Collaborating in the Cloud - AMP PD / Terra

CIRM CNS Workshop - February 24-25, 2022

Matt Bookman (Verily Life Sciences)
David Craig (USC)
Barry Landin (Technome)

Acknowledgements

Acknowledgements

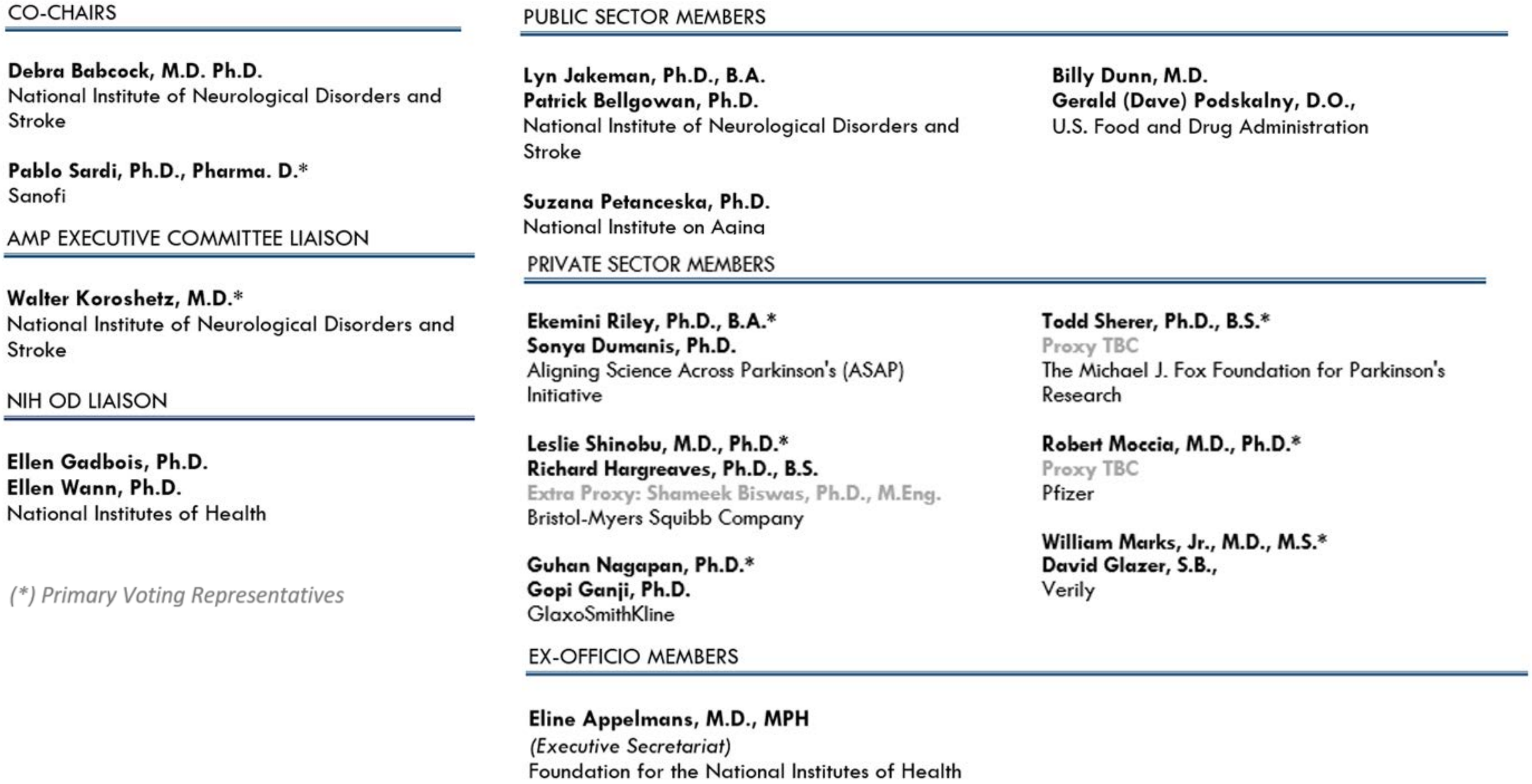
AMP PD is built on Parkinson's disease (PD) related study data that has been collected, harmonized, and made publicly available with the goals of identifying new biomarkers and developing new treatments for PD. Overcoming challenges associated with resources, time, cost, and data availability is made possible through the collective efforts of:

- MJFF and NINDS BioFIND Study
- NIA International Lewy Body Dementia Genetics Consortium Genome Sequencing in Lewy body dementia case-control cohort (LBD)
- MJFF LRRK2 Cohort Consortium (LCC)
- Brigham and Women's Hospital/MGH Harvard Biomarkers Study
- NINDS Parkinson's Disease Biomarkers Program
- MJFF Parkinson's Progression Markers Initiative
- NINDS Study of Isradipine as a Disease Modifying Agent in Subjects With Early Parkinson Disease, Phase 3 (STEADY-PD3).
- The MJFF and NINDS Study of Urate Elevation in Parkinson's Disease, Phase 3 (SURE-PD3)
- Global Parkinson's Genetics Program (GP2)

Special thanks to the study participants who have so graciously contributed their data.

AMP PD Steering Committee Roster

Reminder Proxy Member Nominations Welcome



CO-CHAIRS

Debra Babcock, M.D. Ph.D.
National Institute of Neurological Disorders and Stroke

Pablo Sardi, Ph.D., Pharma. D.*
Sanofi

AMP EXECUTIVE COMMITTEE LIAISON

Walter Koroshetz, M.D.*
National Institute of Neurological Disorders and Stroke

NIH OD LIAISON

Ellen Gadbois, Ph.D.
Ellen Wann, Ph.D.
National Institutes of Health

(*) Primary Voting Representatives

PUBLIC SECTOR MEMBERS

Lyn Jakeman, Ph.D., B.A.
Patrick Bellgowan, Ph.D.
National Institute of Neurological Disorders and Stroke

Suzana Petanceska, Ph.D.
National Institute on Aging

Billy Dunn, M.D.
Gerald (Dave) Podskalny, D.O.,
U.S. Food and Drug Administration

