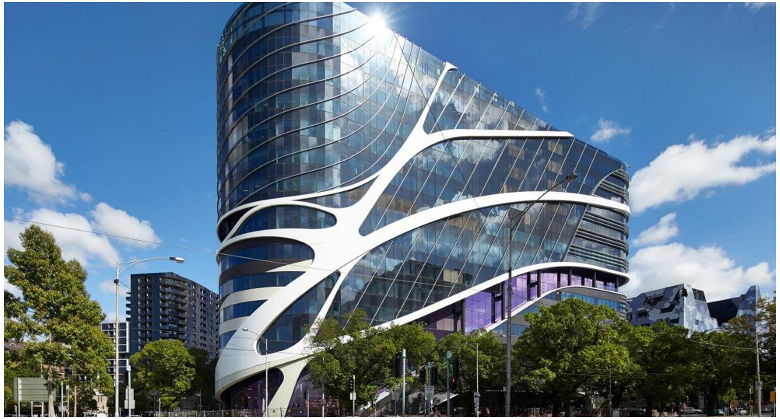
GEN3 DATA COMMONS

We Need:

- a way to SEARCH different catalogues of paediatric data
- a way to easily gain and grant ACCESS to the data
- a way to ANALYSE this data in place if possible

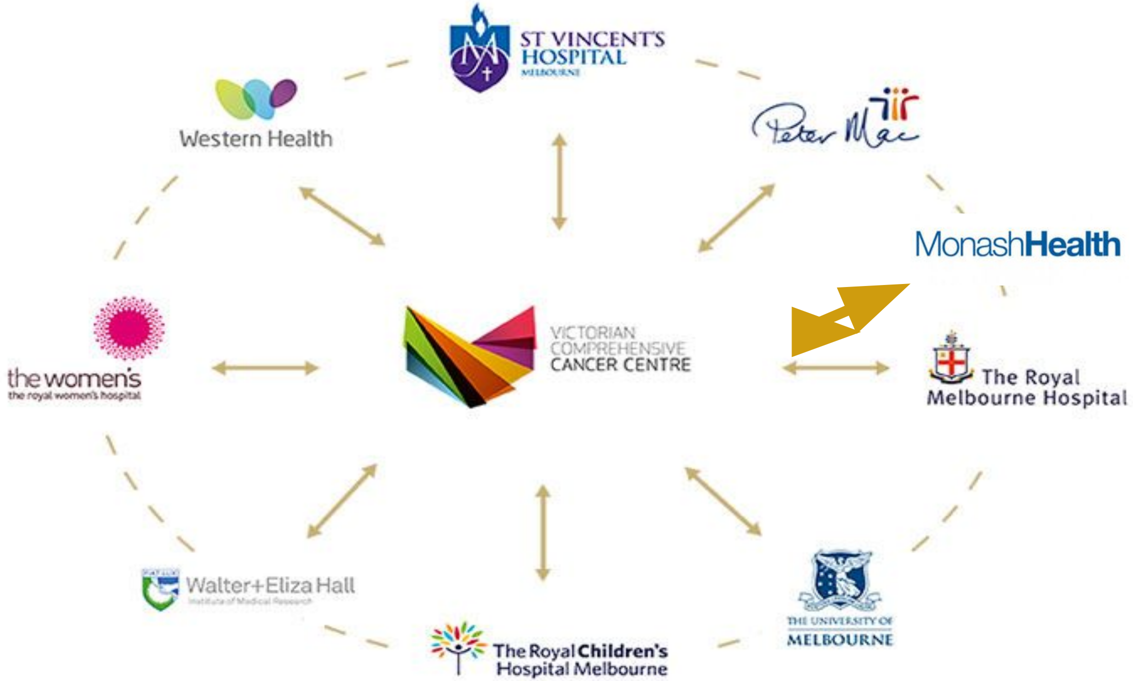


Australian BioCommons



University of Melbourne Centre for Cancer Research

Precision Oncology Program


ST VINCENT'S HOSPITAL MELBOURNE
Western Health
Peter Mac
MonashHealth
the women's
the royal women's hospital
VICTORIAN COMPREHENSIVE CANCER CENTRE
The Royal Melbourne Hospital
Walter+Eliza Hall
Institute of Medical Research
The Royal Children's Hospital Melbourne
THE UNIVERSITY OF MELBOURNE

Sean Grimmond

Precision Oncology Program

Recalcitrant Cancers, Rare Cancers, Cancers of the Unknown Primary

UMCCR Genomics Platform Group

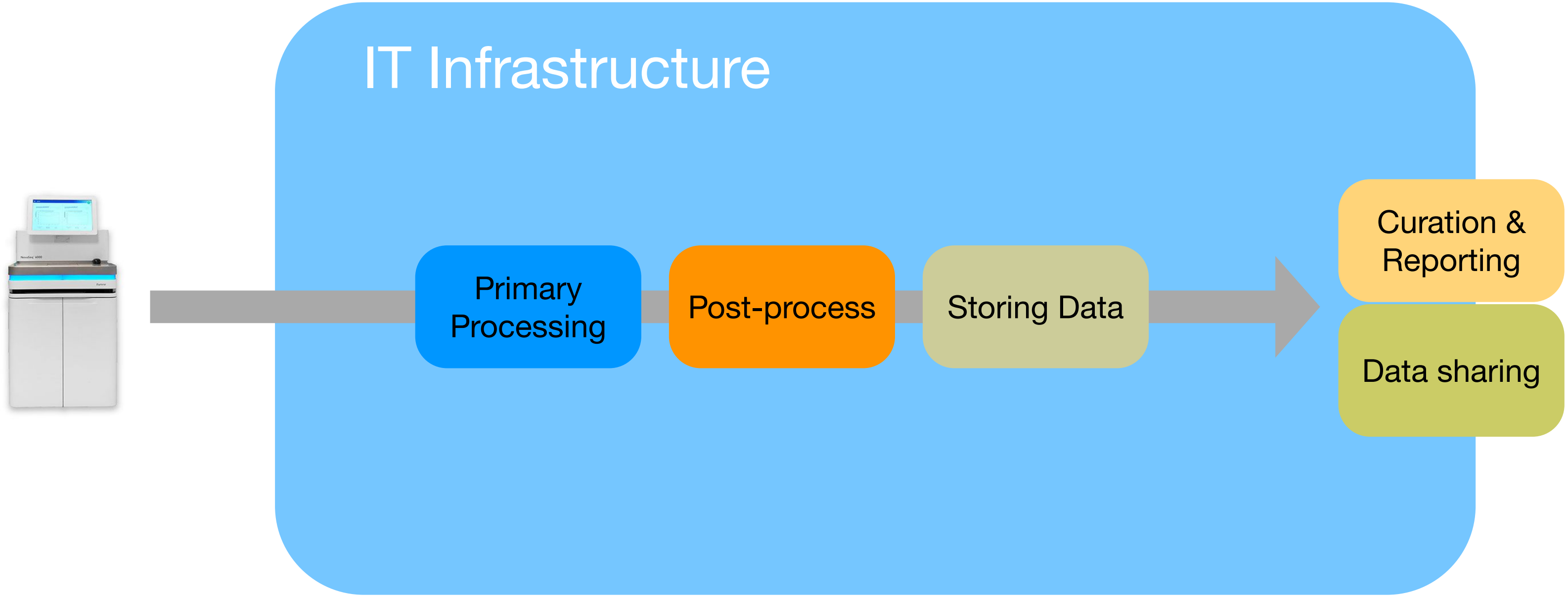
Workflow Development

Technology Assessment

Standards Development & Implementation

Workflow Development: **Rapid WGTS**

Supporting Precision Oncology



IT Infrastructure

Primary Processing

Post-process

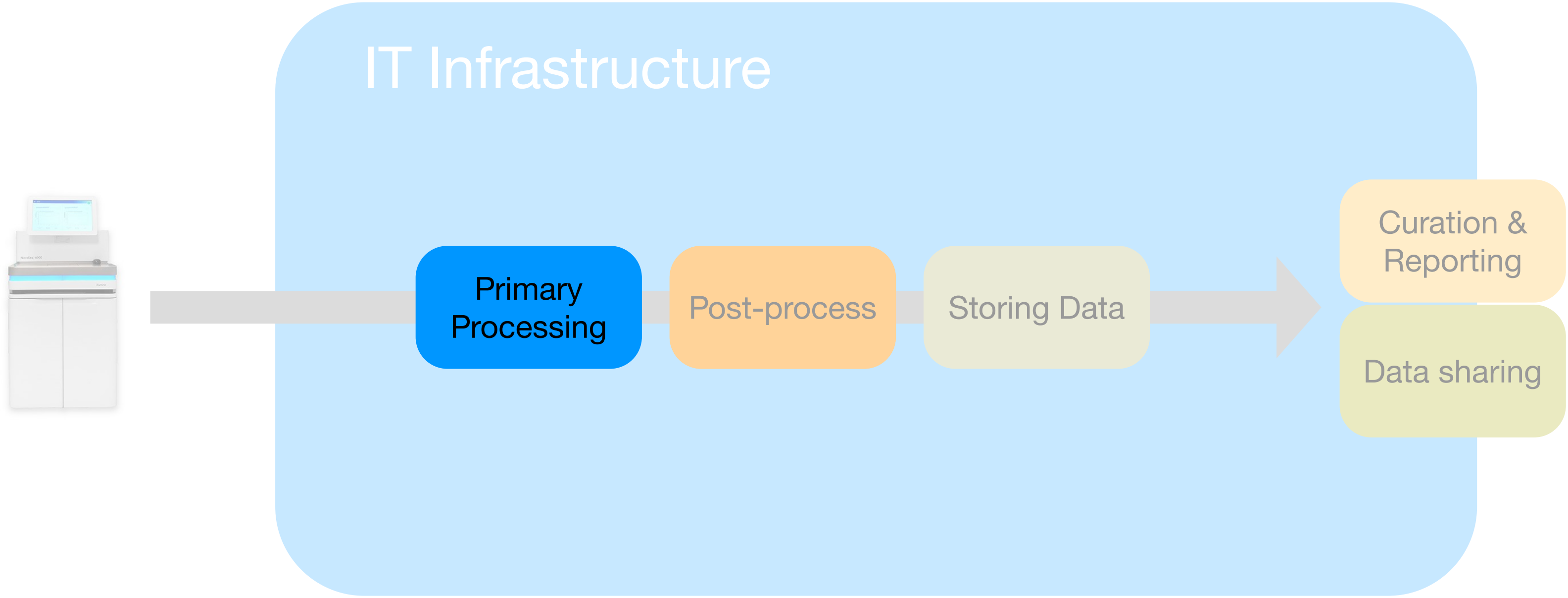
Storing Data

Curation & Reporting

Data sharing

Workflow Development: **Data Flow**

Supporting Precision Oncology



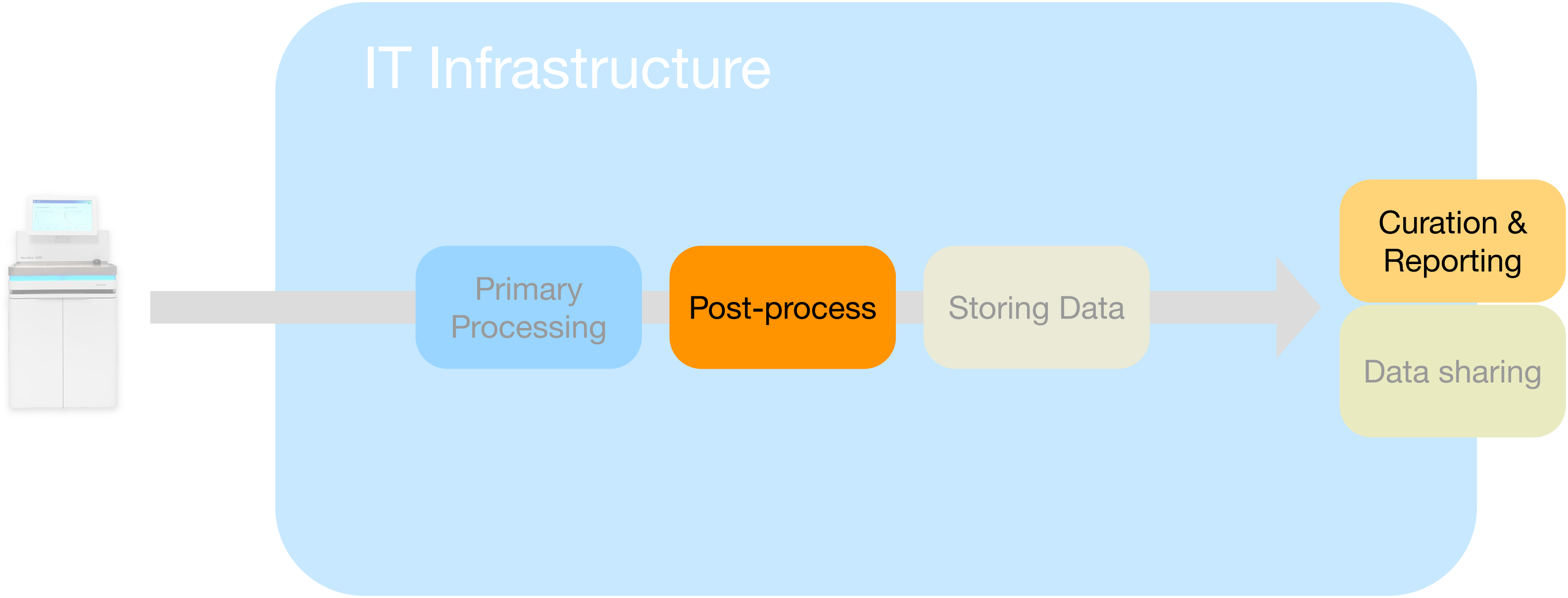
Workflow Development: **Primary Analysis**

Supporting Precision Oncology

Illumina-University of Melbourne Partnership

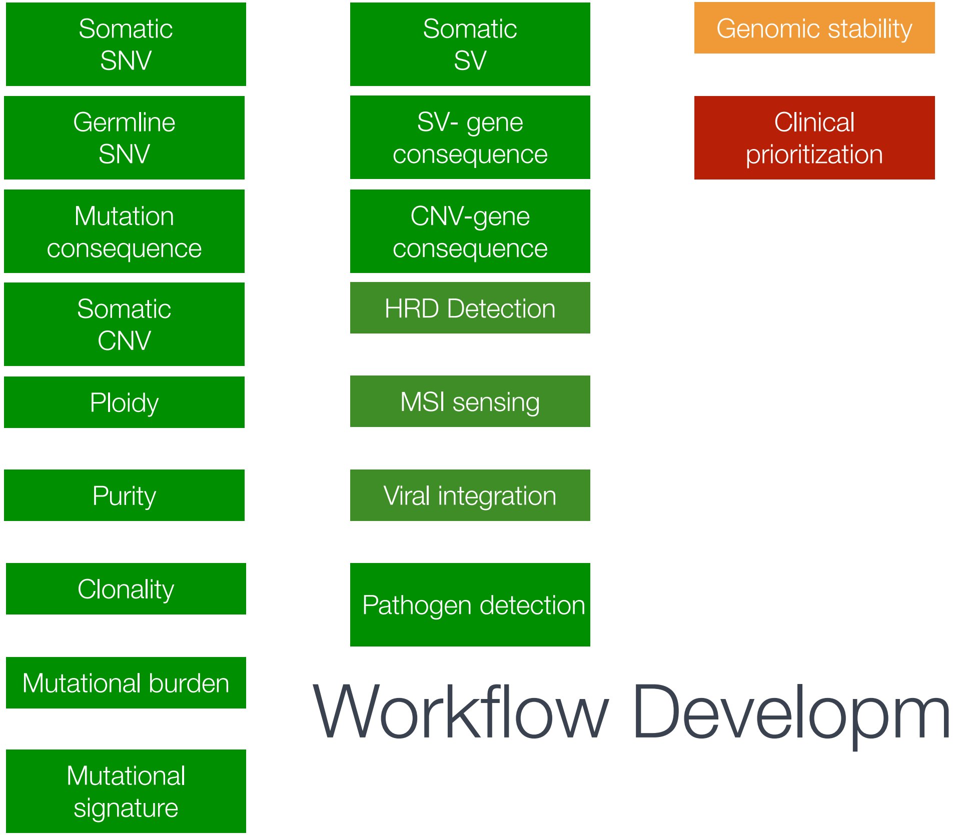
Partnering to provide the infrastructure, expertise, systems and analysis to translate and implement genomics into routine clinical care.

Core Platform: DRAGEN/ICA



Workflow Development: **Post-process / Reporting**

Supporting Precision Oncology



Somatic SNV

Germline SNV

Mutation consequence

Somatic CNV

Ploidy

Purity

Clonality

Mutational burden

Mutational signature

Somatic SV

SV- gene consequence

CNV-gene consequence

HRD Detection

MSI sensing

Viral integration

Pathogen detection

Genomic stability

Clinical prioritization

Workflow Development: **Post-processing**

umccrise

Somatic SNV

Germline SNV

Mutation consequence

Somatic CNV

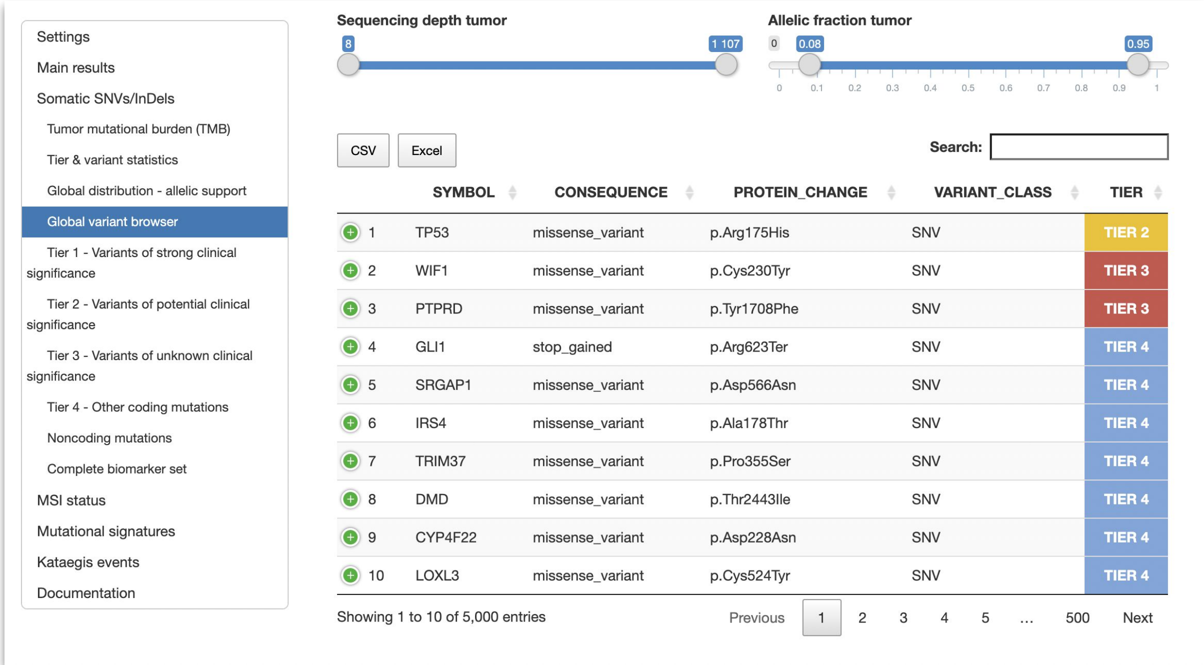
Ploidy

Purity

Clonality

Mutational burden

Mutational signature


Settings
Main results
Somatic SNVs/InDels
Tumor mutational burden (TMB)
Tier & variant statistics
Global distribution - allelic support
Global variant browser
Tier 1 - Variants of strong clinical significance
Tier 2 - Variants of potential clinical significance
Tier 3 - Variants of unknown clinical significance
Tier 4 - Other coding mutations
Noncoding mutations
Complete biomarker set
MSI status
Mutational signatures
Kataegis events
Documentation
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
CSV
Excel
Search:
Previous
1
2
3
4
5
...
500
Next

Sequencing depth tumor: 8 – 1 107

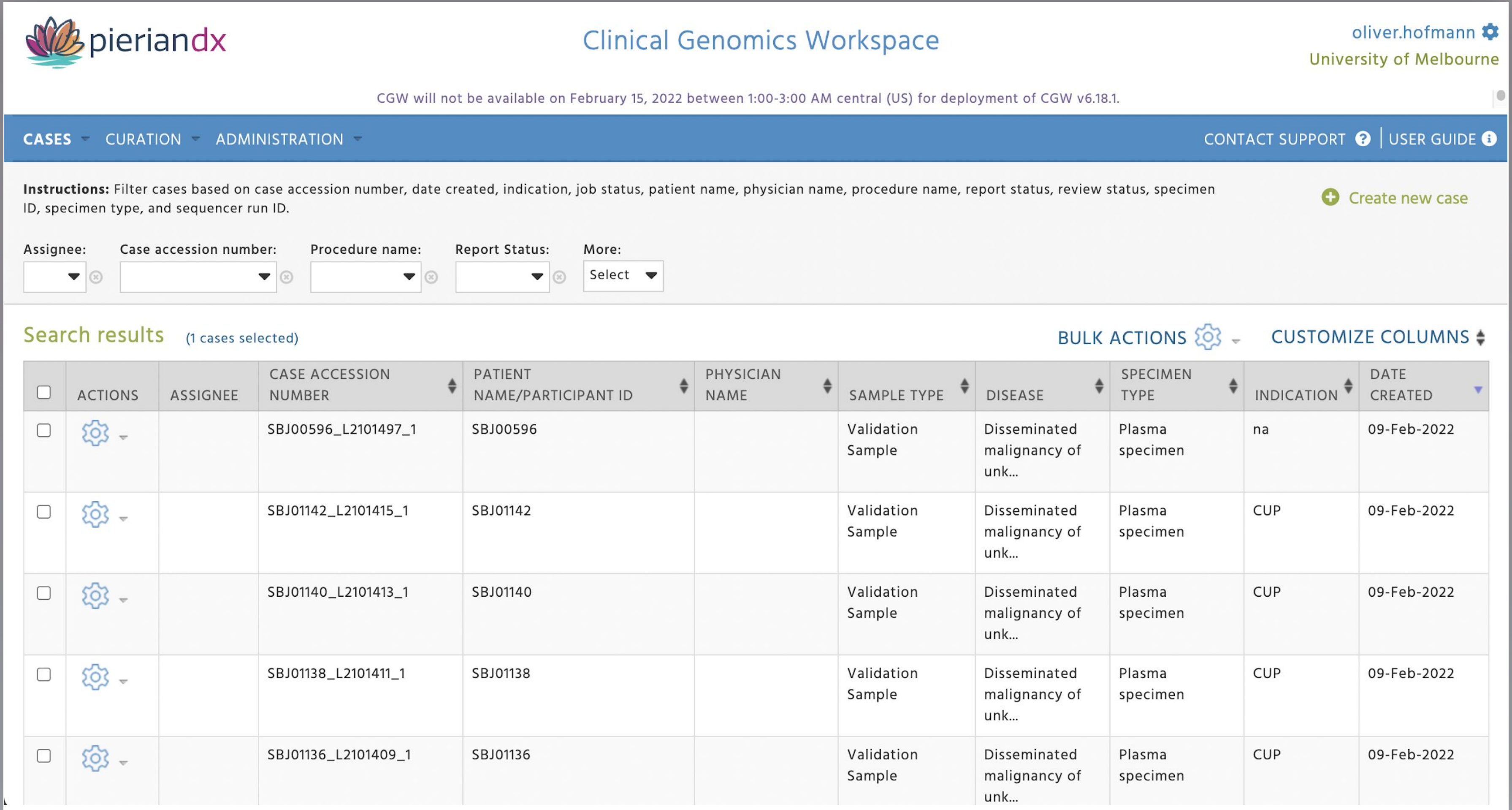
Allelic fraction tumor: 0.08 – 0.95

	SYMBOL	CONSEQUENCE	PROTEIN_CHANGE	VARIANT_CLASS	TIER
1	TP53	missense_variant	p.Arg175His	SNV	TIER 2
2	WIF1	missense_variant	p.Cys230Tyr	SNV	TIER 3
3	PTPRD	missense_variant	p.Tyr1708Phe	SNV	TIER 3
4	GLI1	stop_gained	p.Arg623Ter	SNV	TIER 4
5	SRGAP1	missense_variant	p.Asp566Asn	SNV	TIER 4
6	IRS4	missense_variant	p.Ala178Thr	SNV	TIER 4
7	TRIM37	missense_variant	p.Pro355Ser	SNV	TIER 4
8	DMD	missense_variant	p.Thr2443Ile	SNV	TIER 4
9	CYP4F22	missense_variant	p.Asp228Asn	SNV	TIER 4
10	LOXL3	missense_variant	p.Cys524Tyr	SNV	TIER 4

Showing 1 to 10 of 5,000 entries

Workflow Development: Reporting

MultiQC, PCGR/CPSR, …



CGW will not be available on February 15, 2022 between 1:00-3:00 AM central (US) for deployment of CGW v6.18.1.

Instructions: Filter cases based on case accession number, date created, indication, job status, patient name, physician name, procedure name, report status, review status, specimen ID, specimen type, and sequencer run ID.

Assignee:

Case accession number:

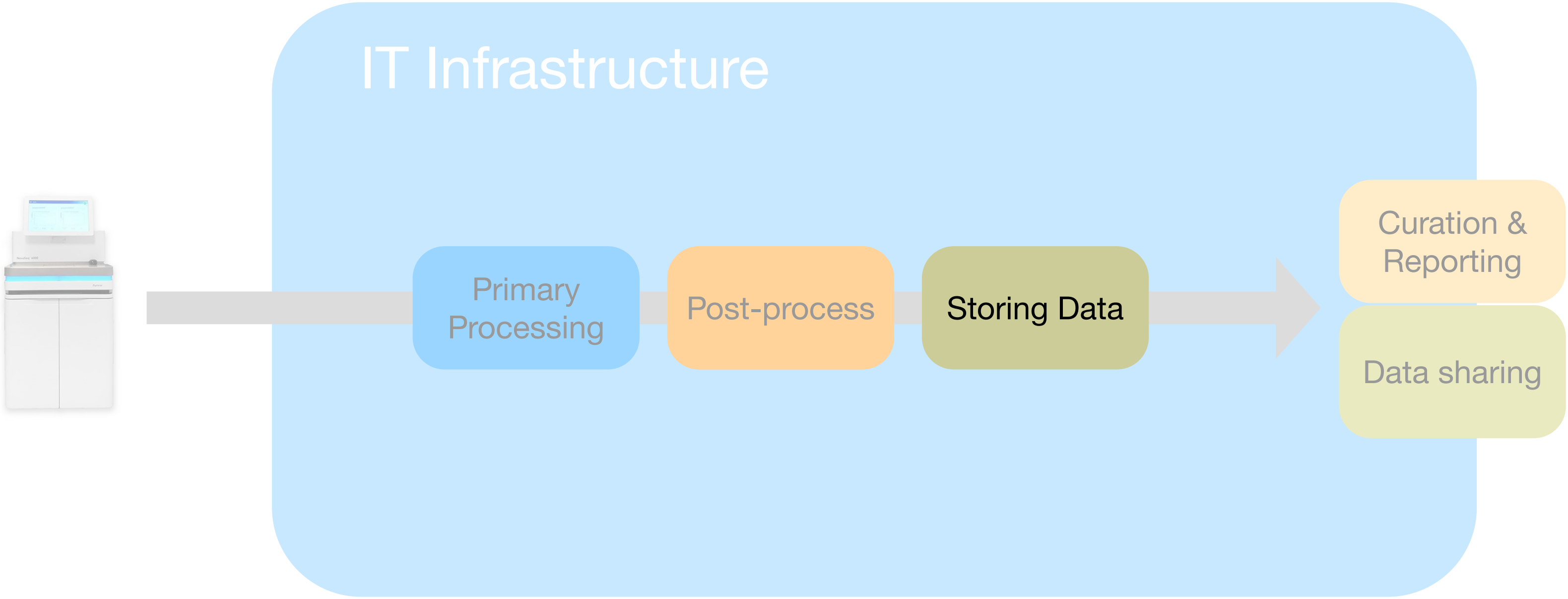
Procedure name:

Report Status:

More: Select

Search results (1 cases selected)

□	ACTIONS	ASSIGNEE	CASE ACCESSION NUMBER	PATIENT NAME/PARTICIPANT ID	PHYSICIAN NAME	SAMPLE TYPE	DISEASE	SPECIMEN TYPE	INDICATION	DATE CREATED
□			SBJ00596_L2101497_1	SBJ00596		Validation Sample	Disseminated malignancy of unk...	Plasma specimen	na	09-Feb-2022
□			SBJ01142_L2101415_1	SBJ01142		Validation Sample	Disseminated malignancy of unk...	Plasma specimen	CUP	09-Feb-2022
□			SBJ01140_L2101413_1	SBJ01140		Validation Sample	Disseminated malignancy of unk...	Plasma specimen	CUP	09-Feb-2022
□			SBJ01138_L2101411_1	SBJ01138		Validation Sample	Disseminated malignancy of unk...	Plasma specimen	CUP	09-Feb-2022
□			SBJ01136_L2101409_1	SBJ01136		Validation Sample	Disseminated malignancy of unk...	Plasma specimen	CUP	09-Feb-2022

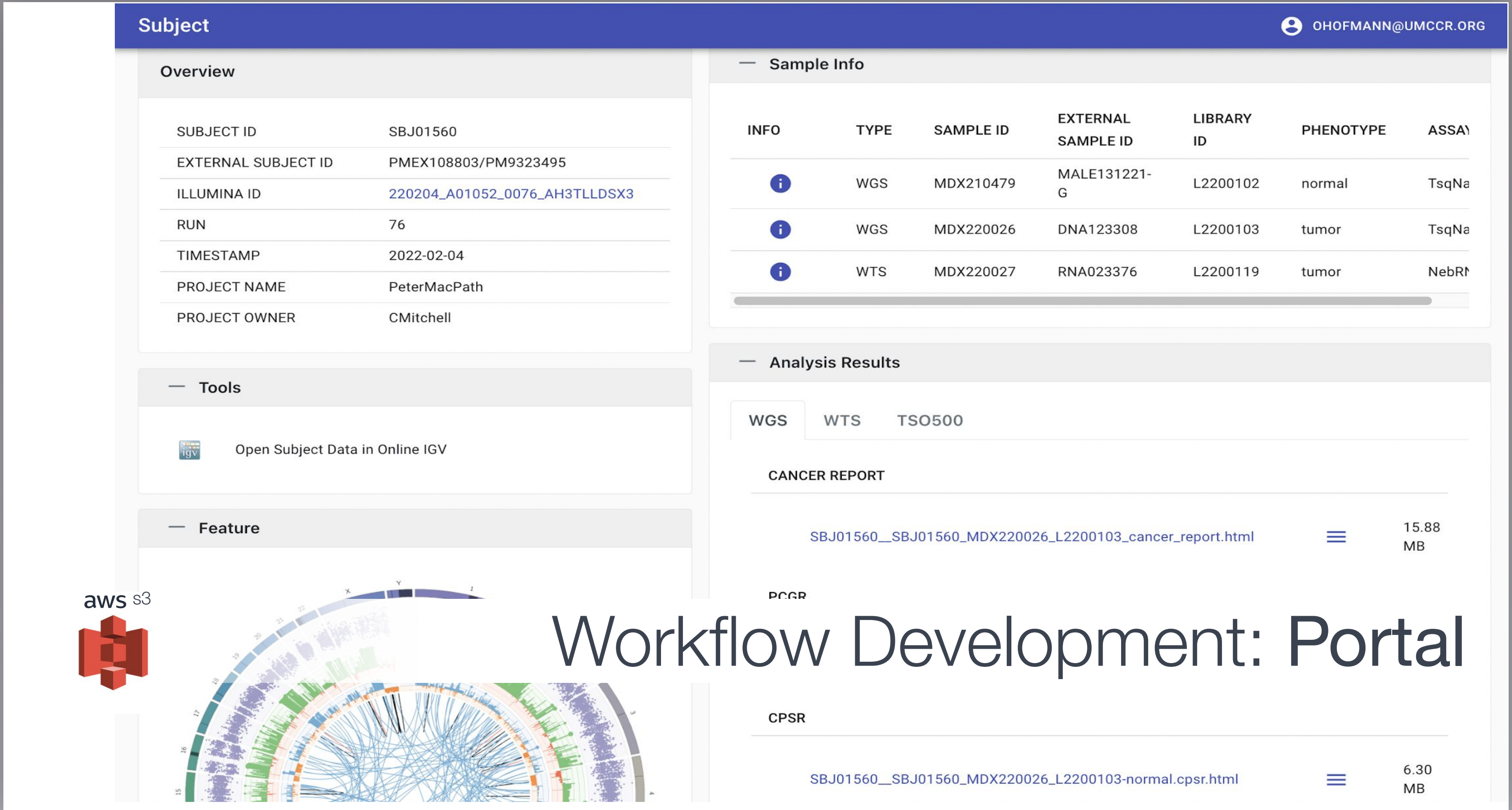


Workflow Development: **Archiving**

Supporting Precision Oncology

Workflow Development: Portal

aws s3



Subject

OHOFMANN@UMCCR.ORG

Overview

SUBJECT ID	SBJ01560
EXTERNAL SUBJECT ID	PMEX108803/PM9323495
ILLUMINA ID	220204_A01052_0076_AH3TLLDSX3
RUN	76
TIMESTAMP	2022-02-04
PROJECT NAME	PeterMacPath
PROJECT OWNER	CMitchell

— Tools

Open Subject Data in Online IGV

— Feature

— Sample Info

INFO	TYPE	SAMPLE ID	EXTERNAL SAMPLE ID	LIBRARY ID	PHENOTYPE	ASSAY
i	WGS	MDX210479	MALE131221-G	L2200102	normal	TsqNa
i	WGS	MDX220026	DNA123308	L2200103	tumor	TsqNa
i	WTS	MDX220027	RNA023376	L2200119	tumor	NebRN

— Analysis Results

WGS WTS TSO500

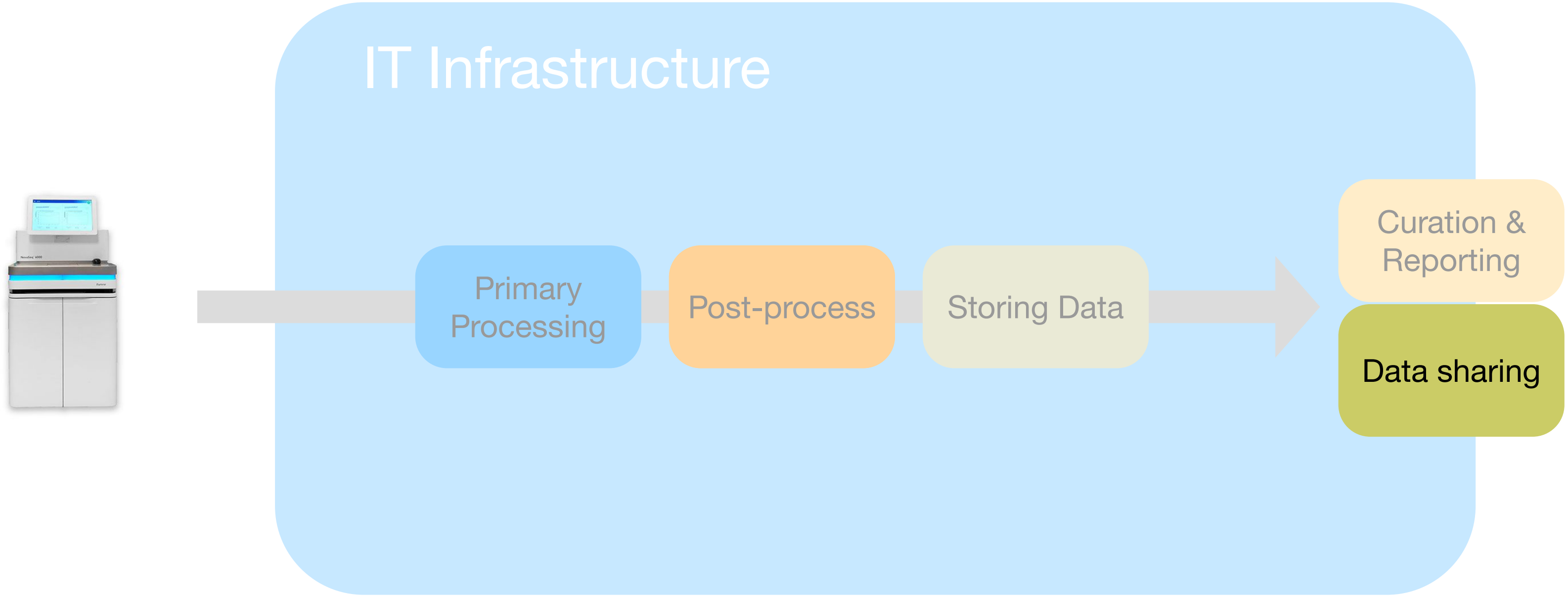
CANCER REPORT

SBJ01560__SBJ01560_MDX220026_L2200103_cancer_report.html	≡	15.88 MB

PCGR

CPSR

SBJ01560__SBJ01560_MDX220026_L2200103-normal.cpsr.html	≡	6.30 MB



Workflow Development: **Data Sharing**

A mandate to share data with the scientific community

Sharing data saves lives

THE GLOBAL ALLIANCE FOR GENOMICS & HEALTH

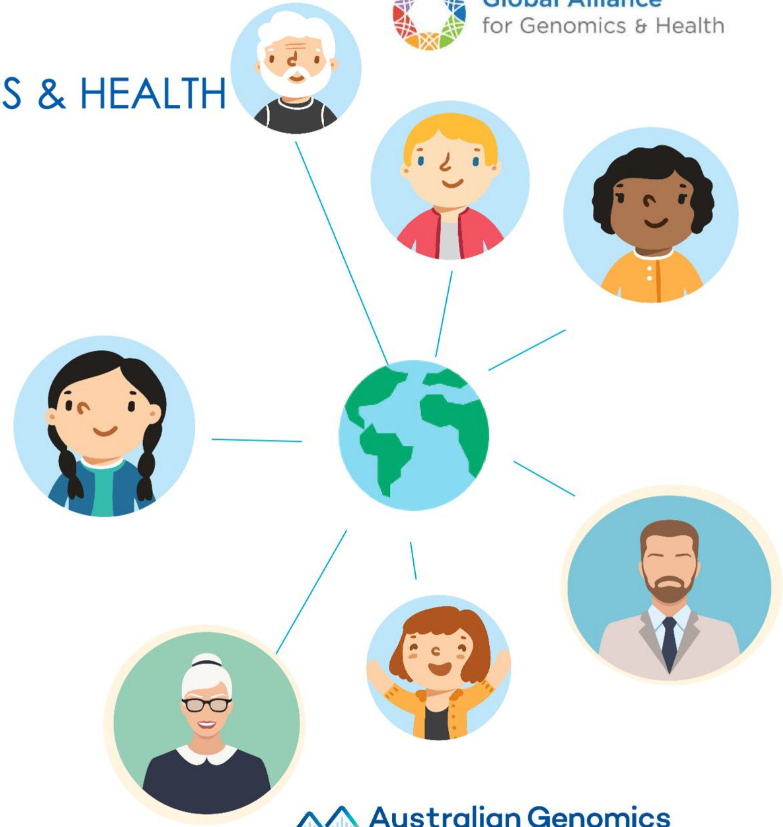
Global Alliance
for Genomics & Health

The need:

- Data from ***millions of samples*** is needed to address questions in rare disease, complex disease and cancer.

The challenge:

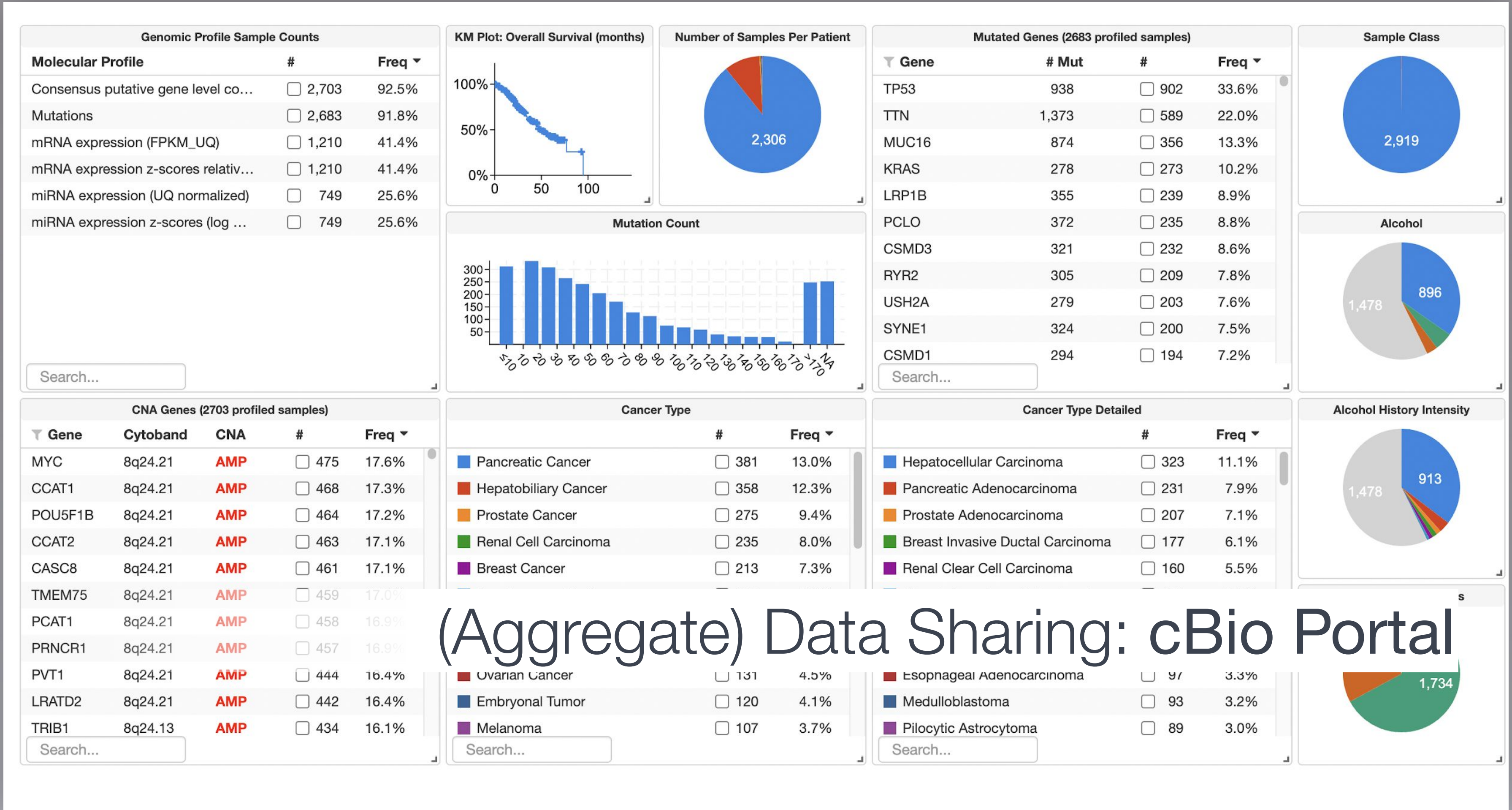
- Data in silos.
- Lack of standard analysis methods.
- Different approaches to regulation, consent and data sharing.