PRIVATE SECTOR MEMBERS

Ekemini Riley, Ph.D., B.A.*
Sonya Dumanis, Ph.D.
Aligning Science Across Parkinson's (ASAP) Initiative

Leslie Shinobu, M.D., Ph.D.*
Richard Hargreaves, Ph.D., B.S.
Extra Proxy: Shameek Biswas, Ph.D., M.Eng.
Bristol-Myers Squibb Company

Guhan Nagapan, Ph.D.*
Gopi Ganji, Ph.D.
GlaxoSmithKline

Todd Sherer, Ph.D., B.S.*
Proxy TBC
The Michael J. Fox Foundation for Parkinson's Research

Robert Moccia, M.D., Ph.D.*
Proxy TBC
Pfizer

William Marks, Jr., M.D., M.S.*
David Glazer, S.B.,
Verily

EX-OFFICIO MEMBERS

Eline Appelmans, M.D., MPH
(Executive Secretariat)
Foundation for the National Institutes of Health

It Takes a Village

Steering Committee:
- Deb Babcock (NIH/NINDS) – Co-Chair
- Pablo Sardi (Sanofi) – Co-Chair
- Billy Dunn (FDA)
- David Glazer (Verily)
- Ekemini Riley (ASAP)
- Ellen Gadbois (NIH/OD)
- Ellen Wann (NIH/OD)
- Gerald (Dave) Podskalny (FDA)
- Gopi Ganji (GSK)
- Guhan Nagappan (GSK)
- Leslie Shinobu (BMS)
- Lyn Jakemen, (NINDS)
- Pat Bellgowan (NINDS)
- Robert Moccia (Pfizer)
- Sonya Dumanis (ASAP)
- Shameek Biswas (BMS)
- Suzana Petanceska (NIA)
- Todd Sherer (MJFF)
- Walter Koroshetz (NIH/NINDS)
- William Marks (Verily)

Ad-hoc/Guest
- Barry Landin (FNIH Contractor)
- Dave Vismer (FNIH Contractor)
- Irit Rapley (BMS)
- Matt Bookman (Verily)
- Richard Hargreaves (BMS)

Metabolomics Working Group
- Deb Babcock (NINDS) – Co-Chair
- Lilu Guo (Sanofi) – Co-Chair
- Barry Landin (FNIH Contractor)
- Chrissa Dwyer (GSK)
- Eline Appelmans (FNIH)
- Kevin Jensen (Celgene/BMS)
- Marcus Bantscheff (GSK)
- Mark Frasier (MJFF)
- Suzana Petanceska (NIA)

Ad-hoc
- Rima Kaddurah-Daouk (Duke)
- Michael Schwarzchild (Harvard)

Data Usage and Policies Working Group
- Alyssa Reimer (MJFF) – Co-Chair
- Deb Babcock (NINDS) – Co-Chair
- Claire Wegel (IU/GP2)
- Clemens Scherzer (BWH)
- Ekemini Riley (ASAP)
- Eline Appelmans (FNIH)
- Matt Bookman (Verily)
- Yuliya Kuras (Harvard)
- Sonja Scholz (NIA – LBD)
- Dave Alonso (MJFF – LCC)

Outreach and Portal Working Group
- Patrick Bellgowan(NINDS) – Co-Chair
- Leslie Kirsch (MJFF) – Co-Chair
- Suzanne Beschamps (FNIH)
- Carl Wonders (NINDS)
- Chris Swanson-Fischer (NINDS)
- David Alonso (MJFF)
- Deb Babcock (NINDS)
- Eline Appelmans (FNIH)
- Kathleen Parkes (Verily)
- Kristine Treece (NINDS Contractor)
- Sumit Dey (QMUL/GP2)

Ad-hoc
- Matt Bookman (Verily)
- Barbara Marebwa (MJFF)
- Amy Adams (NINDS)
- Barry Landin (FNIH Contractor)
- Katherine Thompson (FNIH)
- Nadia Douaji (NINDS)
- Ryan Stewart-Frederick (FNIH Contractor)

RNA Seq Brain Working Group
- Deb Babcock (NINDS) – Co-Chair
- Kevin Jensen (Celgene/BMS)– Co-Chair
- Barry Landin (FNIH Contractor)
- Bradford Casey (MJFF)
- Charissa Dwyer (GSK)
- Eline Appelmans (FNIH)
- David Craig (USC)
- Dinesh Kumar (Sanofi)
- Suzana Petanceska (NIA)

Data Analyses Working Group
- Bradford Casey (MJFF) – Co-Chair
- Hirotaka Iwaki (NIA) – Co-Chair
- Alyssa Reimer (MJFF)
- Andrew Sharp (MSSM)
- Andy Singleton (NIA/LNG)
- Christoph Bussler (Verily)
- Clemens Scherzer (BWH)
- David Alonso (MJFF)
- David Craig (USC)
- David Knowles (NYGenome)
- David Vismer (FNIH Contractor)
- Deb Babcock (NINDS)
- Eline Appelmans (FNIH)
- Elizabeth Hutchins (TGEN)
- Erin Teeple (Sanofi)
- Hampton Leonard (NIA/NINDS)
- Ivo Violich (USC)
- Jie Yuan (BWH)
- Kendall Jenssen (TGEN)
- Mary Makarious (NIA/NINDS)
- Matt Bookman (Verily)
- Meaghan Cogswell (Sanofi)
- Mike Nalls (NIA/NINDS)
- Ninad Amondikar (MJFF)
- Ramsey Magana (GSK)
- Ramiya Sivakumar (USC)
- Ricardo Vialle (MSSM)
- Rob Woodruff (GSK)
- Ruifeng Hu (BWH)
- Samantha Hutten (MJFF)
- Shameek Biswas (BMS)
- Srini Shankara (Sanofi)
- Towfique Raj (MSSM)
- Xianjun Dong (BWH)

Clinical Data Harmonization Working Group
- Alyssa Reimer (MJFF) – Co-Chair
- Deb Babcock (NINDS) – Co-Chair
- Barry Landin (FNIH Contractor)
- Clemens Scherzer (BWH)
- David Pulford (GSK)
- David Vismer (FNIH Contractor)
- Eline Appelmans (FNIH)
- Erin Teeple (Sanofi)
- Mary Makarious (NIA/NINDS)
- Matt Bookman (Verily)
- Nicole Glazer (Celgene/BMS)
- Robert Moccia (Pfizer)
- Yuliya Kuras (Harvard)

WGS and TS Subgroup
- Andy Singleton (NIA/LNG) – WGS Co-Chair
- Shameek Biswas (Celgene/BMS) – WGS Co-Chair
- Kendall Jensen (TGen) – Co-chair
- Deb Babcock (NINDS) – Co-Chair (Temp)
- Ashutosh Pandey (GSK)
- Barry Landin (FNIH Contractor)
- Bradford Casey (MJFF)
- Christine Swanson-Fischer (NINDS)
- Christoph Bussler (Verily)
- Daniel Seaton (GSK)
- David Craig (USC)
- David Pulford (GSK)
- David Vismer (FNIH Contractor)
- Dena Hernandez (NIA)
- Eline Appelmans (FNIH)
- Elizabeth Hutchins (TGen)
- Guhan Nagappan (GSK)
- Hampton Leonard (NIA)
- Hirotaka Iwaki (NIA)
- Ivo Violich (USC)
- Jerrod Schwartz (Verily)
- Jie Yuan (BWH)
- Lyn Jakeman (NINDS)
- Mark Frasier (MJFF)
- Mary Makarious (NIA/NINDS)
- Matt Bookman (Verily)
- Matt Edwards (Verily)
- Meaghan Cogswell (Sanofi)
- Mike Nalls (NIA)
- Ramsey Magana (GSK)
- Robert Moccia (Pfizer)
- Robert Woodruff (GSK)
- Ruifeng Hu (BWH)
- Shawn Levy (Hudson Alpha)
- Sonja Scholz (NIA)
- Srini Shankara (Sanofi)
- Willy Nojopranoto (Verily)
- Xianjun Dong (BWH)

Access Compliance Team
Admin Support Team
Operations Team
Publication Subcommittee
and all the alumni and external experts that have helped us along the way

Proteomics Working Group
- Deb Babcock (NINDS) – Co-Chair
- Samantha Hutten (MJFF) – Co-Chair
- Amilcar Flores-Morales (Sanofi)
- Andy Christoforou (BMS)
- Bailin Zhang (Sanofi)
- Barry Landin (FNIH Contractor)
- Bradford Casey (MJFF)
- Christine Swanson-Fischer (NINDS)
- Eline Appelmans (FNIH)
- Howard Schulman (KOL)
- Lyn Jakeman (NINDS)
- Marcus Bantscheff (GSK)
- Pablo Sardi (Sanofi)

Umbrella Data Working Group
- Deb Babcock (NINDS) – Co-Chair
- Matt Bookman (Verily) – Co-Chair
- Alyssa Reimer (MJFF)
- Andy Singleton (NIA/LNG)
- Arthur Toga (USC)
- Bailin Zhang (Sanofi)
- Barry Landin (FNIH Contractor)
- Bradford Casey (MJFF)
- Codrin Lungu (NINDS)
- David Craig (USC)
- Eline Appelmans (FNIH)
- Erin Teeple (Sanofi)
- Juan Botia (KCL)
- Lungu Codrin (NINDS)
- Mike Nalls (NIA, LNG)
- Pablo Sardi (Sanofi)
- Ritu Kapur (Verily)
- Robert Moccia (Pfizer)
- Shameek Biswas (BMS)
- Willy Nojopranoto (Verily)

Special thanks to all cohort contributors & study participants

FNIH *AMP PD is made possible by the Public and Private Partners and members who have worked tirelessly to make this valuable resource available to the scientific community.*

AMP PD Organization and Objectives

AMP PD Terra

FDA

NIH National Institute of Neurological Disorders

NIH National Institute on Aging

Government

FNIH
Foundation for the National Institutes of Health

THE MICHAEL J. FOX FOUNDATION
FOR PARKINSON'S RESEARCH

Non-profit

Bristol Myers Squibb™

SANOFI

Pfizer

gsk

verily

Industry

Accelerating Medicines Partnership - Parkinson’s Disease (AMP PD)

What: Public/private partnership managed through the FNIH

When: Kickoff 2018; initially funded for 5 years

Who:

- Government: FDA, NINDS, NIA
- Industry: Bristol-Myers Squibb, GSK, Pfizer, Sanofi, Verily
- Nonprofit: Michael J. Fox Foundation (MJFF)

Goal: Create a resource for biomarker discovery for Parkinson's Disease

URL: https://www.amp-pd.org

AMP PD: Create a resource for biomarker discovery

Data:

- **Harmonize** clinical data from 4 (now 8) separate studies
- Generate new, **analysis-ready**, RNASeq, WGS, and proteomics data
- **Integrate existing WGS data** (leveraging functional equivalence)

Researchers:

- **Bring researchers to the data** (where it is)
- **Accelerate researcher onboarding** with getting started code and documentation
- Derive results, generate knowledge, and **return it to the program**
- **Enable collaboration** between researchers, informaticians, software engineers

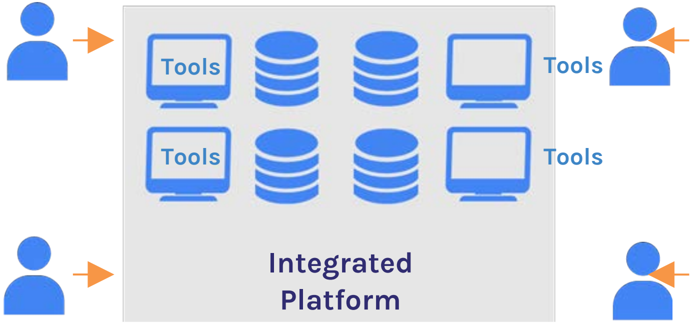
Data Biosphere approach to biomedical research