Australian Genomics
Health Alliance

Slide courtesy of Tiffany Boughtwood, Australian Genomics

(Aggregate) Data Sharing: cBio Portal



Genomic Profile Sample Counts

Molecular Profile	#	Freq
Consensus putative gene level co…	□ 2,703	92.5%
Mutations	□ 2,683	91.8%
mRNA expression (FPKM_UQ)	□ 1,210	41.4%
mRNA expression z-scores relativ…	□ 1,210	41.4%
miRNA expression (UQ normalized)	□ 749	25.6%
miRNA expression z-scores (log …	□ 749	25.6%

Search...

KM Plot: Overall Survival (months)

Number of Samples Per Patient

Mutated Genes (2683 profiled samples)

Gene	# Mut	#	Freq
TP53	938	□ 902	33.6%
TTN	1,373	□ 589	22.0%
MUC16	874	□ 356	13.3%
KRAS	278	□ 273	10.2%
LRP1B	355	□ 239	8.9%
PCLO	372	□ 235	8.8%
CSMD3	321	□ 232	8.6%
RYR2	305	□ 209	7.8%
USH2A	279	□ 203	7.6%
SYNE1	324	□ 200	7.5%
CSMD1	294	□ 194	7.2%

Search...

Sample Class

Mutation Count

Alcohol

CNA Genes (2703 profiled samples)

Gene	Cytoband	CNA	#	Freq
MYC	8q24.21	AMP	□ 475	17.6%
CCAT1	8q24.21	AMP	□ 468	17.3%
POU5F1B	8q24.21	AMP	□ 464	17.2%
CCAT2	8q24.21	AMP	□ 463	17.1%
CASC8	8q24.21	AMP	□ 461	17.1%
TMEM75	8q24.21	AMP	□ 459	17.0%
PCAT1	8q24.21	AMP	□ 458	16.9%
PRNCR1	8q24.21	AMP	□ 457	16.9%
PVT1	8q24.21	AMP	□ 444	16.4%
LRATD2	8q24.21	AMP	□ 442	16.4%
TRIB1	8q24.13	AMP	□ 434	16.1%

Search...

Cancer Type

	#	Freq
🟦 Pancreatic Cancer	□ 381	13.0%
🟥 Hepatobiliary Cancer	□ 358	12.3%
🟧 Prostate Cancer	□ 275	9.4%
🟩 Renal Cell Carcinoma	□ 235	8.0%
🟪 Breast Cancer	□ 213	7.3%
🟥 Ovarian Cancer	□ 131	4.5%
🟦 Embryonal Tumor	□ 120	4.1%
🟪 Melanoma	□ 107	3.7%

Search...

Cancer Type Detailed

	#	Freq
🟦 Hepatocellular Carcinoma	□ 323	11.1%
🟥 Pancreatic Adenocarcinoma	□ 231	7.9%
🟧 Prostate Adenocarcinoma	□ 207	7.1%
🟩 Breast Invasive Ductal Carcinoma	□ 177	6.1%
🟪 Renal Clear Cell Carcinoma	□ 160	5.5%
🟥 Esophageal Adenocarcinoma	□ 97	3.3%
🟦 Medulloblastoma	□ 93	3.2%
🟪 Pilocytic Astrocytoma	□ 89	3.0%

Search...

Alcohol History Intensity

UMCCR Genomics Platform Group

Workflow Development

Technology Assessment

Standards Development & Implementation

UMCCR Genomics Platform Group

Workflow Development

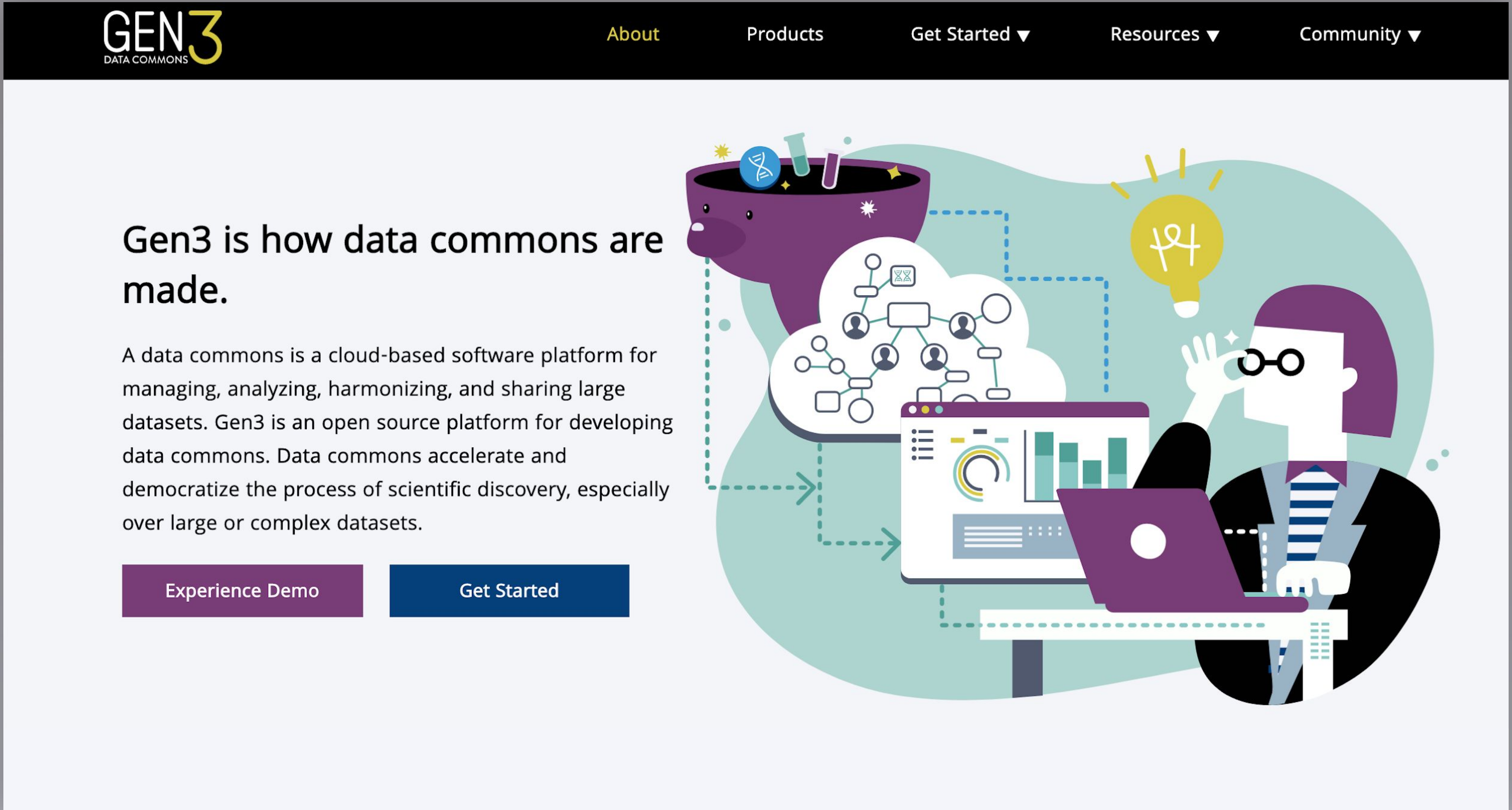
Technology Assessment

Standards Development & Implementation

Victor San Kho Lin

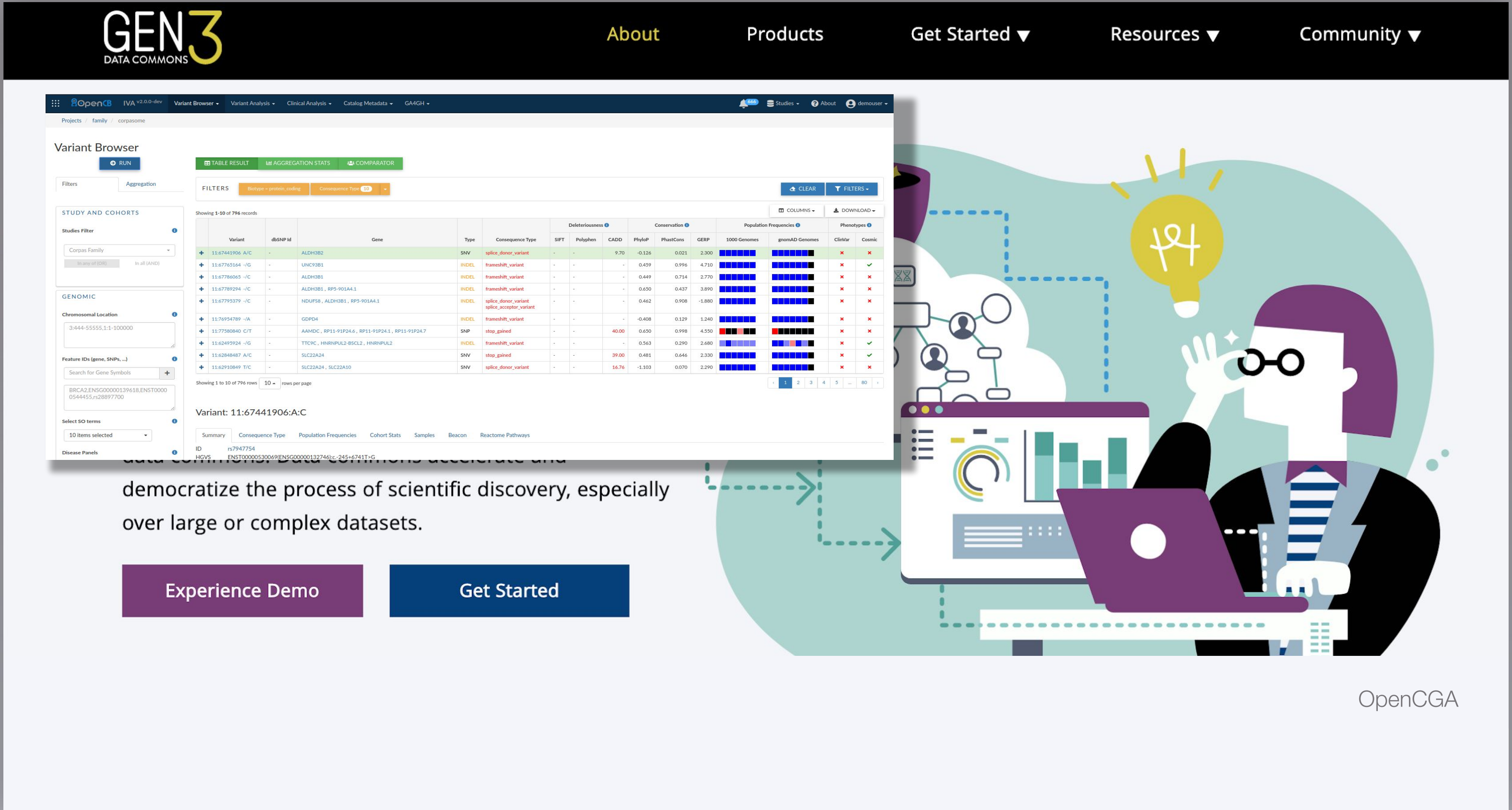
Andrew Patterson

Florian Reisinger



Gen3 is how data commons are made.

A data commons is a cloud-based software platform for managing, analyzing, harmonizing, and sharing large datasets. Gen3 is an open source platform for developing data commons. Data commons accelerate and democratize the process of scientific discovery, especially over large or complex datasets.



GEN3 DATA COMMONS

About | Products | Get Started ▾ | Resources ▾ | Community ▾

OpenCB IVA v2.0.0-dev | Variant Browser ▾ | Variant Analysis ▾ | Clinical Analysis ▾ | Catalog Metadata ▾ | GA4GH ▾ | 666 | Studies ▾ | About | demouser ▾

Projects / family / corpasome

Variant Browser

RUN

TABLE RESULT | AGGREGATION STATS | COMPARATOR

Filters | Aggregation

STUDY AND COHORTS

Studies Filter: Corpas Family

In any of (OR) | In all (AND)

GENOMIC

Chromosomal Location: 3:444-55555,1:1-100000

Feature IDs (gene, SNPs, ...): Search for Gene Symbols +

BRCA2,ENSG00000139618,ENST00000544455,rs28897700

Select SO terms: 10 items selected

Disease Panels

FILTERS: Biotype = protein_coding | Consequence Type 10

CLEAR | FILTERS ▾

COLUMNS ▾ | DOWNLOAD ▾

Showing 1-10 of 796 records

					Deleteriousness			Conservation			Population Frequencies		Phenotypes	
	Variant	dbSNP Id	Gene	Type	Consequence Type	SIFT	Polyphen	CADD	PhyloP	PhastCons	GERP	1000 Genomes	gnomAD Genomes	ClinVar	Cosmic
+	11:67441906 A/C	-	ALDH3B2	SNV	splice_donor_variant	-	-	9.70	-0.126	0.021	2.300	🟦🟦🟦🟦🟦🟦	🟦🟦🟦🟦🟦🟦⬛	✗	✗
+	11:67765164 -/G	-	UNC93B1	INDEL	frameshift_variant	-	-	-	0.459	0.996	4.710	🟦🟦🟦🟦🟦🟦	🟦🟦🟦🟦🟦🟦⬛	✗	✓
+	11:67786065 -/C	-	ALDH3B1	INDEL	frameshift_variant	-	-	-	0.449	0.714	2.770	🟦🟦🟦🟦🟦🟦	🟦🟦🟦🟦🟦🟦⬛	✗	✗
+	11:67789294 -/C	-	ALDH3B1 , RP5-901A4.1	INDEL	frameshift_variant	-	-	-	0.650	0.437	3.890	🟦🟦🟦🟦🟦🟦	🟦🟦🟦🟦🟦🟦⬛	✗	✗
+	11:67795379 -/C	-	NDUFS8 , ALDH3B1 , RP5-901A4.1	INDEL	splice_donor_variant splice_acceptor_variant	-	-	-	0.462	0.908	-1.880	🟦🟦🟦🟦🟦🟦	🟦🟦🟦🟦🟦🟦⬛	✗	✗
+	11:76954789 -/A	-	GDPD4	INDEL	frameshift_variant	-	-	-	-0.408	0.129	1.240	🟦🟦🟦🟦🟦🟦	🟦🟦🟦🟦🟦🟦⬛	✗	✗
+	11:77580840 C/T	-	AAMDC , RP11-91P24.6 , RP11-91P24.1 , RP11-91P24.7	SNP	stop_gained	-	-	40.00	0.650	0.998	4.550	🟥⬛⬛🟥⬛⬛	🟥⬛⬛⬛⬛⬛⬛	✗	✗
+	11:62495924 -/G	-	TTC9C , HNRNPUL2-BSCL2 , HNRNPUL2	INDEL	frameshift_variant	-	-	-	0.563	0.290	2.680	🟦🟦🟦🟦🟦🟦	🟦🟦🟦🟥🟦🟦⬛	✗	✓
+	11:62848487 A/C	-	SLC22A24	SNV	stop_gained	-	-	39.00	0.481	0.646	2.330	🟦🟦🟦🟦🟦🟦	🟦🟦🟦🟦🟦🟦⬛	✗	✓
+	11:62910849 T/C	-	SLC22A24 , SLC22A10	SNV	splice_donor_variant	-	-	16.76	-1.103	0.070	2.290	🟦🟦🟦🟦🟦🟦	🟦🟦🟦🟦🟦🟦⬛	✗	✗

Showing 1 to 10 of 796 rows 10 ▴ rows per page

‹ 1 2 3 4 5 ... 80 ›

Variant: 11:67441906:A:C

Summary | Consequence Type | Population Frequencies | Cohort Stats | Samples | Beacon | Reactome Pathways

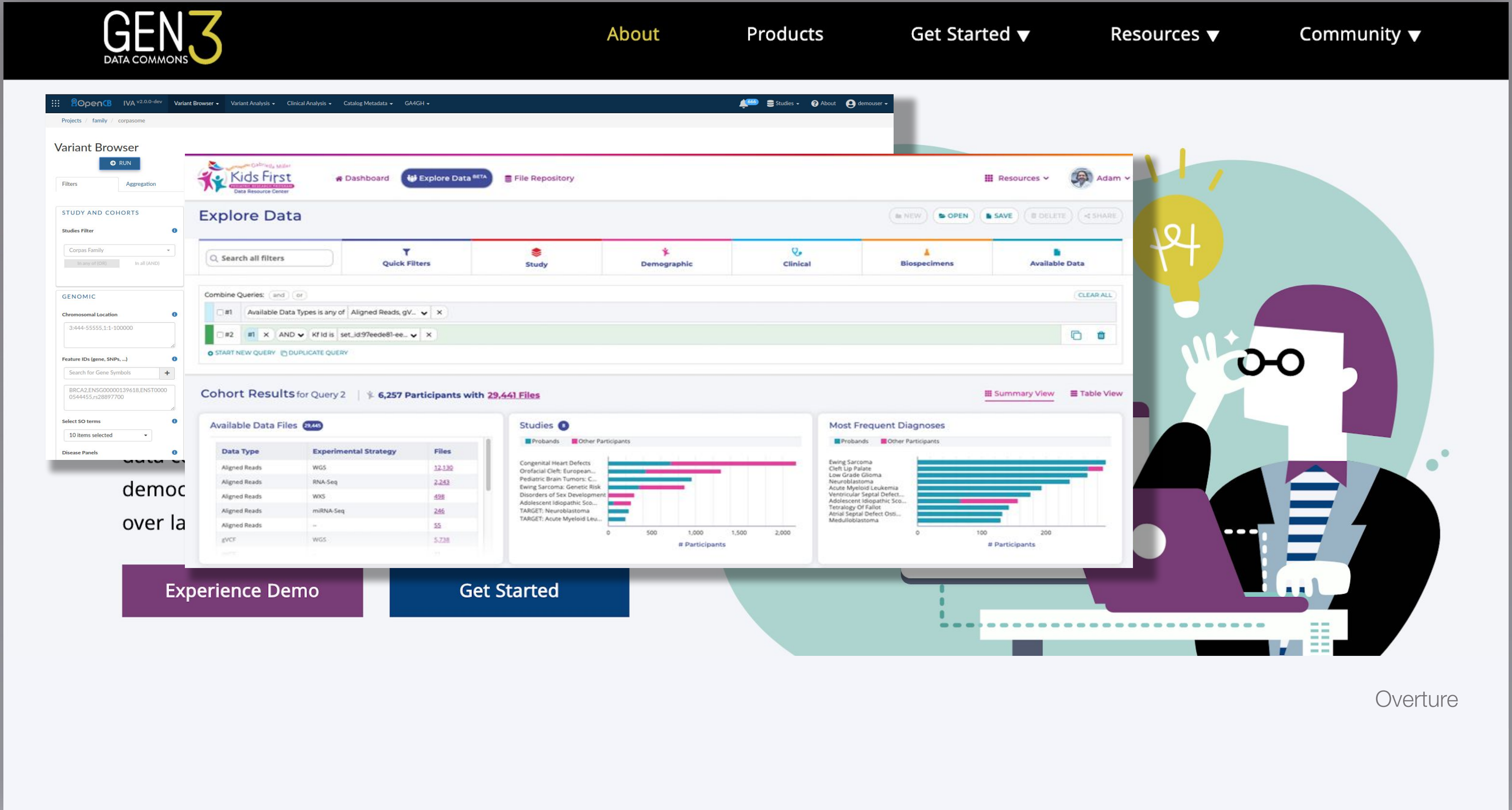
ID rs7947754

HGVS ENST00000530069(ENSG00000132746):c.-245+6741T>G

data commons. Data commons accelerate and democratize the process of scientific discovery, especially over large or complex datasets.

Experience Demo | Get Started

OpenCGA



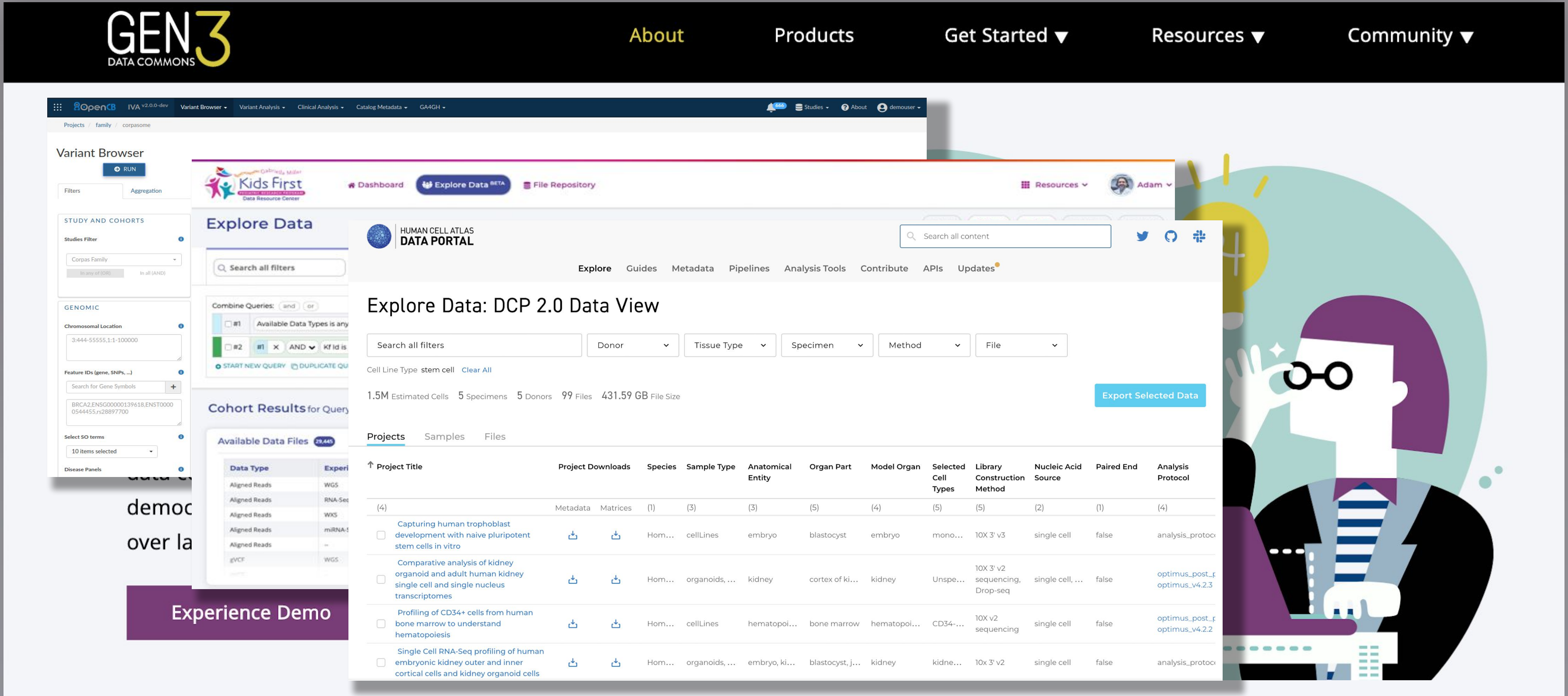
Data Type	Experimental Strategy	Files
Aligned Reads	WGS	12,130
Aligned Reads	RNA-Seq	2,243
Aligned Reads	WXS	498
Aligned Reads	miRNA-Seq	246
Aligned Reads	--	55
gVCF	WGS	5,738

data c...

democ...

over la...

Overture



Variant Browser

RUN

Filters | Aggregation

STUDY AND COHORTS

Studies Filter: Corpas Family (In any of (OR) / In all (AND))

GENOMIC

Chromosomal Location: 3:444-55555,1:1-100000

Feature IDs (gene, SNPs, ...): Search for Gene Symbols +

BRCA2,ENSG00000139618,ENST00000544455,rs28897700

Select SO terms: 10 items selected

Disease Panels

democ

over la

Experience Demo

Explore Data

Search all filters

Combine Queries: and or

□ #1 Available Data Types is any

□ #2 #1 × AND ⌄ Kf Id is

START NEW QUERY DUPLICATE QU

Cohort Results for Query

Available Data Files 29,445

Data Type	Experi
Aligned Reads	WGS
Aligned Reads	RNA-Seq
Aligned Reads	WXS
Aligned Reads	miRNA-S
Aligned Reads	--
gVCF	WGS

Explore Data: DCP 2.0 Data View

Search all filters | Donor | Tissue Type | Specimen | Method | File

Cell Line Type stem cell Clear All

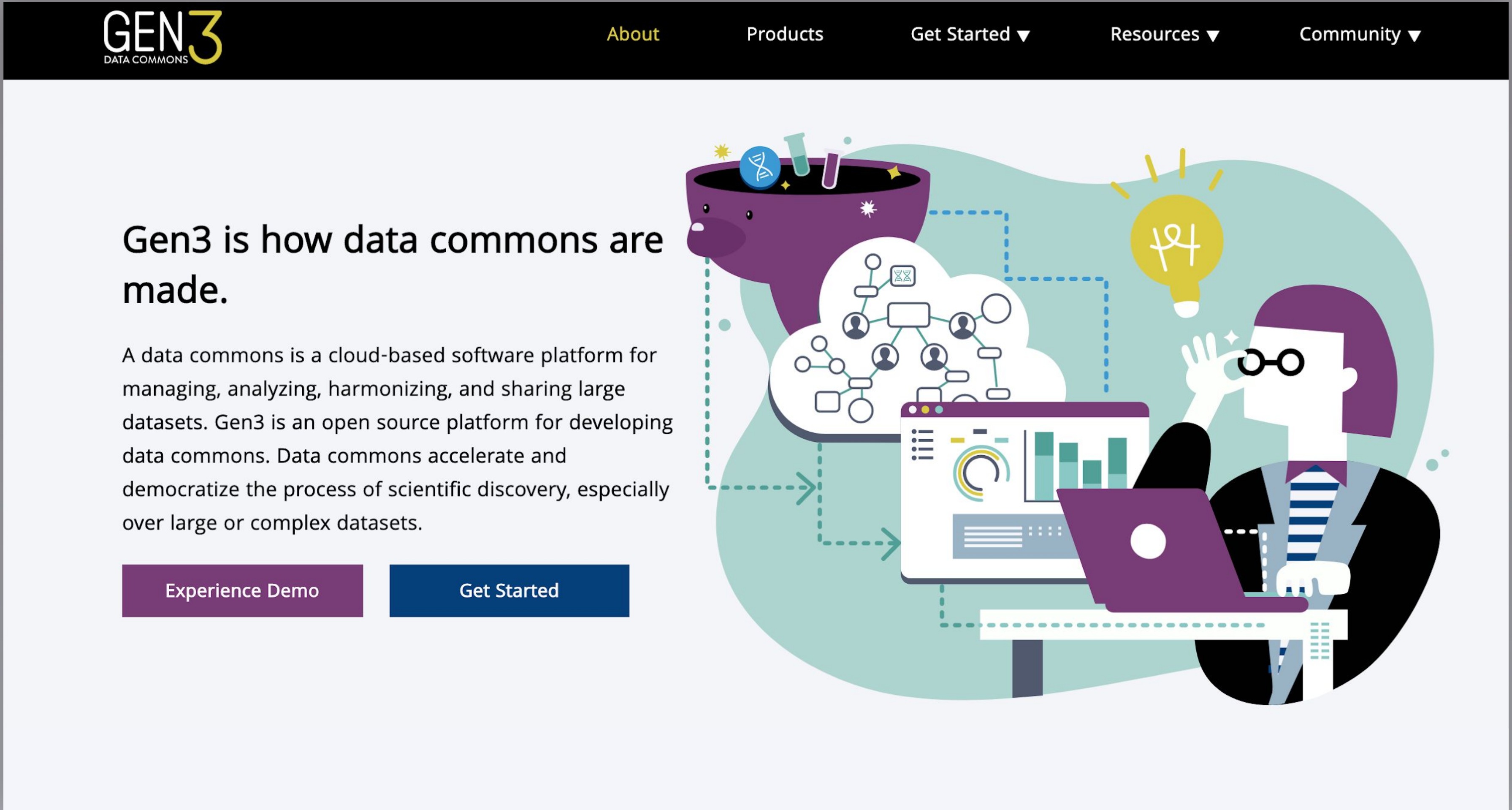
1.5M Estimated Cells 5 Specimens 5 Donors 99 Files 431.59 GB File Size

Export Selected Data

Projects | Samples | Files

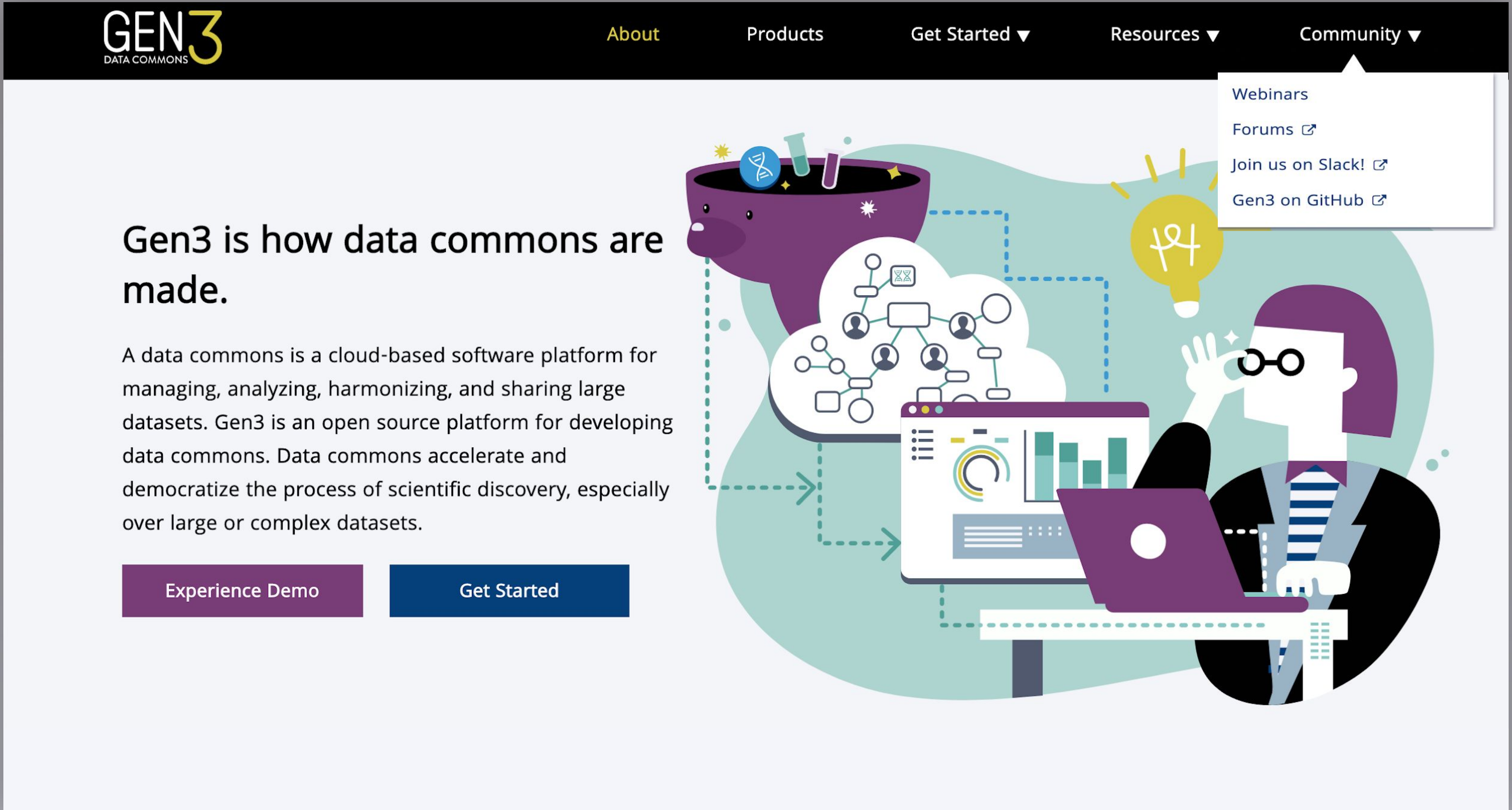
	Project Title	Project Downloads		Species	Sample Type	Anatomical Entity	Organ Part	Model Organ	Selected Cell Types	Library Construction Method	Nucleic Acid Source	Paired End	Analysis Protocol
	(4)	Metadata	Matrices	(1)	(3)	(3)	(5)	(4)	(5)	(5)	(2)	(1)	(4)
□	Capturing human trophoblast development with naive pluripotent stem cells in vitro	⤓	⤓	Hom...	cellLines	embryo	blastocyst	embryo	mono...	10X 3' v3	single cell	false	analysis_protoc
□	Comparative analysis of kidney organoid and adult human kidney single cell and single nucleus transcriptomes	⤓	⤓	Hom...	organoids, ...	kidney	cortex of ki...	kidney	Unspe...	10X 3' v2 sequencing, Drop-seq	single cell, ...	false	optimus_post_p optimus_v4.2.3
□	Profiling of CD34+ cells from human bone marrow to understand hematopoiesis	⤓	⤓	Hom...	cellLines	hematopoi...	bone marrow	hematopoi...	CD34-...	10X v2 sequencing	single cell	false	optimus_post_p optimus_v4.2.2
□	Single Cell RNA-Seq profiling of human embryonic kidney outer and inner cortical cells and kidney organoid cells	⤓	⤓	Hom...	organoids, ...	embryo, ki...	blastocyst, j...	kidney	kidne...	10x 3' v2	single cell	false	analysis_protoc

HCA-Store



Gen3 is how data commons are made.

A data commons is a cloud-based software platform for managing, analyzing, harmonizing, and sharing large datasets. Gen3 is an open source platform for developing data commons. Data commons accelerate and democratize the process of scientific discovery, especially over large or complex datasets.



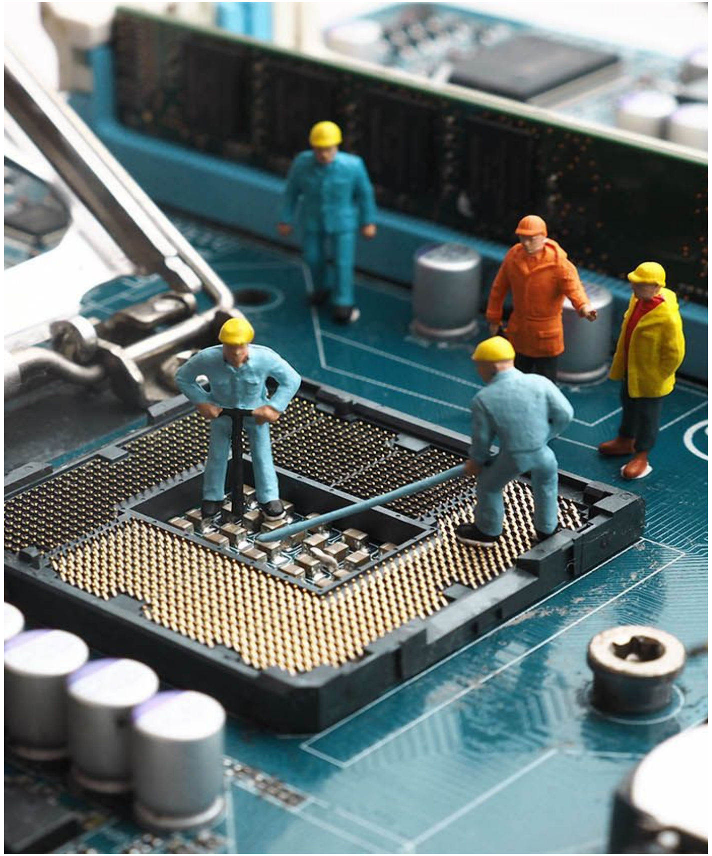
Gen3 is how data commons are made.

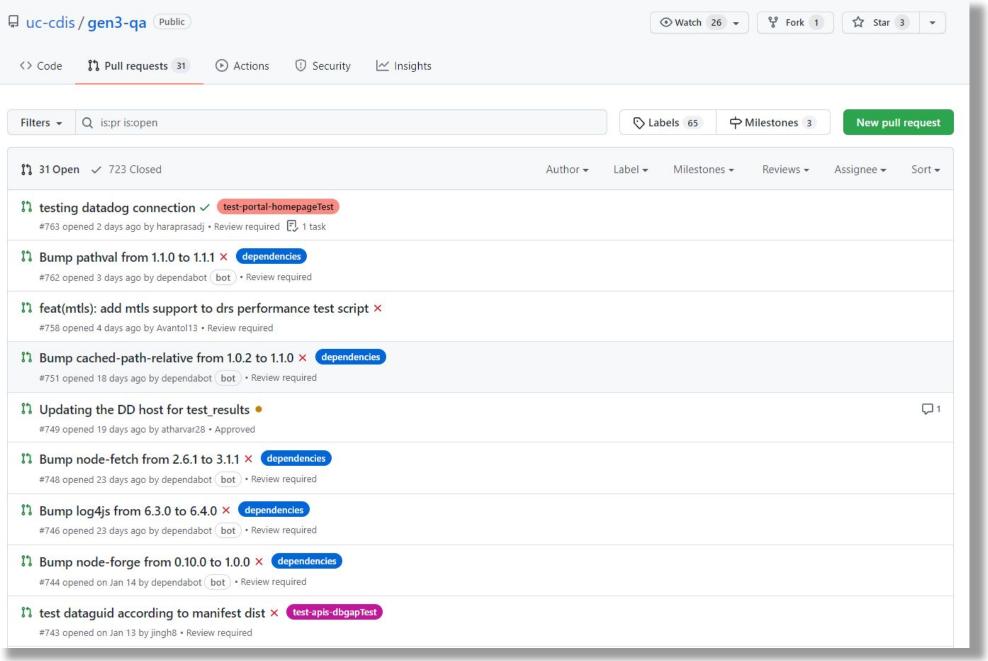
A data commons is a cloud-based software platform for managing, analyzing, harmonizing, and sharing large datasets. Gen3 is an open source platform for developing data commons. Data commons accelerate and democratize the process of scientific discovery, especially over large or complex datasets.