Traditional approach

Bring data to researchers and build tools

Data Biosphere approach

Bring research to data and share tools



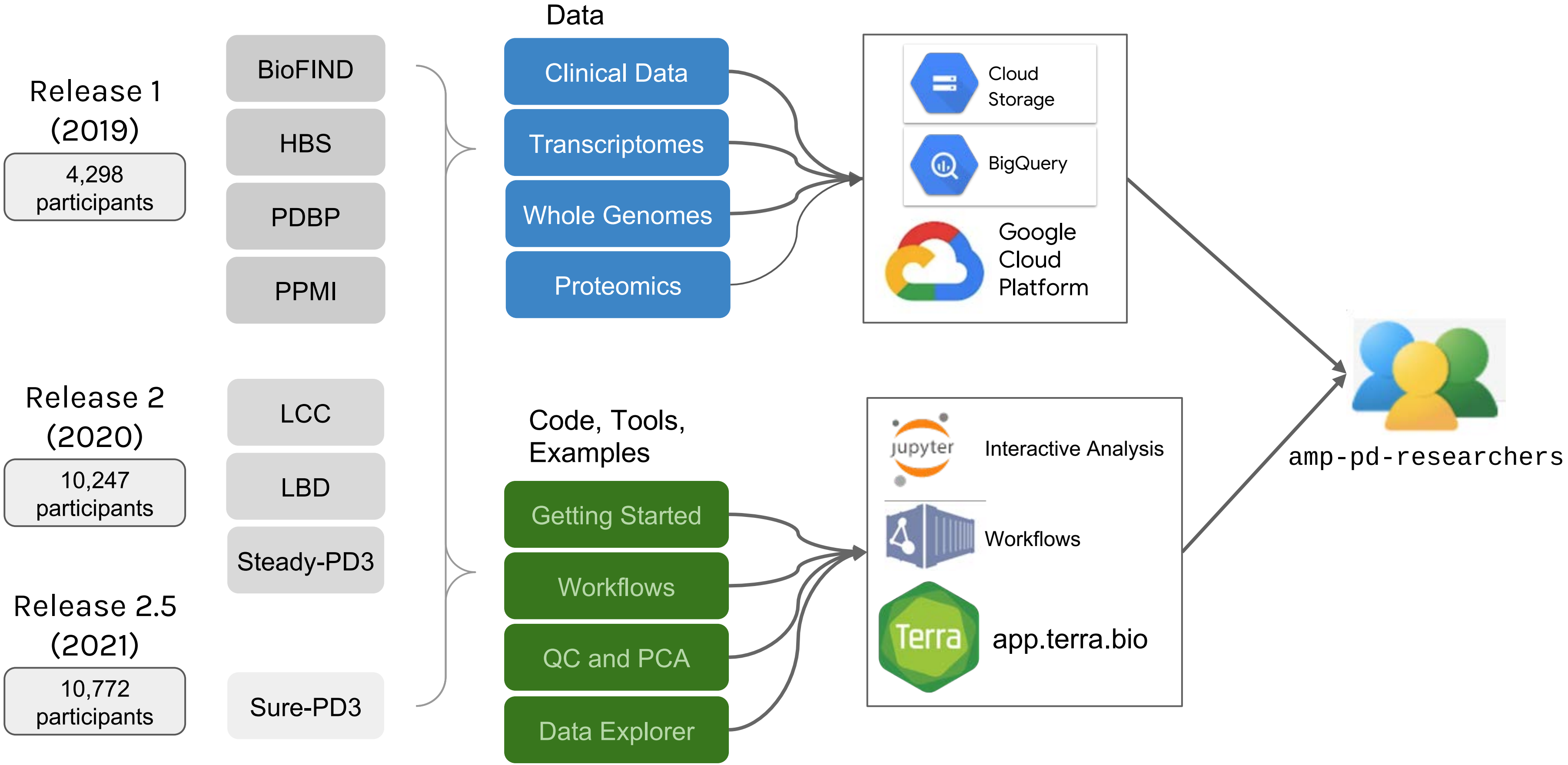
Challenges

Data sharing = data copying + poor security
Bespoke & unsupported tools
Huge infrastructure needed
Discourages shared research

Advantages

More accessible and secure data
Shared tool ecosystem
Decreased cost of storage & compute
Facilitates collaboration

AMP PD data & researcher access



Release 1
(2019)
4,298 participants

Release 2
(2020)
10,247 participants

Release 2.5
(2021)
10,772 participants

BioFIND
HBS
PDBP
PPMI
LCC
LBD
Steady-PD3
Sure-PD3

Data
- Clinical Data
- Transcriptomes
- Whole Genomes
- Proteomics

Code, Tools, Examples
- Getting Started
- Workflows
- QC and PCA
- Data Explorer

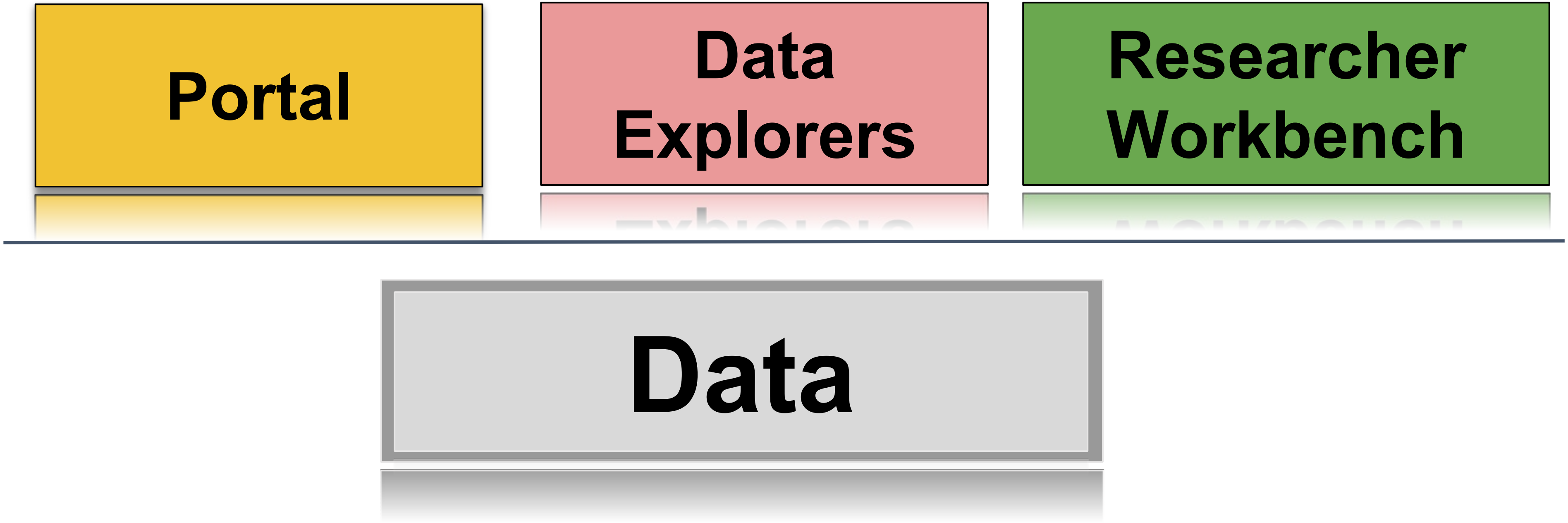
Cloud Storage
BigQuery
Google Cloud Platform