Actively Maintained Code

“A piece of software is being sustained if people are using it, fixing it, and improving it rather than replacing it.”

software-carpentry.org/blog/2014/08/sustainability.html



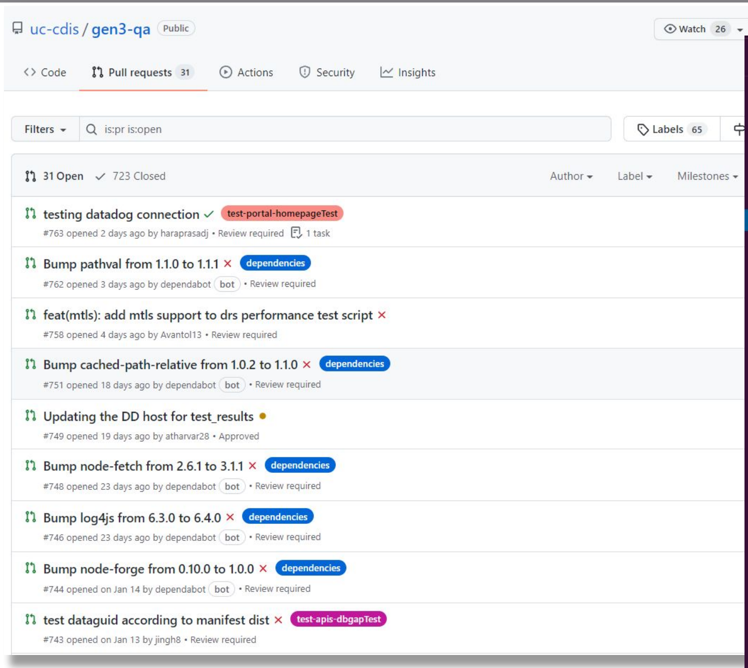


31 Open ✓ 723 Closed

- testing datadog connection ✓ test-portal-homepageTest
  #763 opened 2 days ago by haraprasadj • Review required 1 task
- Bump pathval from 1.1.0 to 1.1.1 ✕ dependencies
  #762 opened 3 days ago by dependabot bot • Review required
- feat(mtls): add mtls support to drs performance test script ✕
  #758 opened 4 days ago by Avantol13 • Review required
- Bump cached-path-relative from 1.0.2 to 1.1.0 ✕ dependencies
  #751 opened 18 days ago by dependabot bot • Review required
- Updating the DD host for test_results ● 1
  #749 opened 19 days ago by atharvar28 • Approved
- Bump node-fetch from 2.6.1 to 3.1.1 ✕ dependencies
  #748 opened 23 days ago by dependabot bot • Review required
- Bump log4js from 6.3.0 to 6.4.0 ✕ dependencies
  #746 opened 23 days ago by dependabot bot • Review required
- Bump node-forge from 0.10.0 to 1.0.0 ✕ dependencies
  #744 opened on Jan 14 by dependabot bot • Review required
- test dataguid according to manifest dist ✕ test-apis-dbgapTest
  #743 opened on Jan 13 by jingh8 • Review required

Active development on Github

github.com/uc-cdis



uc-cdis / gen3-qa Public

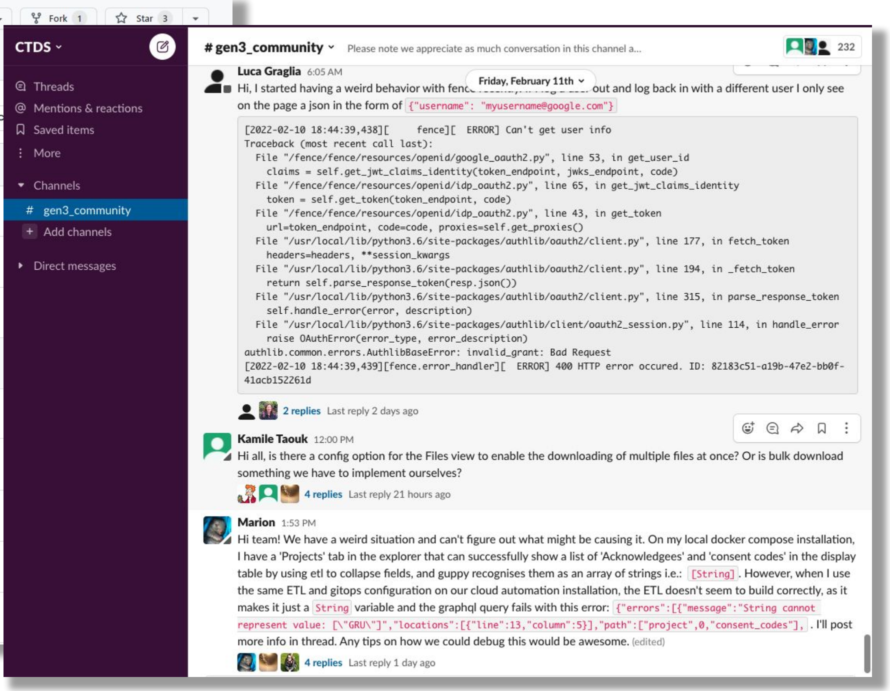
Watch 26 Fork 1 Star 3

Code | Pull requests 31 | Actions | Security | Insights

Filters is:pr is:open Labels 65

31 Open ✓ 723 Closed Author Label Milestones

- testing datadog connection ✓ test-portal-homepageTest
  #763 opened 2 days ago by haraprasadj • Review required 1 task
- Bump pathval from 1.1.0 to 1.1.1 ✕ dependencies
  #762 opened 3 days ago by dependabot bot • Review required
- feat(mtls): add mtls support to drs performance test script ✕
  #758 opened 4 days ago by Avantol13 • Review required
- Bump cached-path-relative from 1.0.2 to 1.1.0 ✕ dependencies
  #751 opened 18 days ago by dependabot bot • Review required
- Updating the DD host for test_results ●
  #749 opened 19 days ago by atharvar28 • Approved
- Bump node-fetch from 2.6.1 to 3.1.1 ✕ dependencies
  #748 opened 23 days ago by dependabot bot • Review required
- Bump log4js from 6.3.0 to 6.4.0 ✕ dependencies
  #746 opened 23 days ago by dependabot bot • Review required
- Bump node-forge from 0.10.0 to 1.0.0 ✕ dependencies
  #744 opened on Jan 14 by dependabot bot • Review required
- test dataguid according to manifest dist ✕ test-apis-dbgapTest
  #743 opened on Jan 13 by jingh8 • Review required



Luca Graglia 6:05 AM

Hi, I started having a weird behavior with fenc ... out and log back in with a different user I only see on the page a json in the form of `{"username": "myusername@google.com"}`

```
[2022-02-10 18:44:39,438][     fence][  ERROR] Can't get user info
Traceback (most recent call last):
  File "/fence/fence/resources/openid/google_oauth2.py", line 53, in get_user_id
    claims = self.get_jwt_claims_identity(token_endpoint, jwks_endpoint, code)
  File "/fence/fence/resources/openid/idp_oauth2.py", line 65, in get_jwt_claims_identity
    token = self.get_token(token_endpoint, code)
  File "/fence/fence/resources/openid/idp_oauth2.py", line 43, in get_token
    url=token_endpoint, code=code, proxies=self.get_proxies()
  File "/usr/local/lib/python3.6/site-packages/authlib/oauth2/client.py", line 177, in fetch_token
    headers=headers, **session_kwargs
  File "/usr/local/lib/python3.6/site-packages/authlib/oauth2/client.py", line 194, in _fetch_token
    return self.parse_response_token(resp.json())
  File "/usr/local/lib/python3.6/site-packages/authlib/oauth2/client.py", line 315, in parse_response_token
    self.handle_error(error, description)
  File "/usr/local/lib/python3.6/site-packages/authlib/client/oauth2_session.py", line 114, in handle_error
    raise OAuthError(error_type, error_description)
authlib.common.errors.AuthlibBaseError: invalid_grant: Bad Request
[2022-02-10 18:44:39,439][fence.error_handler][  ERROR] 400 HTTP error occured. ID: 82183c51-a19b-47e2-bb0f-
41acb152261d
```

2 replies Last reply 2 days ago

Kamile Taouk 12:00 PM

Hi all, is there a config option for the Files view to enable the downloading of multiple files at once? Or is bulk download something we have to implement ourselves?

4 replies Last reply 21 hours ago

Marion 1:53 PM

Hi team! We have a weird situation and can't figure out what might be causing it. On my local docker compose installation, I have a 'Projects' tab in the explorer that can successfully show a list of 'Acknowledgees' and 'consent codes' in the display table by using etl to collapse fields, and guppy recognises them as an array of strings i.e.: `[String]`. However, when I use the same ETL and gitops configuration on our cloud automation installation, the ETL doesn't seem to build correctly, as it makes it just a `String` variable and the graphql query fails with this error: `{"errors":[{"message":"String cannot represent value: [\"GRU\"]","locations":[{"line":13,"column":5}],"path":["project",0,"consent_codes"],` . I'll post more info in thread. Any tips on how we could debug this would be awesome. (edited)

4 replies Last reply 1 day ago

Lively Slack channel

ACCOUNT

BloodPAC
BLOOD PROFILING ATLAS IN CANCER

CANINE
Data Commons

CHICAGOLAND COVID-19 COMMONS

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

Environmental
Data Commons

GenoMEL
the Melanoma Genetics Consortium

BDGC

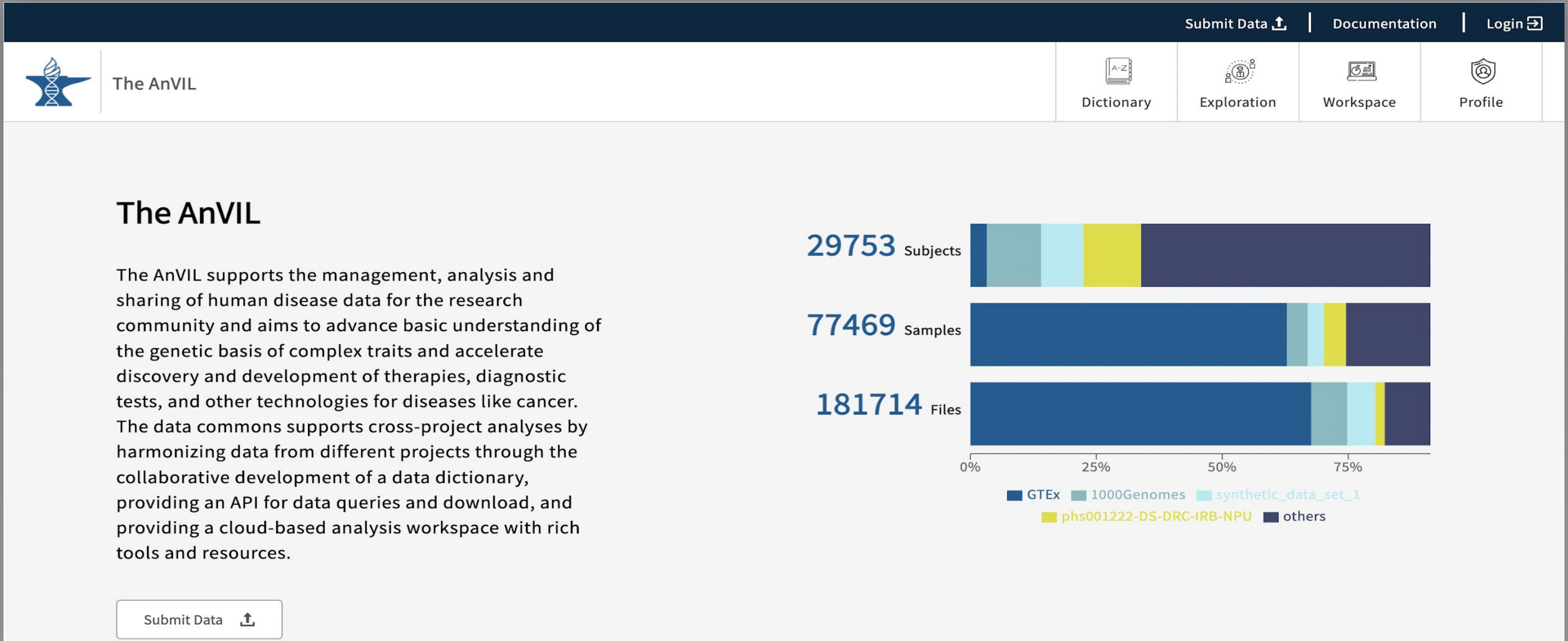
Gabriella Miller
Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER.

NIH National Institute of Allergy and Infectious Diseases
AccessClinicalData@NIAID

NIH National Heart, Lung, and Blood Institute
BioData CATALYST

VPO Veterans Precision Oncology Data Commons

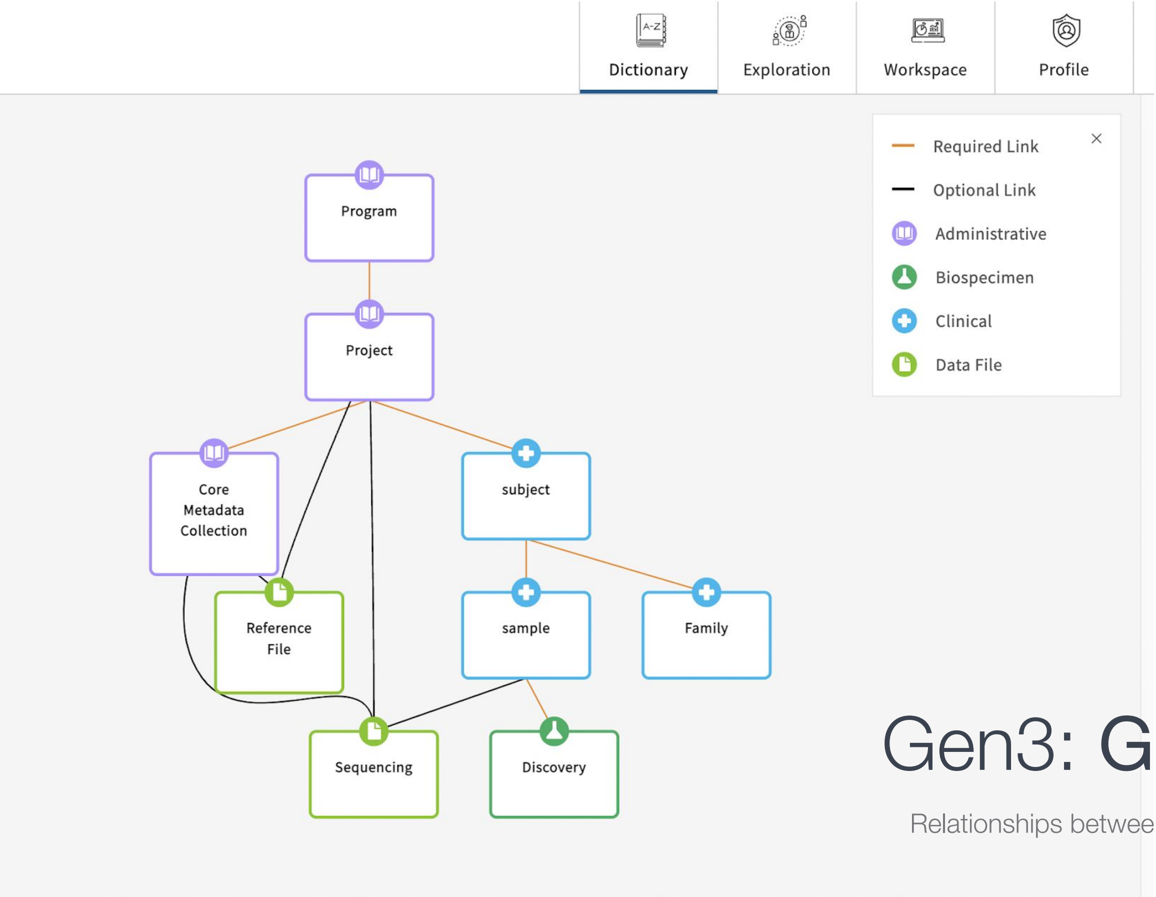


The AnVIL

The AnVIL supports the management, analysis and sharing of human disease data for the research community and aims to advance basic understanding of the genetic basis of complex traits and accelerate discovery and development of therapies, diagnostic tests, and other technologies for diseases like cancer. The data commons supports cross-project analyses by harmonizing data from different projects through the collaborative development of a data dictionary, providing an API for data queries and download, and providing a cloud-based analysis workspace with rich tools and resources.

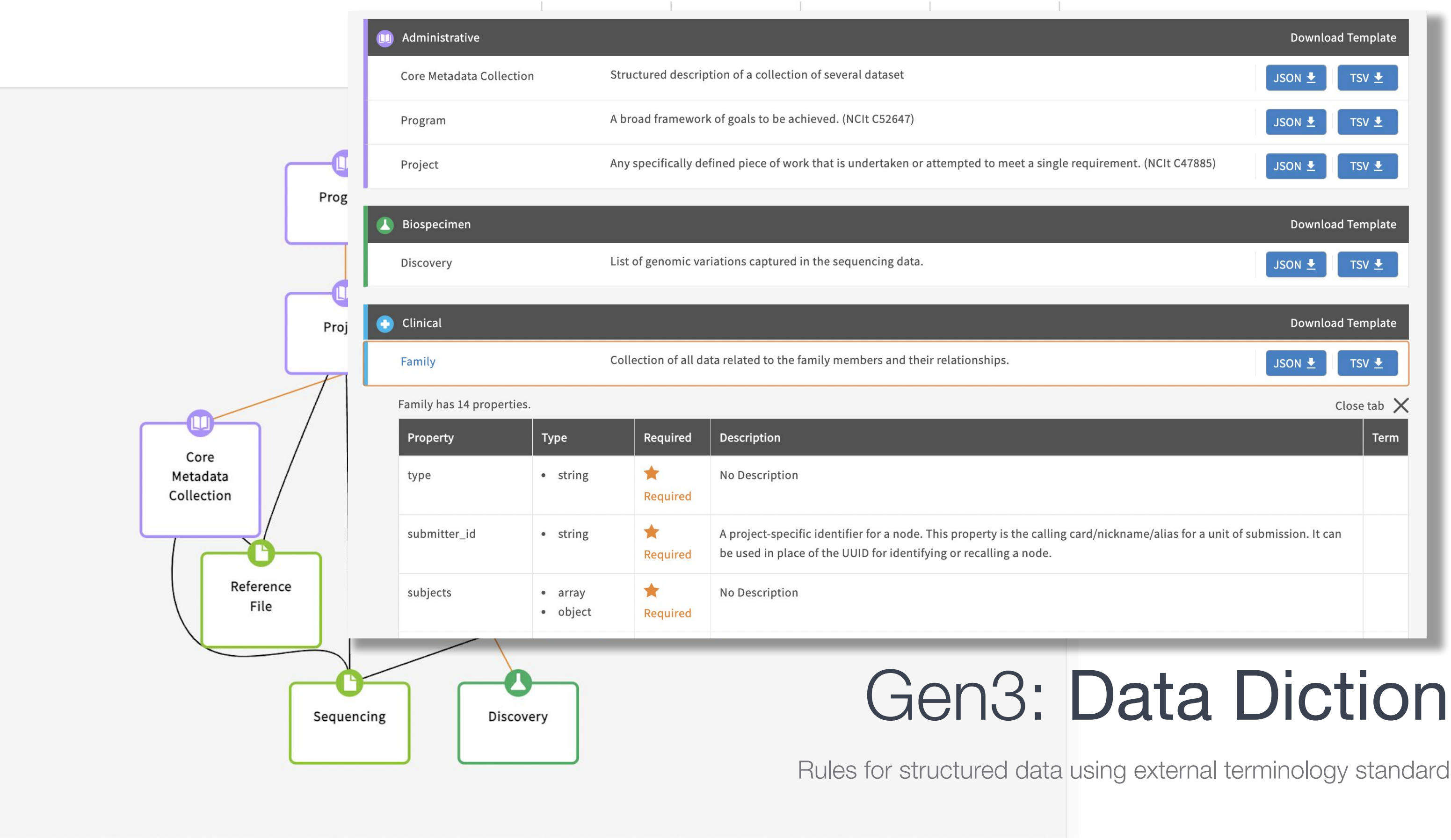
NHGRI Analysis Visualization & Informatics Lab-space

gen3.theanvil.io/



Gen3: Graph Data Model

Relationships between subjects, clinical, biological and molecular data



Administrative		Download Template
Core Metadata Collection	Structured description of a collection of several dataset	JSON ⬇ TSV ⬇
Program	A broad framework of goals to be achieved. (NCIt C52647)	JSON ⬇ TSV ⬇
Project	Any specifically defined piece of work that is undertaken or attempted to meet a single requirement. (NCIt C47885)	JSON ⬇ TSV ⬇

Biospecimen		Download Template
Discovery	List of genomic variations captured in the sequencing data.	JSON ⬇ TSV ⬇

Clinical		Download Template
Family	Collection of all data related to the family members and their relationships.	JSON ⬇ TSV ⬇

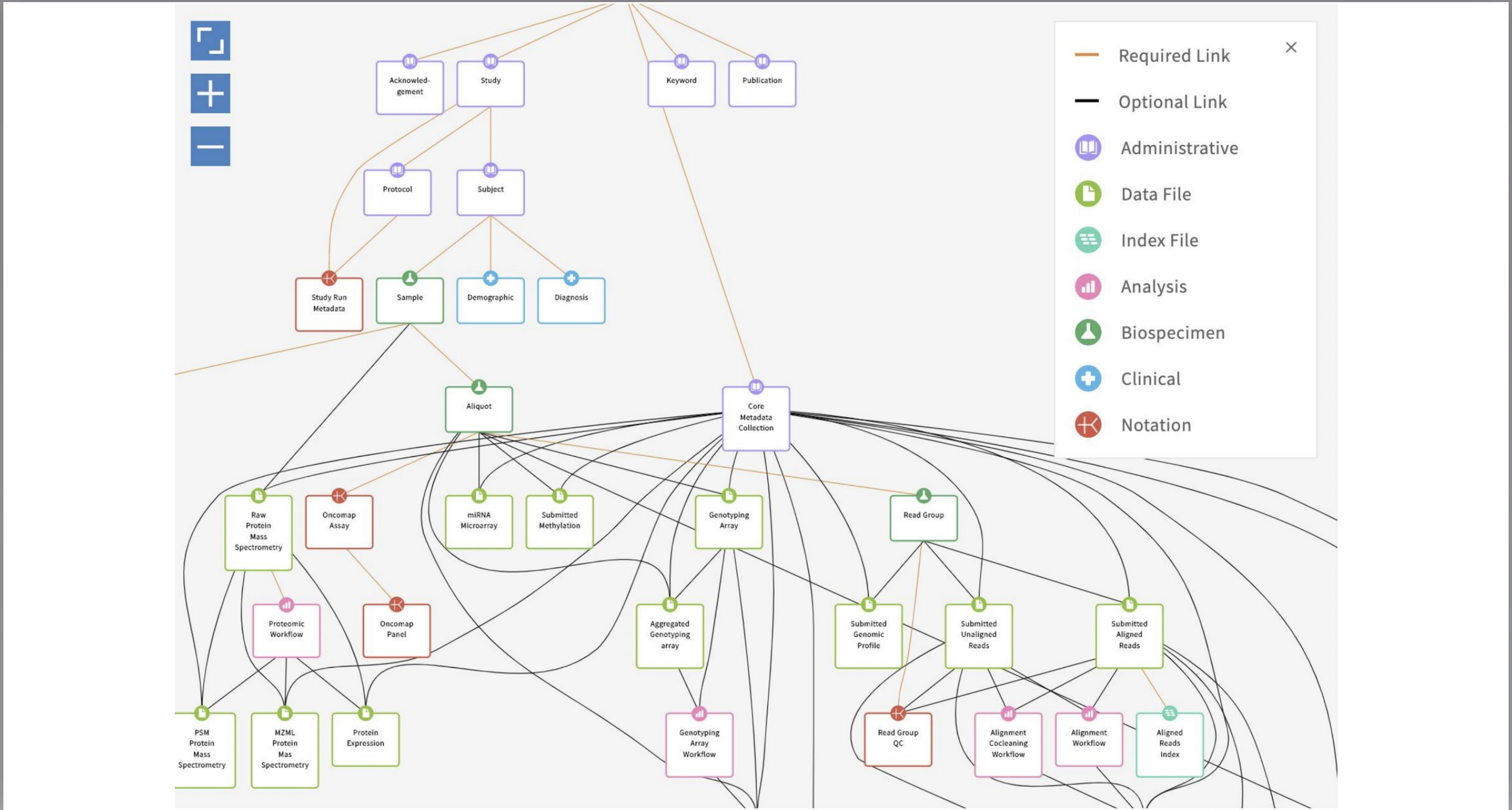
Family has 14 properties.

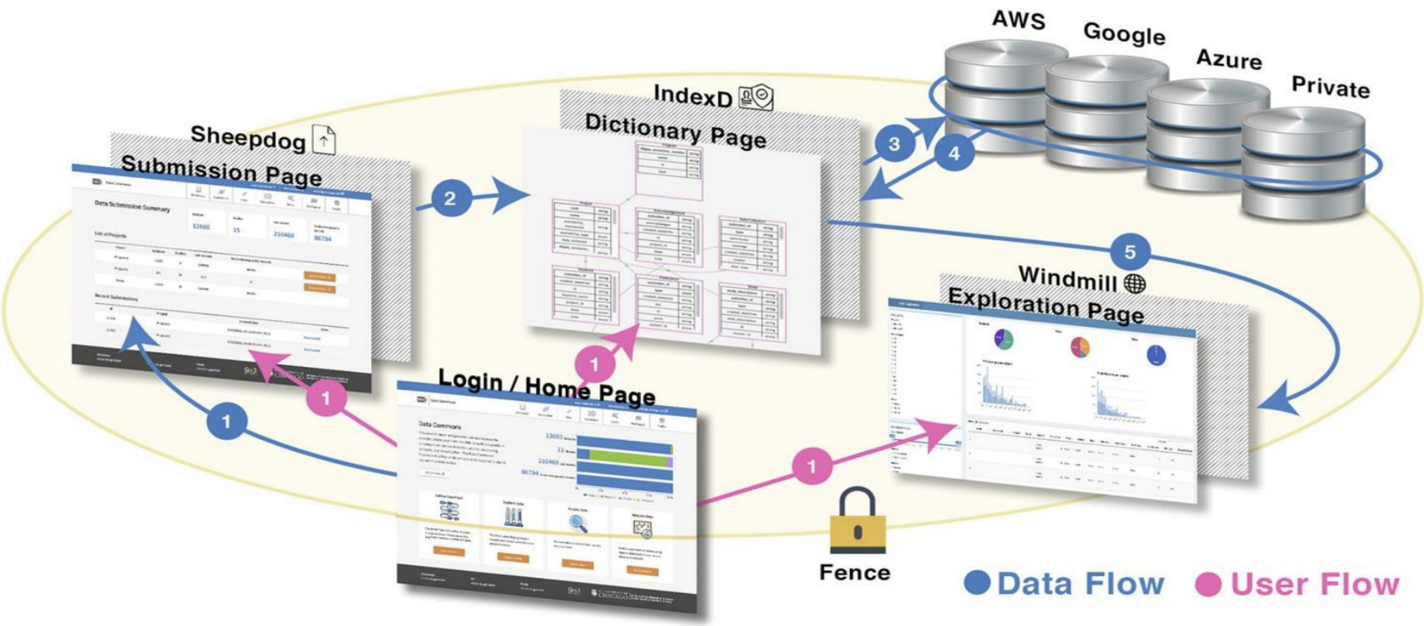
Close tab ✕

Property	Type	Required	Description	Term
type	• string	★ Required	No Description	
submitter_id	• string	★ Required	A project-specific identifier for a node. This property is the calling card/nickname/alias for a unit of submission. It can be used in place of the UUID for identifying or recalling a node.	
subjects	• array • object	★ Required	No Description	

Gen3: Data Dictionaries

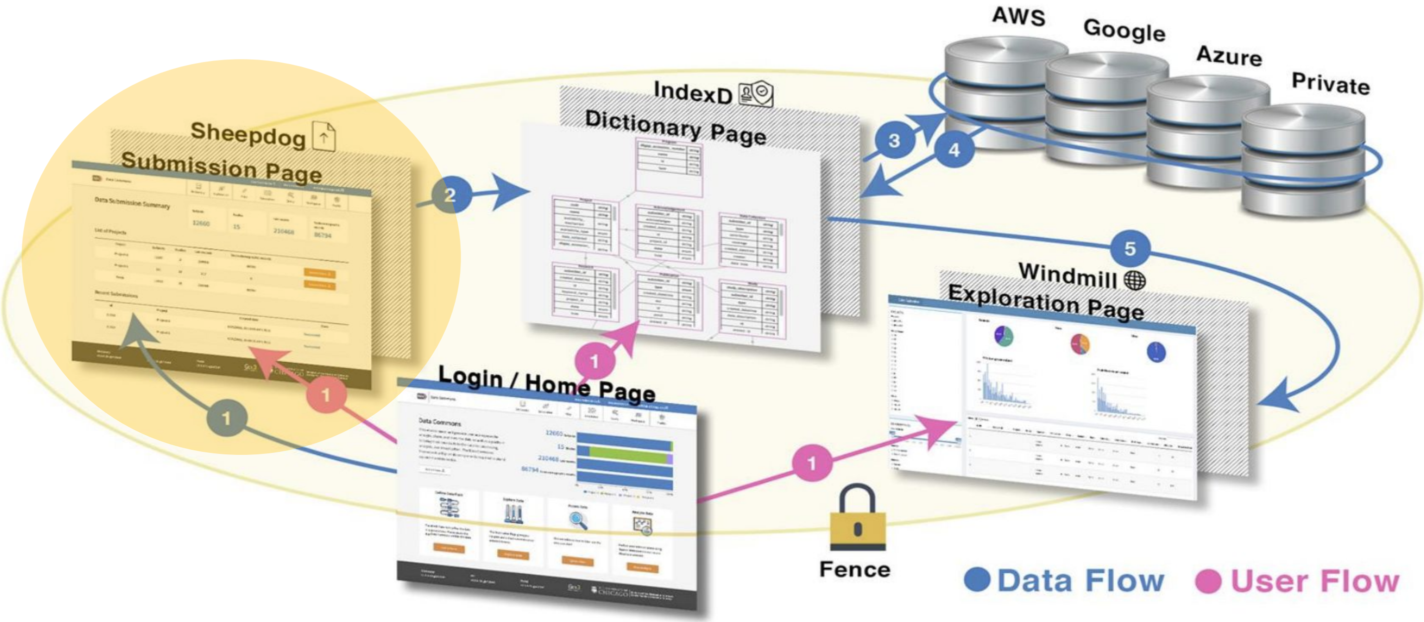
Rules for structured data using external terminology standards / ontologies


Acknowled-gement
Study
Keyword
Publication
Protocol
Subject
Study Run Metadata
Sample
Demographic
Diagnosis
Aliquot
Core Metadata Collection
Raw Protein Mass Spectrometry
Oncomap Assay
miRNA Microarray
Submitted Methylation
Genotyping Array
Read Group
Proteomic Workflow
Oncomap Panel
Aggregated Genotyping array
Submitted Genomic Profile
Submitted Unaligned Reads
Submitted Aligned Reads
PSM Protein Mass Spectrometry
MZML Protein Mas Spectrometry
Protein Expression
Genotyping Array Workflow
Read Group QC
Alignment Cocleaning Workflow
Alignment Workflow
Aligned Reads Index
Required Link
Optional Link
Administrative
Data File
Index File
Analysis
Biospecimen
Clinical
Notation



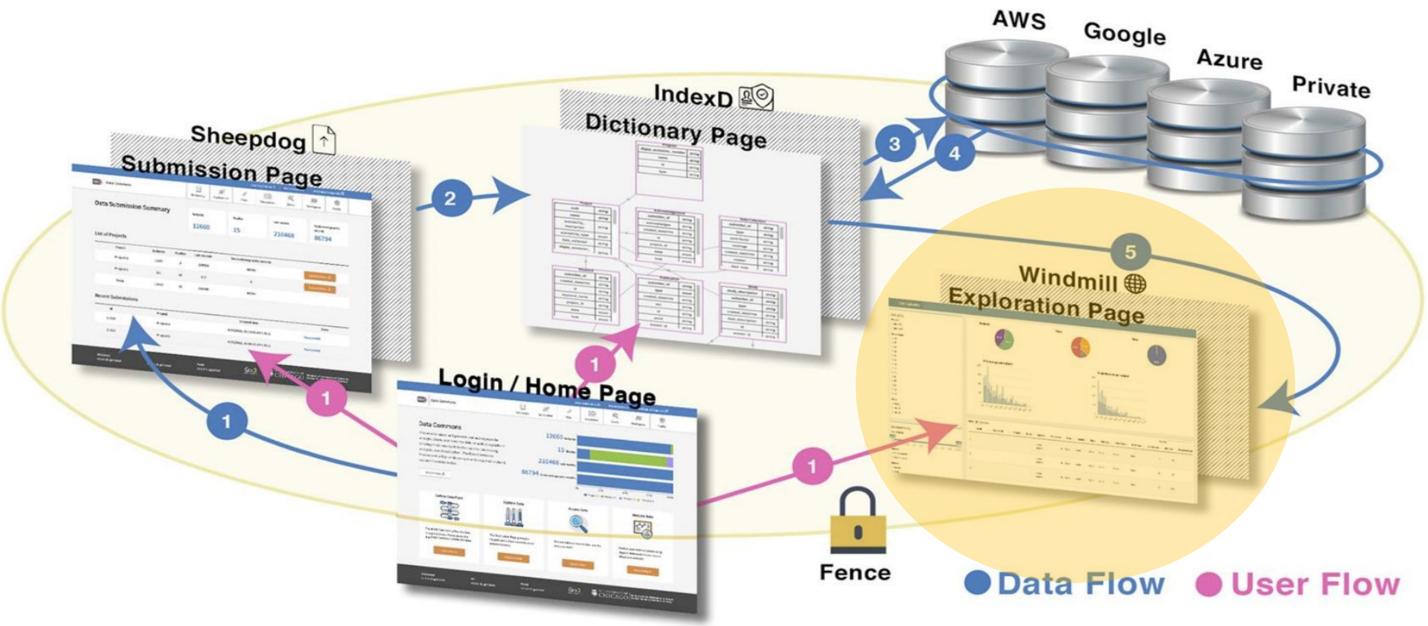
Gen3: Microservices

Modular components with defined interfaces



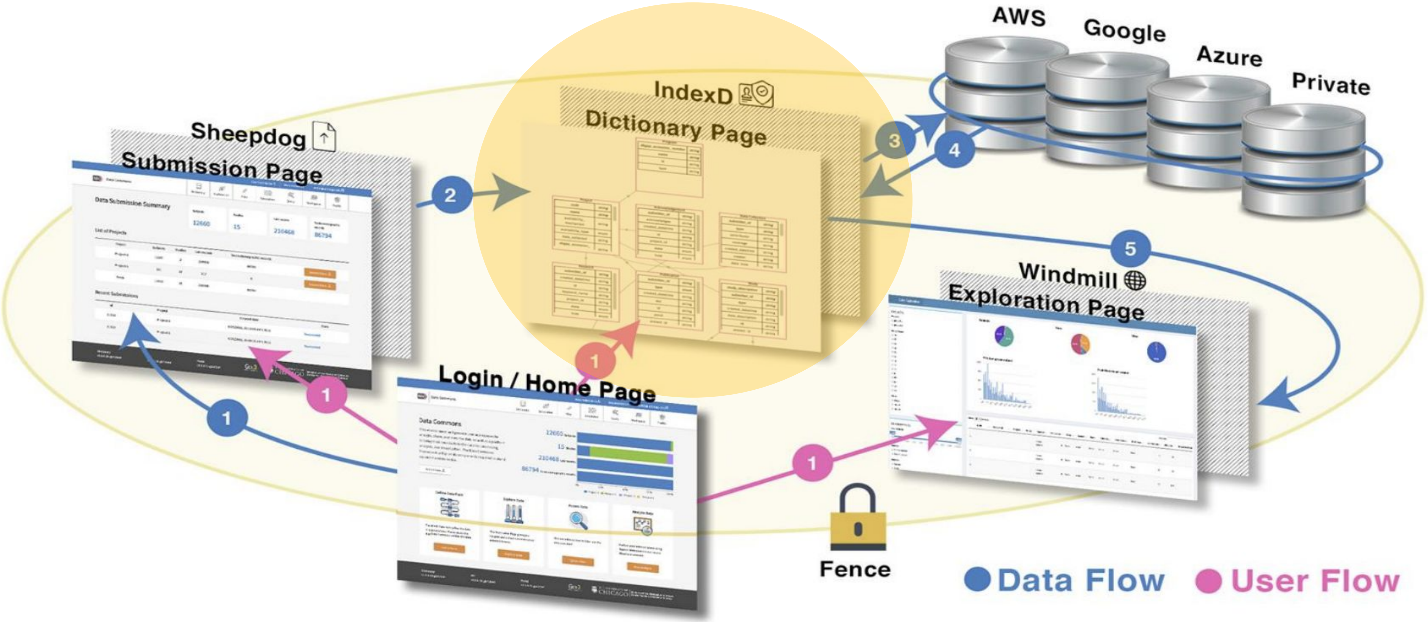
Gen3: Sheepdog

Data ingestion and validation service (UI and API)



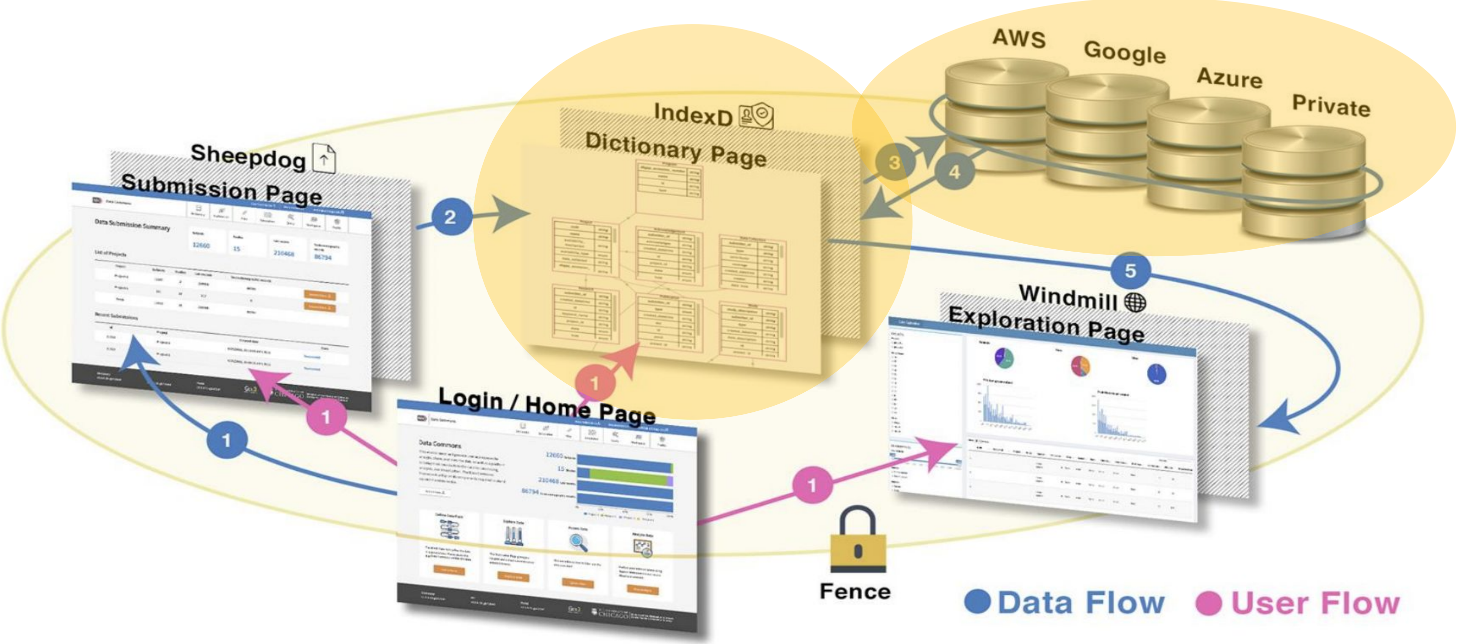
Gen3: Windmill

Web portal for data submission, query, exploration, and analysis



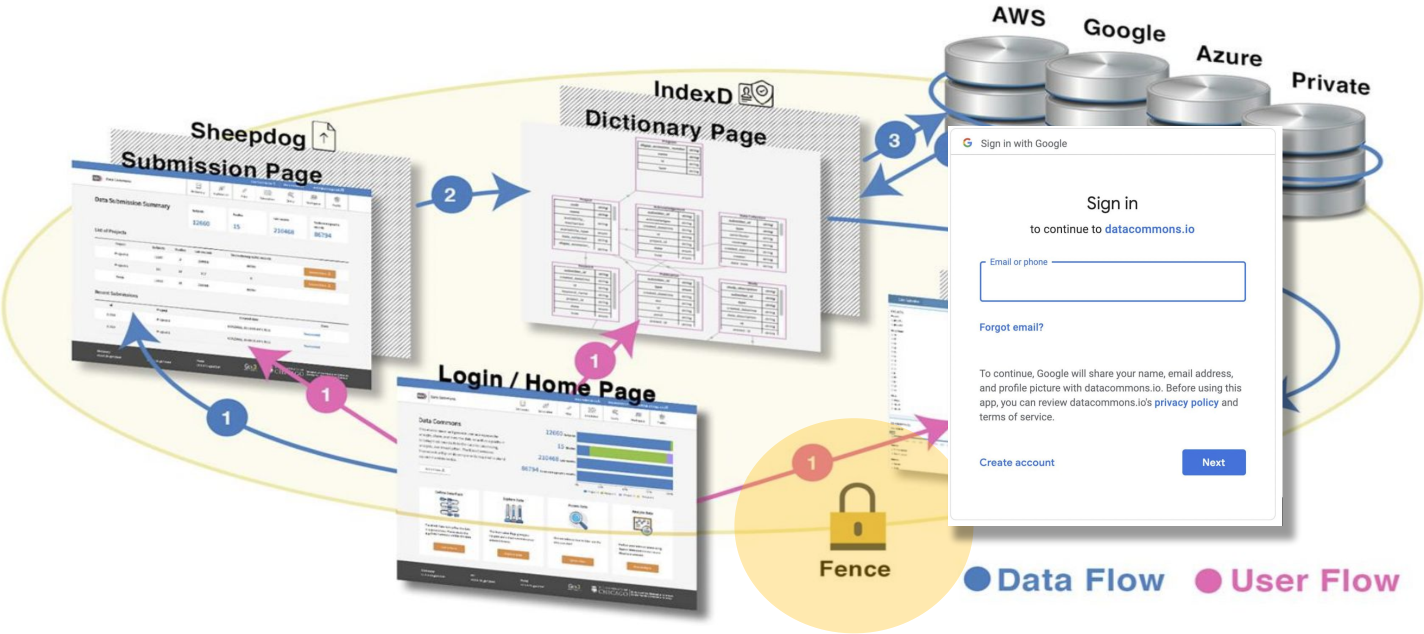
Gen3: IndexD

ID management, checksum and size catalogue



Gen3: IndexD

Supports multiple URLs for stored objects



Sign in with Google

Sign in

to continue to datacommons.io

Email or phone

Forgot email?

To continue, Google will share your name, email address, and profile picture with datacommons.io. Before using this app, you can review datacommons.io's privacy policy and terms of service.