Interactive Analysis
Workflows
app.terra.bio

amp-pd-researchers

AMP PD Platform Architecture

AMP PD Architecture Overview



Portal

Data Explorers

Researcher Workbench

Data

AMP PD Architecture Overview

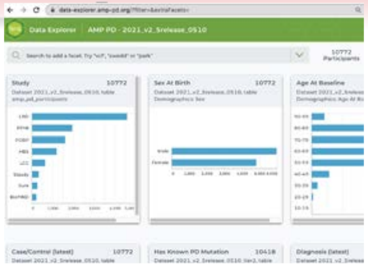
Portal

an Accelerating Medicines Partnership® (AMP®) program

Accelerating Medicines Partnership®
Parkinson's Disease

Collaborating toward biomarker discovery to advance the development of Parkinson's disease therapies

Data Explorers



Researcher Workbench

Welcome to Terra

Terra is a cloud-native platform for biomedical researchers to **access data**, **run analysis tools**, and **collaborate**.

Find how-to's, documentation, video tutorials, and discussion forums

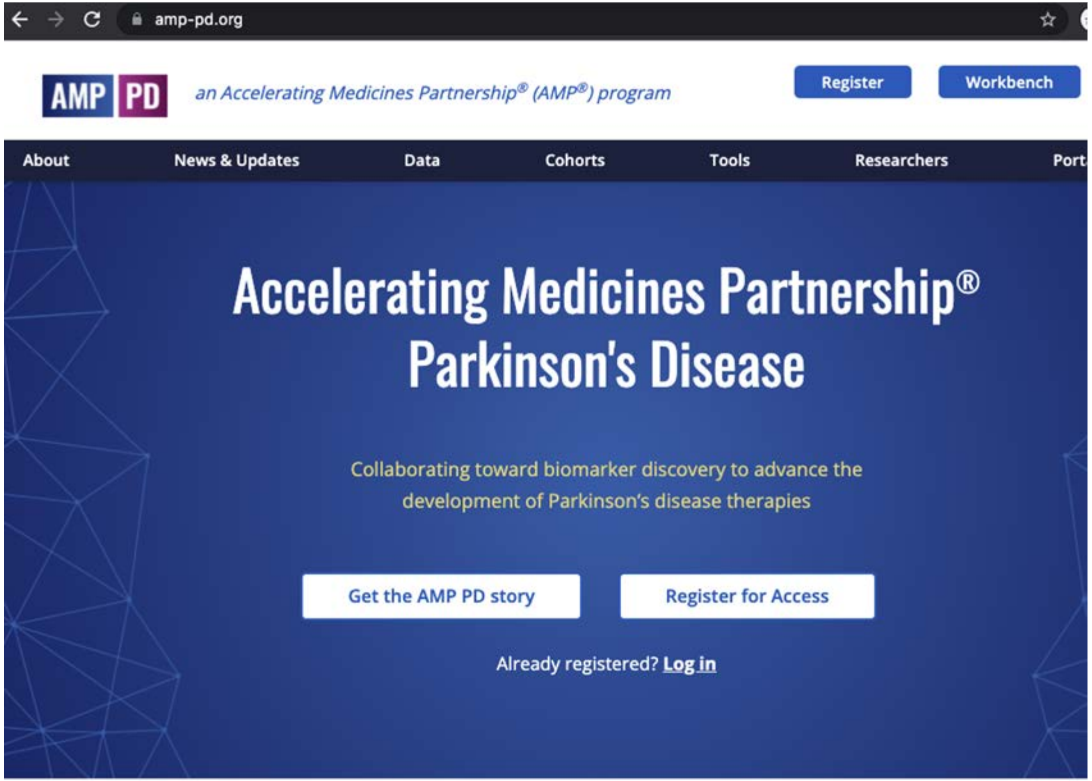
Learn more about the Terra platform and our co-branded sites

Data

Google Cloud Platform

AMP PD Key Components

AMP PD public web site



amp-pd.org

AMP PD an Accelerating Medicines Partnership® (AMP®) program

Accelerating Medicines Partnership® Parkinson's Disease

Collaborating toward biomarker discovery to advance the development of Parkinson's disease therapies

Already registered? Log in

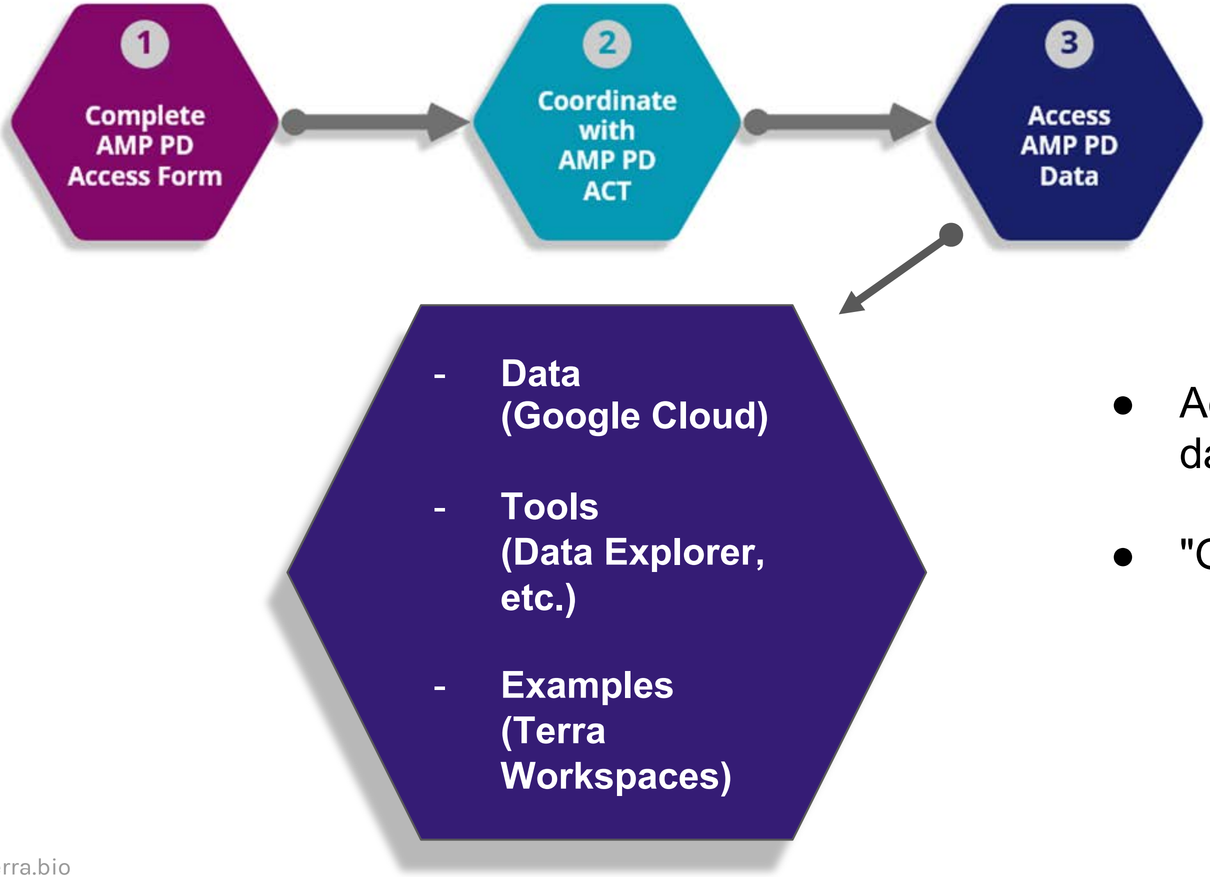
- Information about AMP PD organization and objectives
- Details about contributing studies
- Data summary dashboard
- Data governance policies

- Forms to register for access

- Links to the Data Explorer
- Links to Terra

Portal

AMP PD researcher access



1 Complete AMP PD Access Form

2 Coordinate with AMP PD ACT

3 Access AMP PD Data

- Data (Google Cloud)
- Tools (Data Explorer, etc.)
- Examples (Terra Workspaces)

- Access Control managed by data stewards (AMP PD)
- "Granted Access" =
  data + tools + code

Portal

AMP PD data organized by type

Cloud Storage

Clinical

- Demographics**.csv**
- MDS_UDPRS**.csv**
- MOCA**.csv**
- …

Transcriptomics

- **BAMs (.bam)**
- Quantification (.sf)
- Feature Counts

Whole Genome Sequencing

- **CRAMs**
- VCFs
- QC Metrics

BigQuery

Clinical

- Demographics
- MDS_UDPRS
- MOCA
- …

Transcriptomics

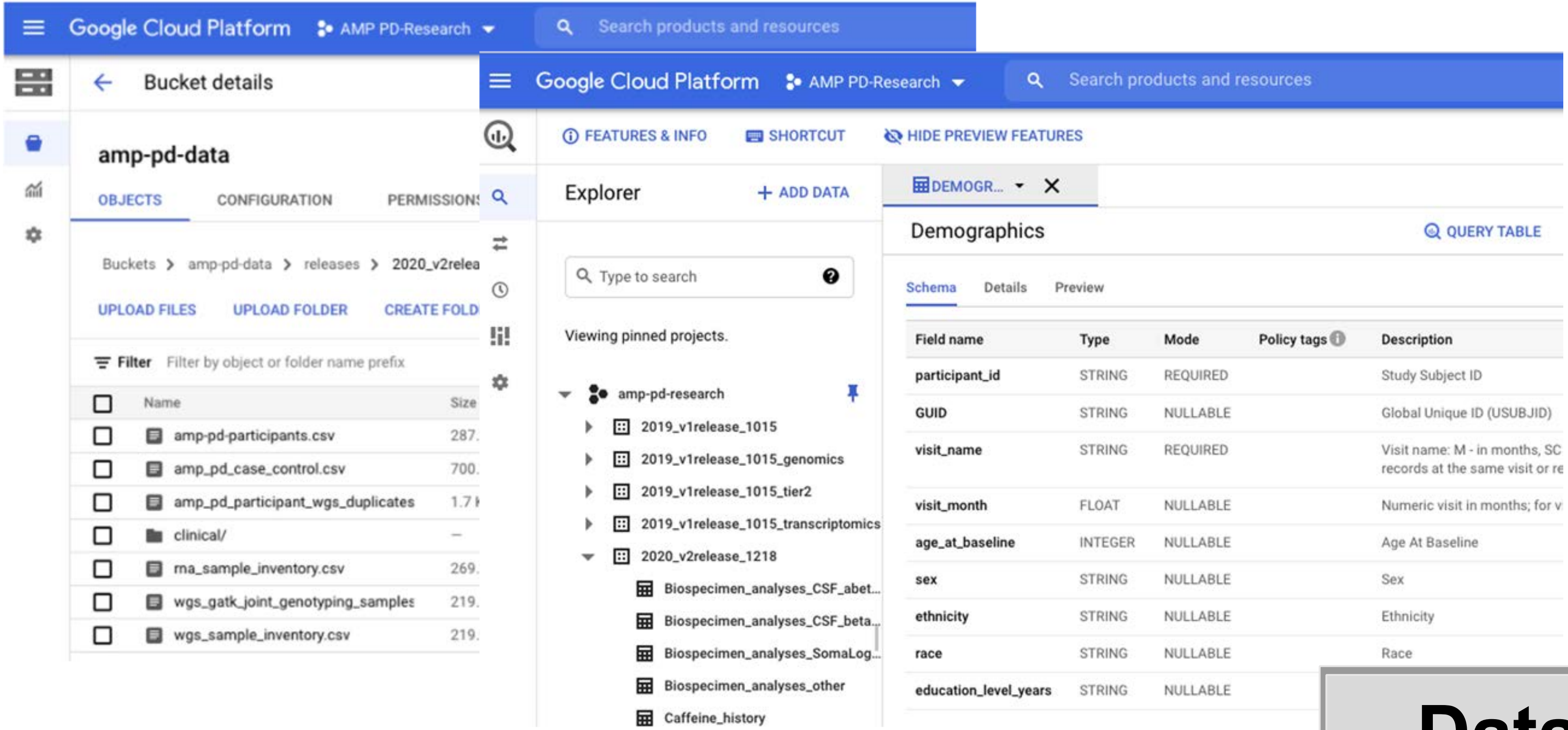
- Quantification (TPMs)
- Feature Counts

Whole Genome Sequencing

- Variants
- QC Metrics

Data

AMP PD data access with cloud native tools



amp-pd-data

Buckets > amp-pd-data > releases > 2020_v2relea

☐	Name	Size
☐	amp-pd-participants.csv	287.
☐	amp_pd_case_control.csv	700.
☐	amp_pd_participant_wgs_duplicates	1.7 k
☐	clinical/	–
☐	rna_sample_inventory.csv	269.
☐	wgs_gatk_joint_genotyping_samples	219.
☐	wgs_sample_inventory.csv	219.

Viewing pinned projects.

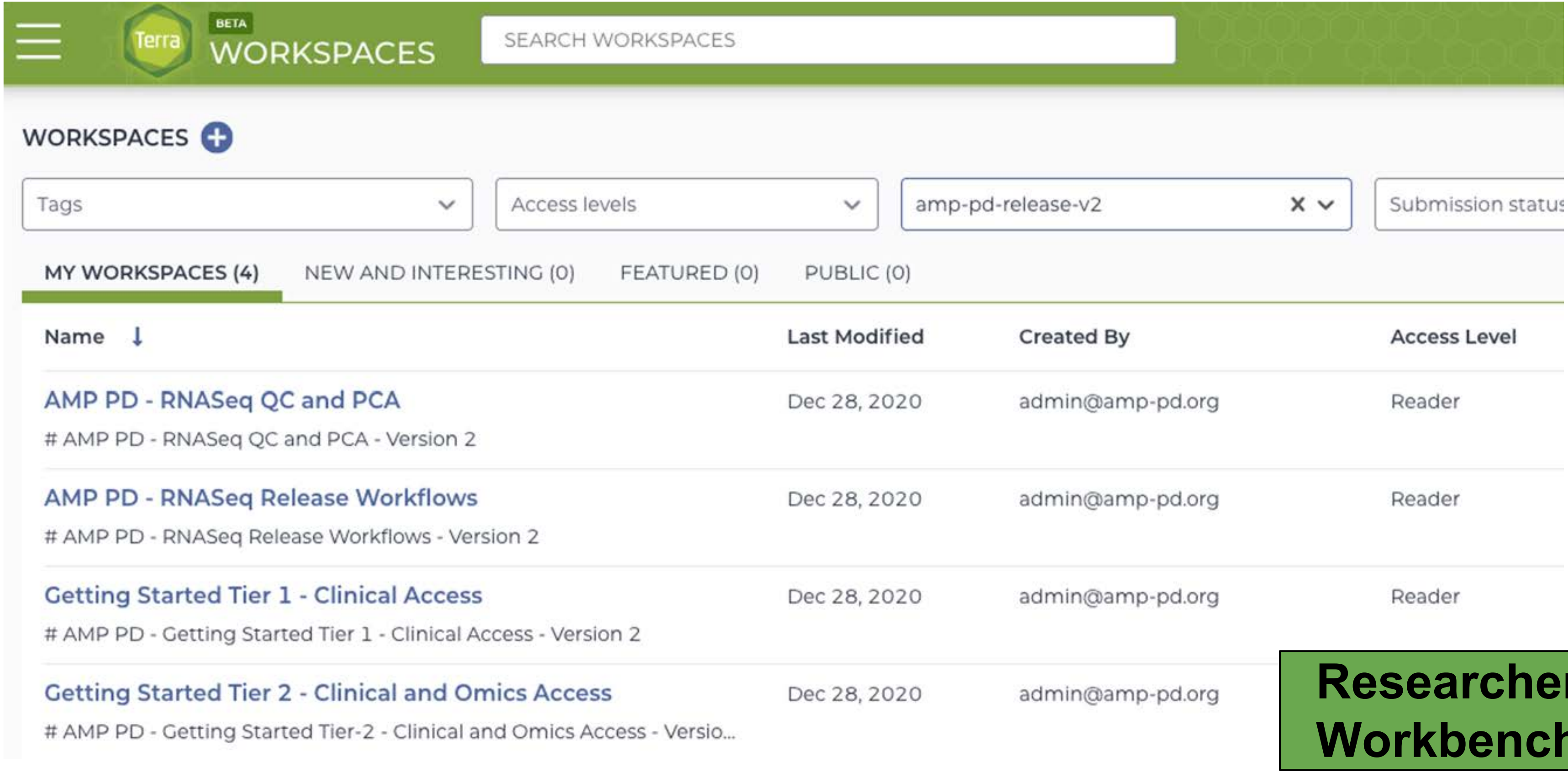
- amp-pd-research
  - 2019_v1release_1015
  - 2019_v1release_1015_genomics
  - 2019_v1release_1015_tier2
  - 2019_v1release_1015_transcriptomics
  - 2020_v2release_1218
    - Biospecimen_analyses_CSF_abet...
    - Biospecimen_analyses_CSF_beta...
    - Biospecimen_analyses_SomaLog...
    - Biospecimen_analyses_other
    - Caffeine_history

Demographics

Field name	Type	Mode	Policy tags	Description
participant_id	STRING	REQUIRED		Study Subject ID
GUID	STRING	NULLABLE		Global Unique ID (USUBJID)
visit_name	STRING	REQUIRED		Visit name: M - in months, SC records at the same visit or re
visit_month	FLOAT	NULLABLE		Numeric visit in months; for v
age_at_baseline	INTEGER	NULLABLE		Age At Baseline
sex	STRING	NULLABLE		Sex
ethnicity	STRING	NULLABLE		Ethnicity
race	STRING	NULLABLE		Race
education_level_years	STRING	NULLABLE		

Data

AMP PD Getting Started Workspaces



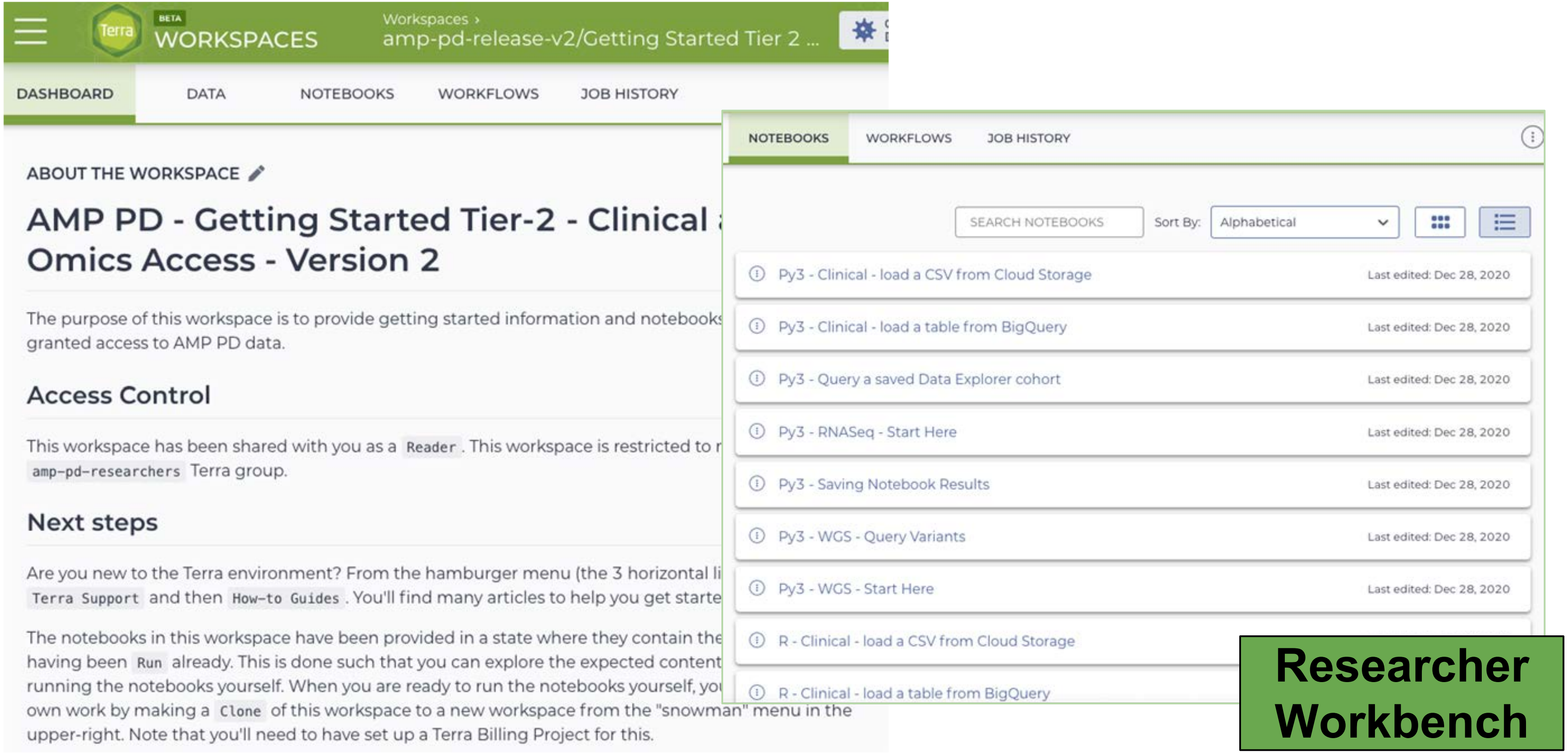
Filter: amp-pd-release-v2

Name	Last Modified	Created By	Access Level
AMP PD - RNASeq QC and PCA # AMP PD - RNASeq QC and PCA - Version 2	Dec 28, 2020	admin@amp-pd.org	Reader
AMP PD - RNASeq Release Workflows # AMP PD - RNASeq Release Workflows - Version 2	Dec 28, 2020	admin@amp-pd.org	Reader
Getting Started Tier 1 - Clinical Access # AMP PD - Getting Started Tier 1 - Clinical Access - Version 2	Dec 28, 2020	admin@amp-pd.org	Reader
Getting Started Tier 2 - Clinical and Omics Access # AMP PD - Getting Started Tier-2 - Clinical and Omics Access - Versio...	Dec 28, 2020	admin@amp-pd.org	

Researcher Workbench

terra.bio

AMP PD Getting Started Workspaces



ABOUT THE WORKSPACE

AMP PD - Getting Started Tier-2 - Clinical
Omics Access - Version 2

The purpose of this workspace is to provide getting started information and notebooks
granted access to AMP PD data.

Access Control

This workspace has been shared with you as a `Reader`. This workspace is restricted to
`amp-pd-researchers` Terra group.

Next steps

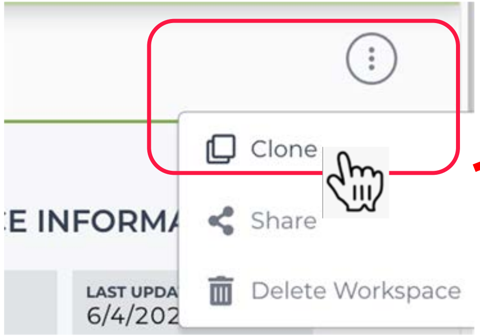
Are you new to the Terra environment? From the hamburger menu (the 3 horizontal li
`Terra Support` and then `How-to Guides`. You'll find many articles to help you get starte