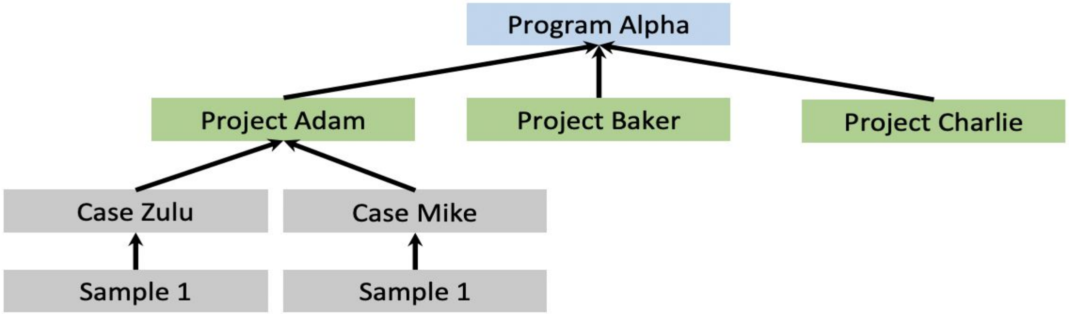
Create account

Next

Gen3: Fence

Authentication and Authorisation – OpenID Connect with support for Google, eRA Commons, eduGain, …

Gen3 Auth

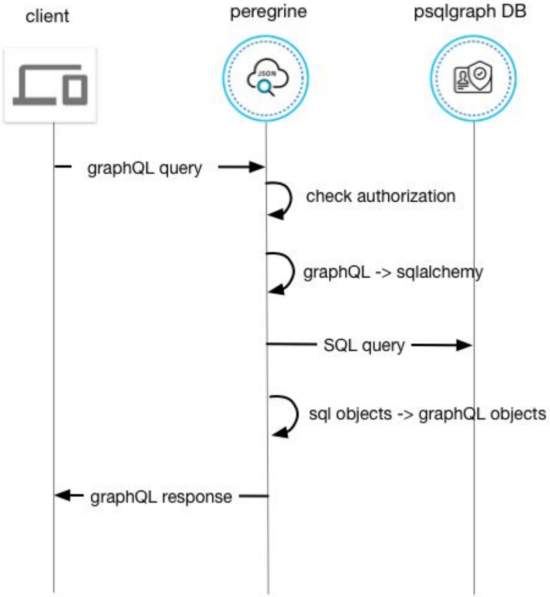


Gen3: DAC

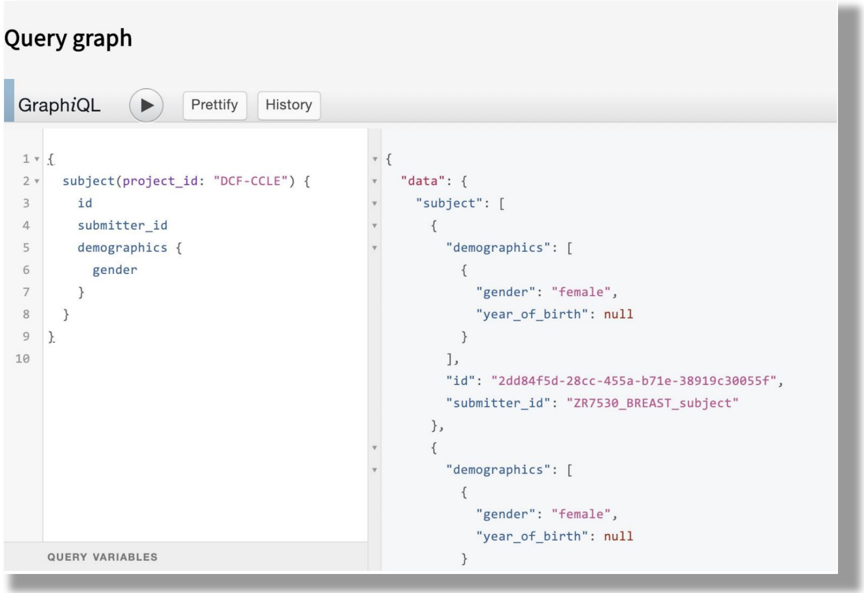
Role-Based Data Access Control (RBAC) engine

Gen3: Peregrine

Graph-based metadata queries

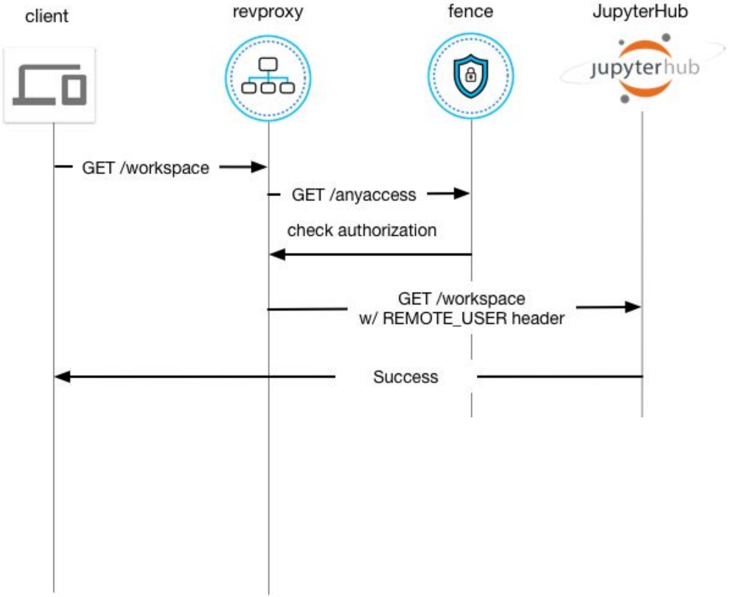


Gen3 rich query

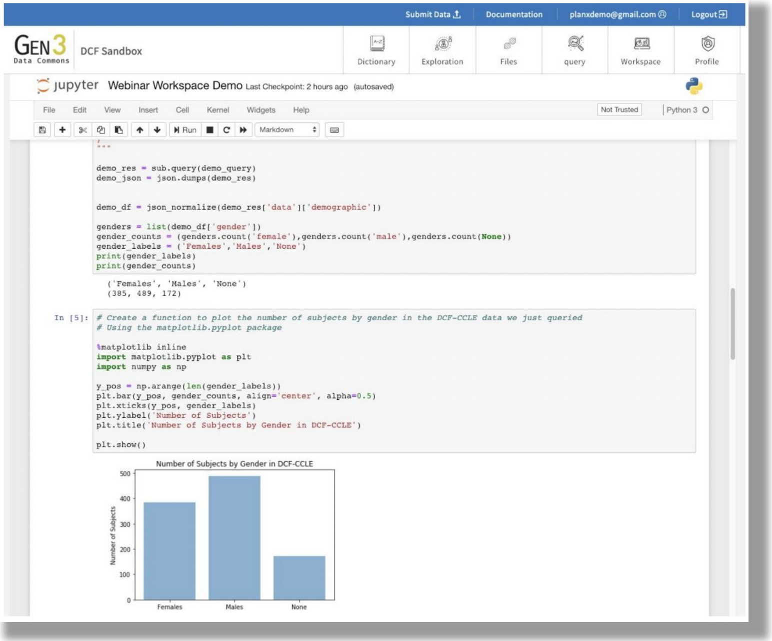


```
{
  subject(project_id: "DCF-CCLE") {
    id
    submitter_id
    demographics {
      gender
    }
  }
}
```

```
{
  "data": {
    "subject": [
      {
        "demographics": [
          {
            "gender": "female",
            "year_of_birth": null
          }
        ],
        "id": "2dd84f5d-28cc-455a-b71e-38919c30055f",
        "submitter_id": "ZR7530_BREAST_subject"
      },
      {
        "demographics": [
          {
            "gender": "female",
            "year_of_birth": null
          }
```



Gen3 Lightweight Workspaces with JupyterHub



```
"""

demo_res = sub.query(demo_query)
demo_json = json.dumps(demo_res)


demo_df = json_normalize(demo_res['data']['demographic'])

genders = list(demo_df['gender'])
gender_counts = (genders.count('female'),genders.count('male'),genders.count(None))
gender_labels = ('Females','Males','None')
print(gender_labels)
print(gender_counts)
```

```
('Females', 'Males', 'None')
(385, 489, 172)
```

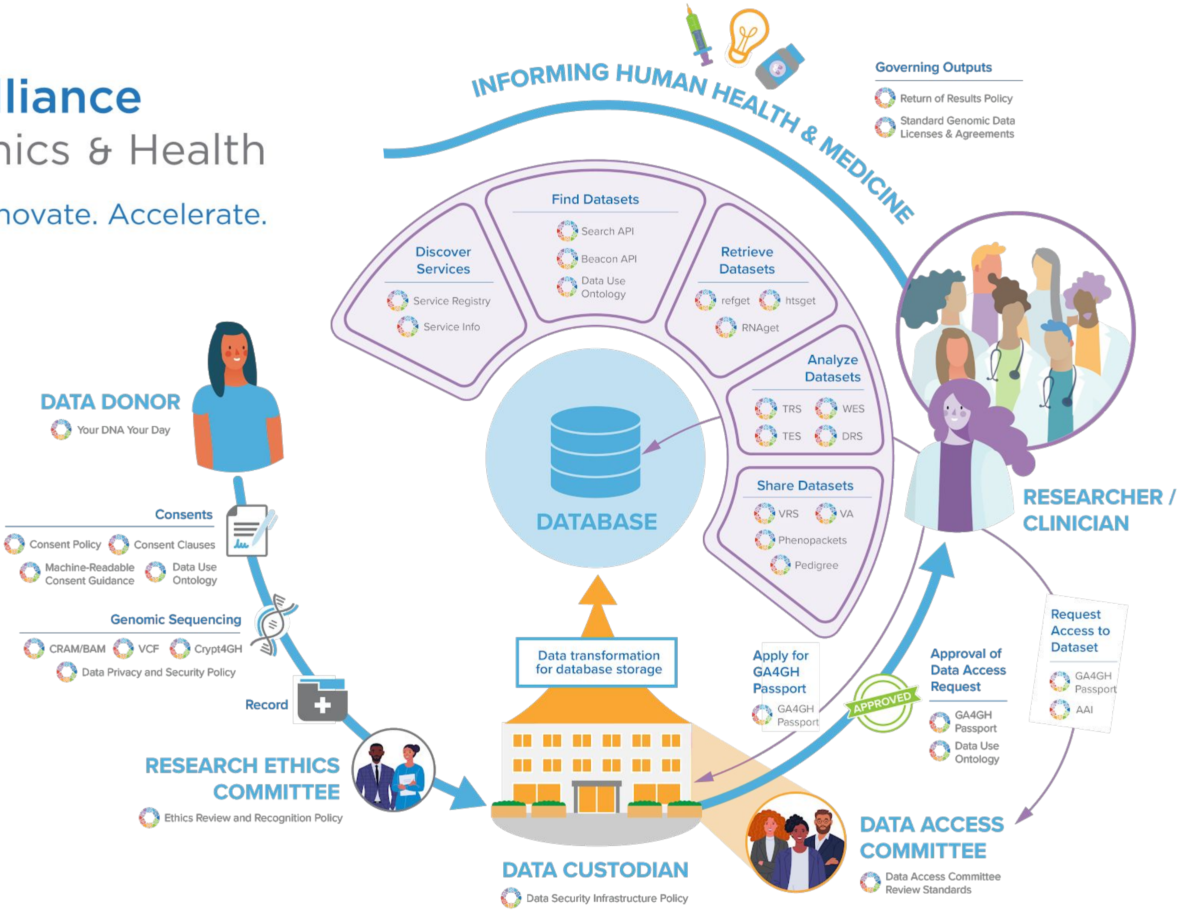
```
In [5]: # Create a function to plot the number of subjects by gender in the DCF-CCLE data we just queried
        # Using the matplotlib.pyplot package

        %matplotlib inline
        import matplotlib.pyplot as plt
        import numpy as np

        y_pos = np.arange(len(gender_labels))
        plt.bar(y_pos, gender_counts, align='center', alpha=0.5)
        plt.xticks(y_pos, gender_labels)
        plt.ylabel('Number of Subjects')
        plt.title('Number of Subjects by Gender in DCF-CCLE')

        plt.show()
```

Gen3: **Lightweight Workspaces**

Basic support for Jupyter notebooks for analysis and visualization in R, Python

# Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.



INFORMING HUMAN HEALTH & MEDICINE

Governing Outputs
- Return of Results Policy
- Standard Genomic Data Licenses & Agreements

DATA DONOR
- Your DNA Your Day

Consents
- Consent Policy
- Consent Clauses
- Machine-Readable Consent Guidance
- Data Use Ontology

Genomic Sequencing
- CRAM/BAM
- VCF
- Crypt4GH
- Data Privacy and Security Policy

Record

RESEARCH ETHICS COMMITTEE
- Ethics Review and Recognition Policy

Discover Services
- Service Registry
- Service Info

Find Datasets
- Search API
- Beacon API
- Data Use Ontology

Retrieve Datasets
- refget
- htsget
- RNAget

Analyze Datasets
- TRS
- WES
- TES
- DRS

Share Datasets
- VRS
- VA
- Phenopackets
- Pedigree

DATABASE

Data transformation for database storage

DATA CUSTODIAN
- Data Security Infrastructure Policy

Apply for GA4GH Passport
- GA4GH Passport

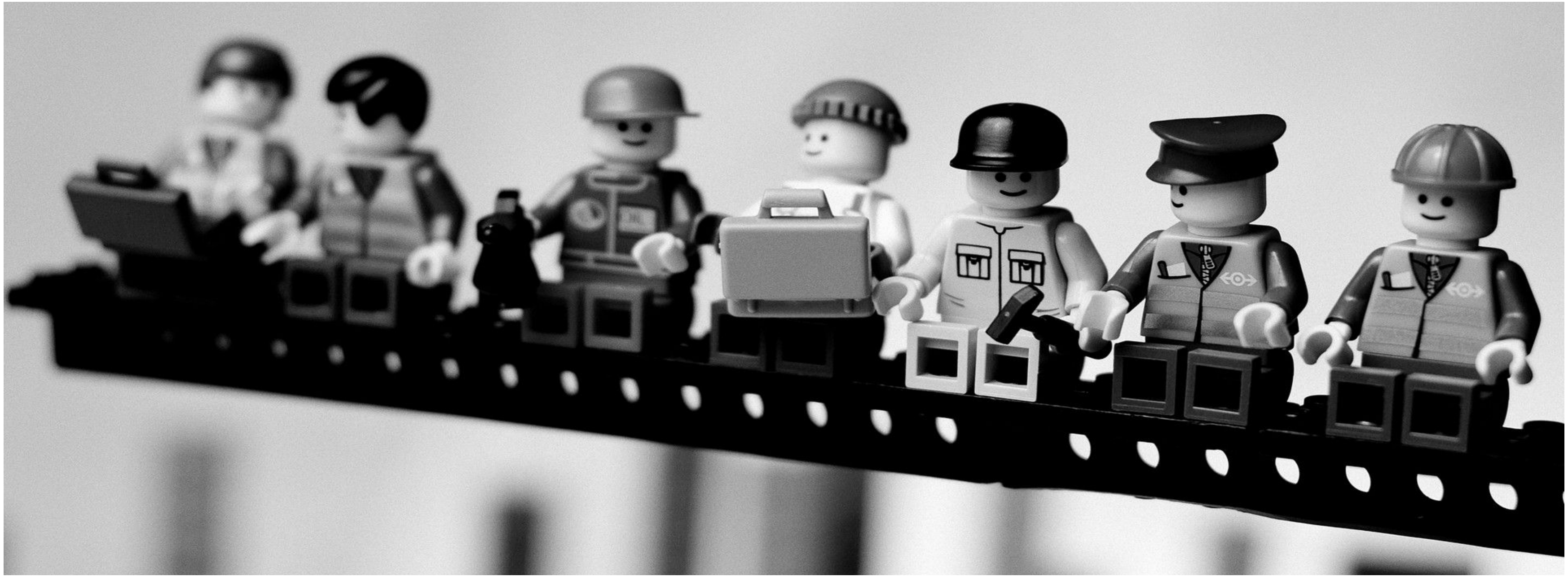
APPROVED

Approval of Data Access Request
- GA4GH Passport
- Data Use Ontology

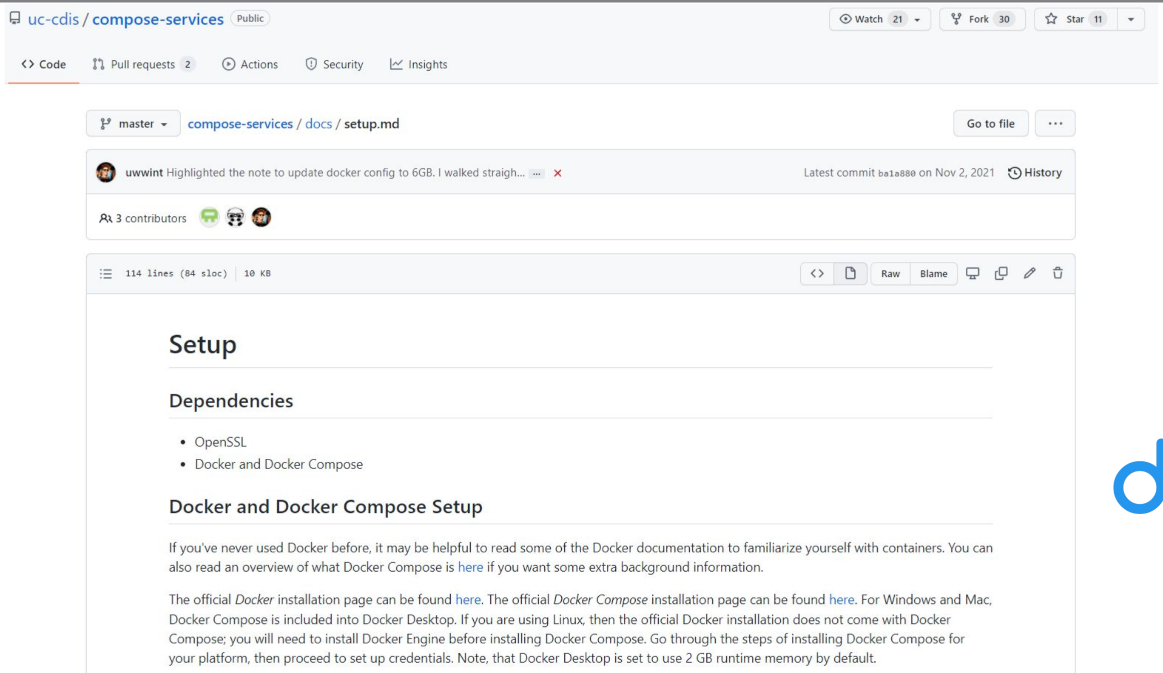
Request Access to Dataset
- GA4GH Passport
- AAI

DATA ACCESS COMMITTEE
- Data Access Committee Review Standards

RESEARCHER / CLINICIAN



Setting up Gen3



uc-cdis / compose-services Public
Watch 21
Fork 30
Star 11
Code
Pull requests 2
Actions
Security
Insights
master
compose-services / docs / setup.md
Go to file
uwwint Highlighted the note to update docker config to 6GB. I walked straigh... ...
Latest commit ba1a880 on Nov 2, 2021
History
3 contributors
114 lines (84 sloc) | 10 KB
Raw
Blame

Setup

Dependencies

- OpenSSL
- Docker and Docker Compose

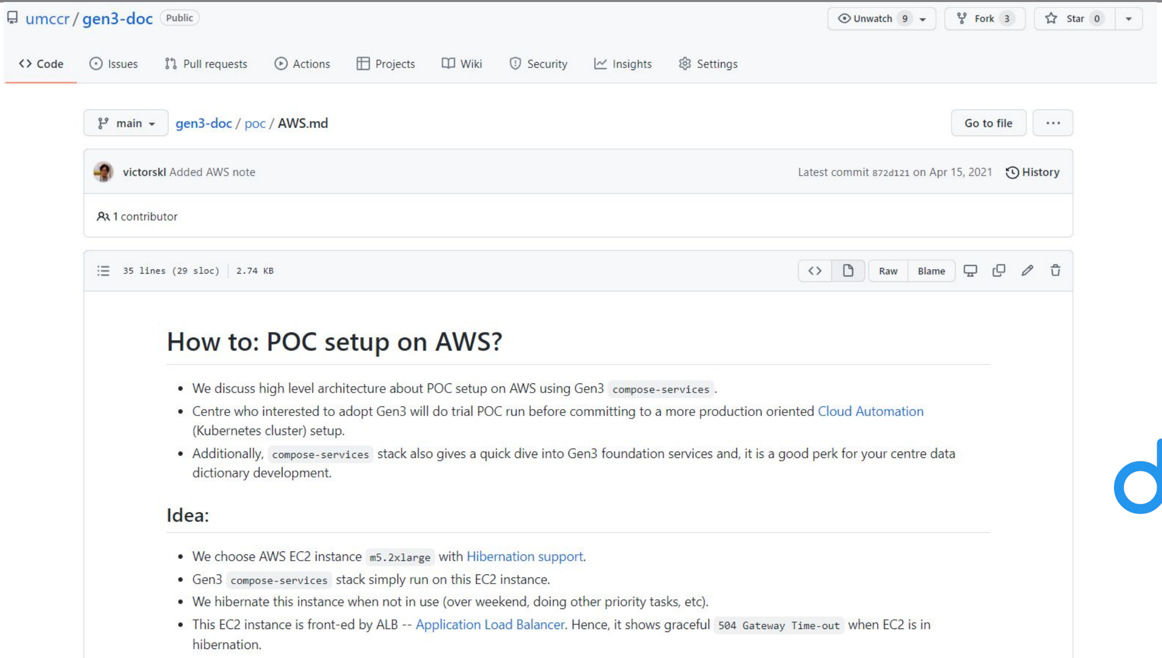
Docker and Docker Compose Setup

If you've never used Docker before, it may be helpful to read some of the Docker documentation to familiarize yourself with containers. You can also read an overview of what Docker Compose is here if you want some extra background information.

The official *Docker* installation page can be found here. The official *Docker Compose* installation page can be found here. For Windows and Mac, Docker Compose is included into Docker Desktop. If you are using Linux, then the official Docker installation does not come with Docker Compose; you will need to install Docker Engine before installing Docker Compose. Go through the steps of installing Docker Compose for your platform, then proceed to set up credentials. Note, that Docker Desktop is set to use 2 GB runtime memory by default.

Quick: Compose-Services

github.com/uc-cdis/compose-services



How to: POC setup on AWS?

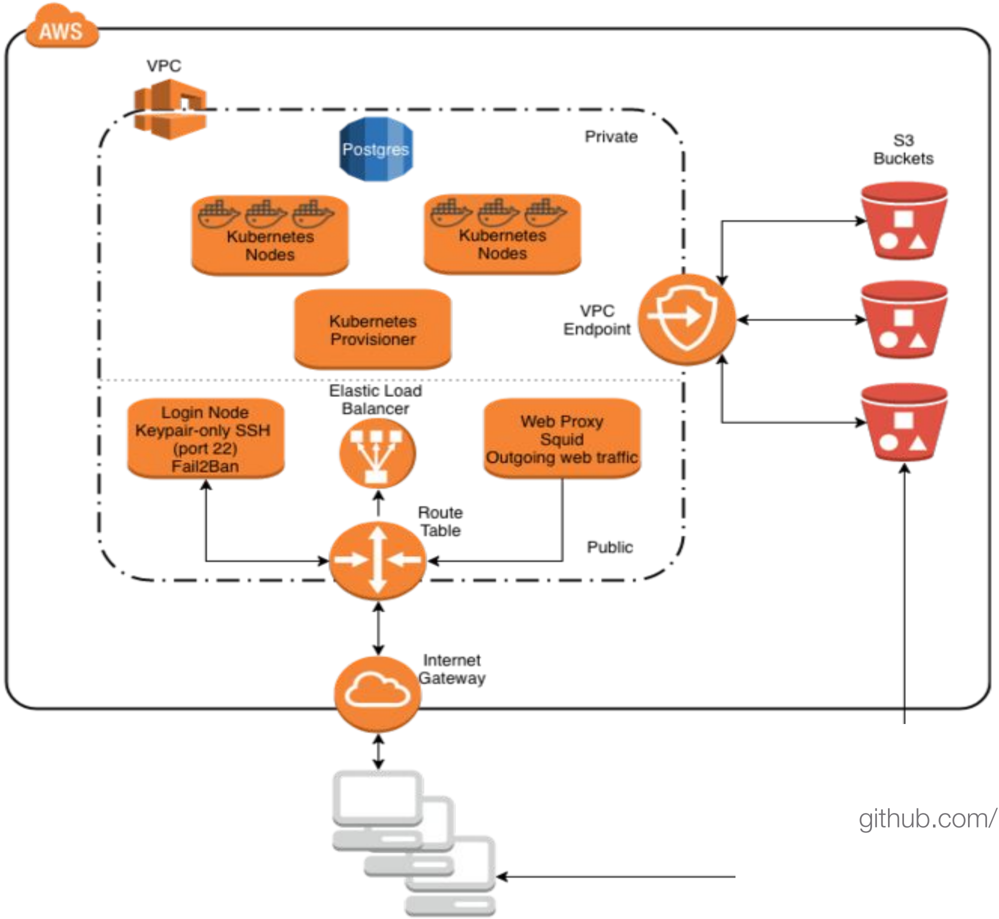
- We discuss high level architecture about POC setup on AWS using Gen3 `compose-services`.
- Centre who interested to adopt Gen3 will do trial POC run before committing to a more production oriented Cloud Automation (Kubernetes cluster) setup.
- Additionally, `compose-services` stack also gives a quick dive into Gen3 foundation services and, it is a good perk for your centre data dictionary development.

Idea:

- We choose AWS EC2 instance `m5.2xlarge` with Hibernation support.
- Gen3 `compose-services` stack simply run on this EC2 instance.
- We hibernate this instance when not in use (over weekend, doing other priority tasks, etc).
- This EC2 instance is front-ed by ALB -- Application Load Balancer. Hence, it shows graceful `504 Gateway Time-out` when EC2 is in hibernation.

Intermediate: Compose-Services on AWS

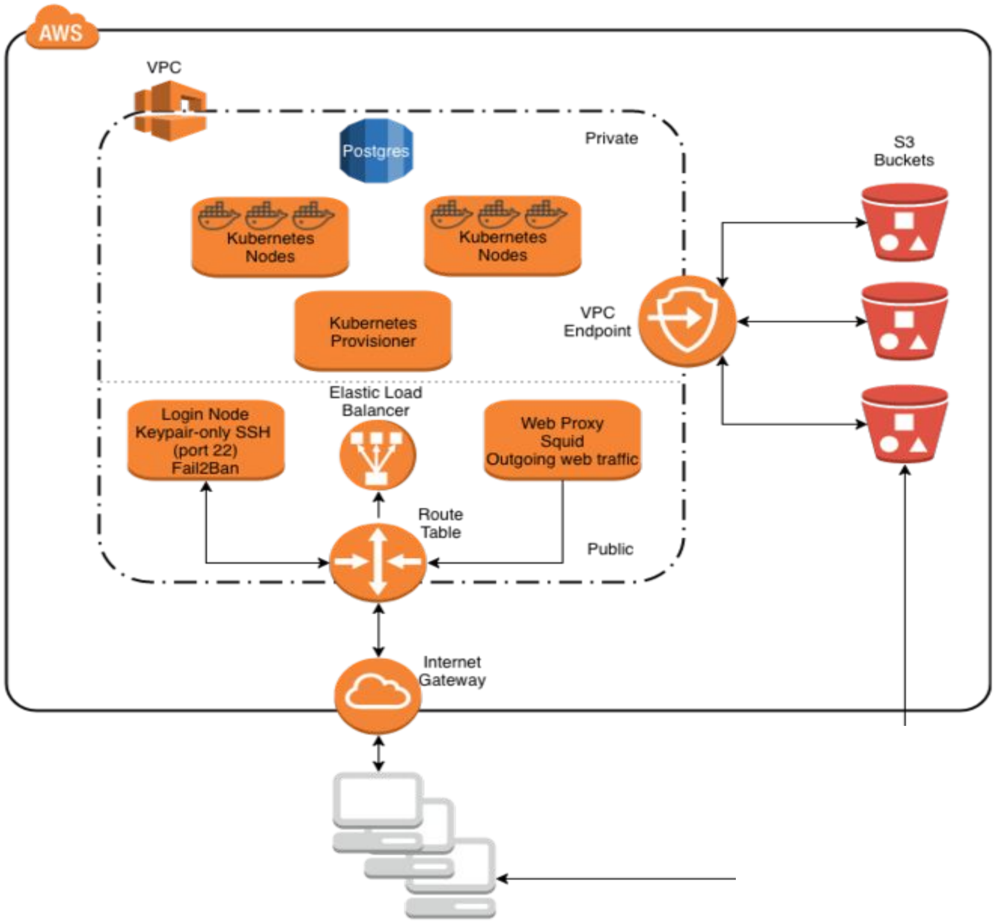
github.com/umccr/gen3-doc/blob/main/poc/AWS.md

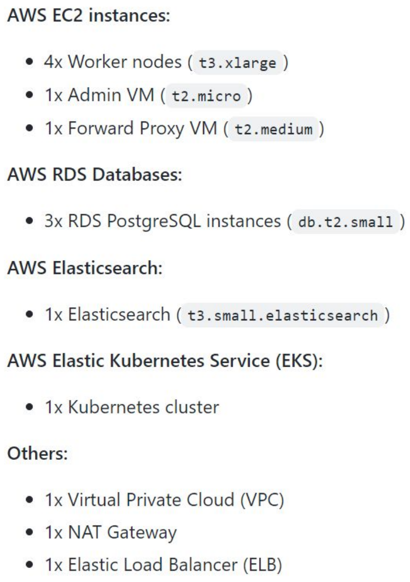


Scalable: Cloud

github.com/uc-cdis/cloud-automation and github.com/umccr/gen3-doc

Scalable: Cloud





AWS EC2 instances:

- 4x Worker nodes (`t3.xlarge`)
- 1x Admin VM (`t2.micro`)
- 1x Forward Proxy VM (`t2.medium`)

AWS RDS Databases:

- 3x RDS PostgreSQL instances (`db.t2.small`)

AWS Elasticsearch:

- 1x Elasticsearch (`t3.small.elasticsearch`)

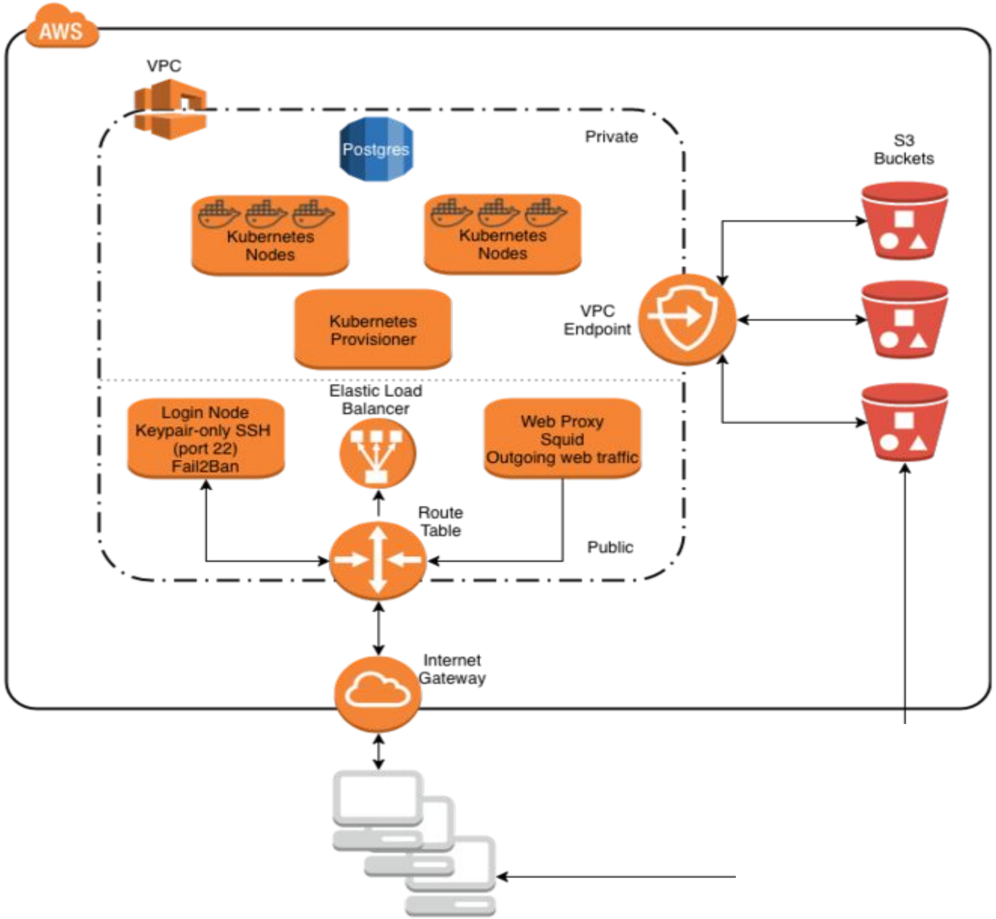
AWS Elastic Kubernetes Service (EKS):

- 1x Kubernetes cluster

Others:

- 1x Virtual Private Cloud (VPC)
- 1x NAT Gateway
- 1x Elastic Load Balancer (ELB)

github.com/umccr/gen3-doc/blob/main/cloud/AWS.md



AWS Services

Terraform

Kubernetes, Docker, Linux

ElasticSearch

PostgreSQL

GraphQL, Graph and DAG, ETL process

ReactJS SPA

Identity Provider (IdP), Federated AuthN/Z,

Single-SignOn (SSO) setup

…

Good troubleshooting skills

Scalable: Cloud

github.com/umccr/gen3-doc/blob/main/cloud/AWS.md

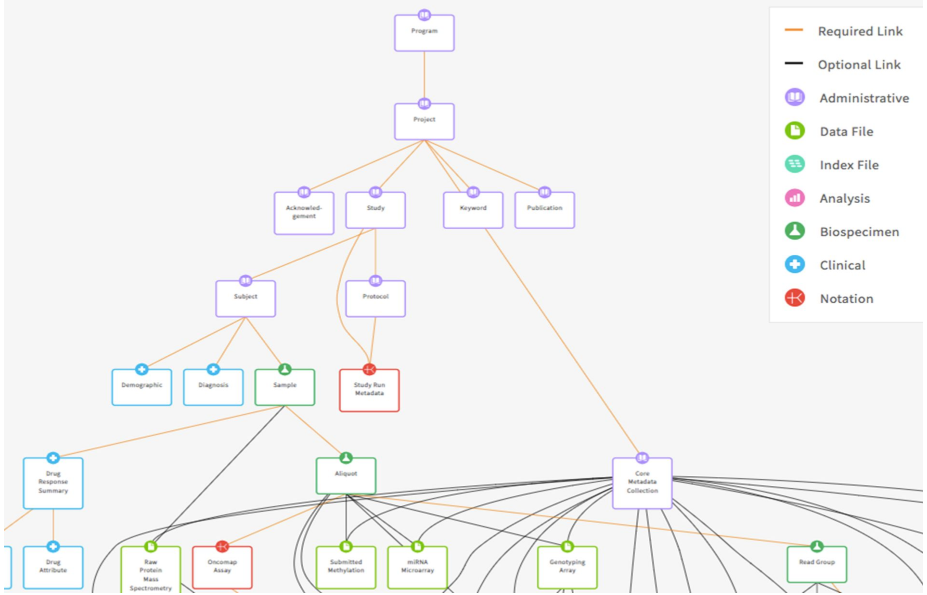
Quickstart: Sample Data Models

Define a data model

Generate a commons with a Gen3 API

Load data into the commons

Start exploring



Limitations and Difficulties

- Complicated infrastructure
- Data models are complex; one per instance
- Lack of granular control over permissions and data access

A national approach to genomics information management (NAGIM)

- The vision for human genomics data sharing in Australia requires considerable coordination and collaboration.
- The NAGIM Blueprint sets out a series of principles to guide decision-making on the responsible collection, storage, use and management of genomic data.
- Australian Genomics is developing recommendations for implementing NAGIM.
- In 2021 Australian Genomics led an implementation prototyping phase in response to NAGIM.
- A panel of external assessors are evaluating prototype submissions presently.

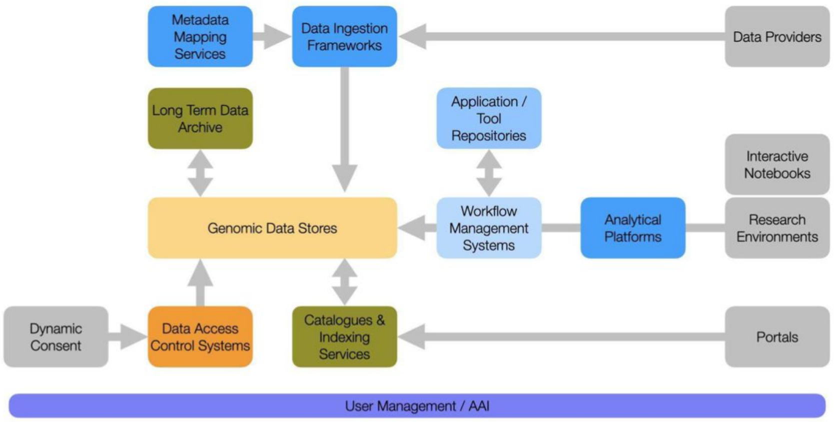
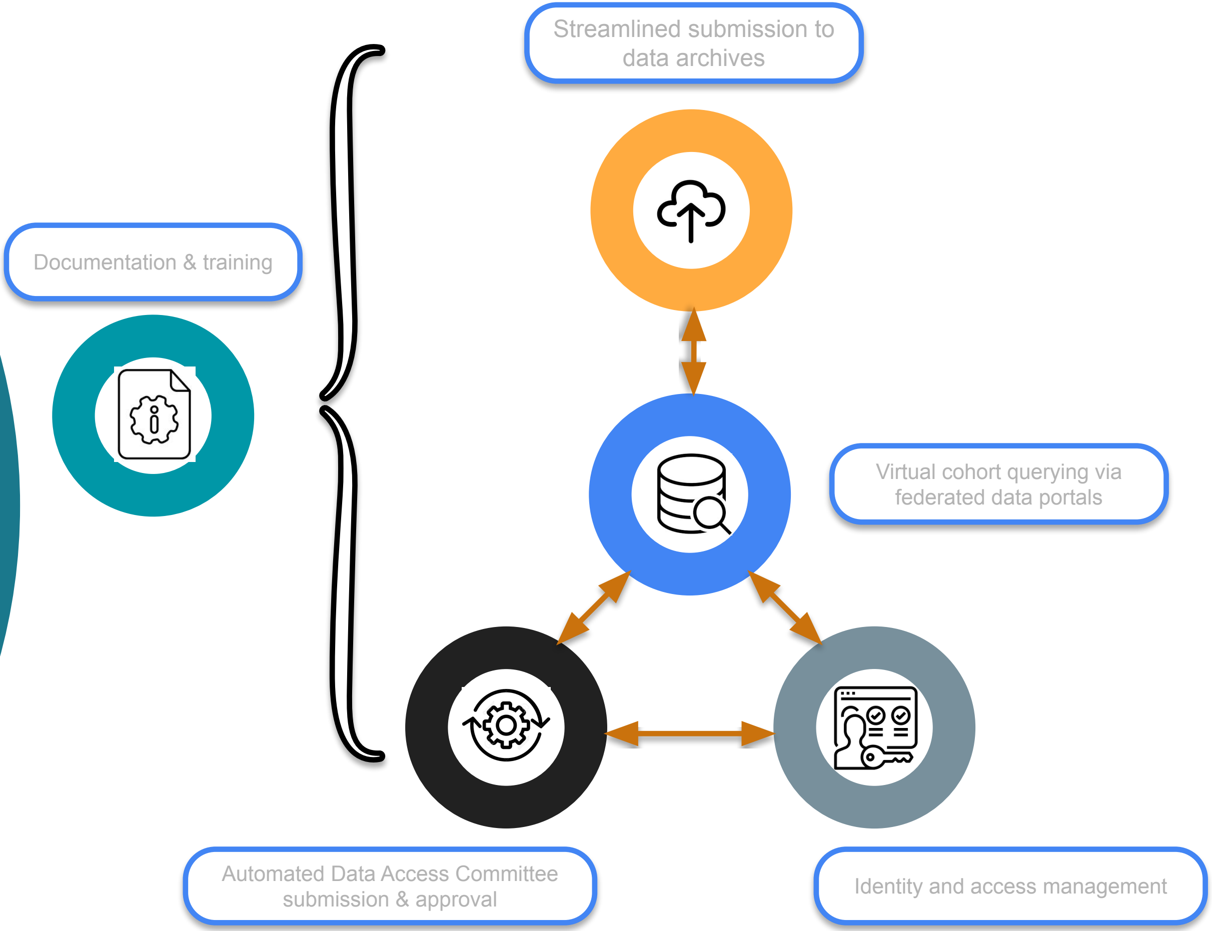

Metadata Mapping Services
Data Ingestion Frameworks
Data Providers
Long Term Data Archive
Application / Tool Repositories
Interactive Notebooks
Genomic Data Stores
Workflow Management Systems
Analytical Platforms
Research Environments
Dynamic Consent
Data Access Control Systems
Catalogues & Indexing Services
Portals
User Management / AAI

Figure 1. *Components of a national genomics infrastructure*

A National Approach to Genomic Information Management, Australian Genomics Implementation Recommendations Progress Report, November 2021

HUMAN GENOMES PLATFORM PROJECT



Streamlined submission to data archives

Documentation & training

Virtual cohort querying via federated data portals

Automated Data Access Committee submission & approval

Identity and access management

Project Partners

- Australian Access Federation
- Australian BioCommons
- ARDC Australian Research Data Commons
- Bioplatforms Australia
- Garvan Institute of Medical Research
- NCI Australia
- QIMR Berghofer Medical Research Institute
- The University of Melbourne
- zero Childhood Cancer

Jess Holliday and Marion Shadbolt

Supporting Australian Cardiovascular Disease Research

- We have begun working with partners from the Australian Cardiovascular Alliance to establish systems to support identification of biomarkers of increased risk of heart attack.
- We're currently mid way through an 8 month project to establish a new Gen3 instance and populate with 3 coronary artery disease cohorts.
- Data harmonisation across the cohorts is underway.
- We've populated the instance with synthetic data to allow functionality testing.

Baker
HEART & DIABETES INSTITUTE

THE UNIVERSITY OF SYDNEY

ACvA
Australian Cardiovascular Alliance

BIOPLATFORMS AUSTRALIA

Australian BioCommons

Acknowledgements

Australian BioCommons &

Melbourne Bioinformatics

- Jeff Christiansen
- Lisa Phippard
- Jess Holliday
- Marion Shadbolt
- Steven Manos
- Uwe Winter
- Andrew Lonie
- Nuwan Goonasekera

UMCCR

- Victor San-Kho Lin
- Florian Reisinger
- Andrew Patterson
- (Grant Lee)
- (Lavinia Gordon)

Australian Genomics

- Tiffany Boughtwood
- Marie-Jo Brion
- Sarah Casauria

Children's Cancer Institute, Sydney
Zero Childhood Cancer program

- Marie Wong-Erasmus
- Kamile Taouk
- Mark Cowley
- Vanessa Tyrrell

Children's Hospital of Philadelphia

- Allison Heath
- Adam Resnick
- Miguel Brown
  Yuankun Zhu
- Bailey Farrow

Questions?

Australian BioCommons

NEXT …

Conservation genomics in the age of extinction

Dr Carolyn Hogg, University of Sydney

8 March 2022

biocommons.org.au/events

Australian BioCommons

Tell us what you thought …

Feedback survey

Thanks for joining us!

The Australian BioCommons is enabled by NCRIS via Bioplatforms Australia funding

BIOPLATFORMS AUSTRALIA

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative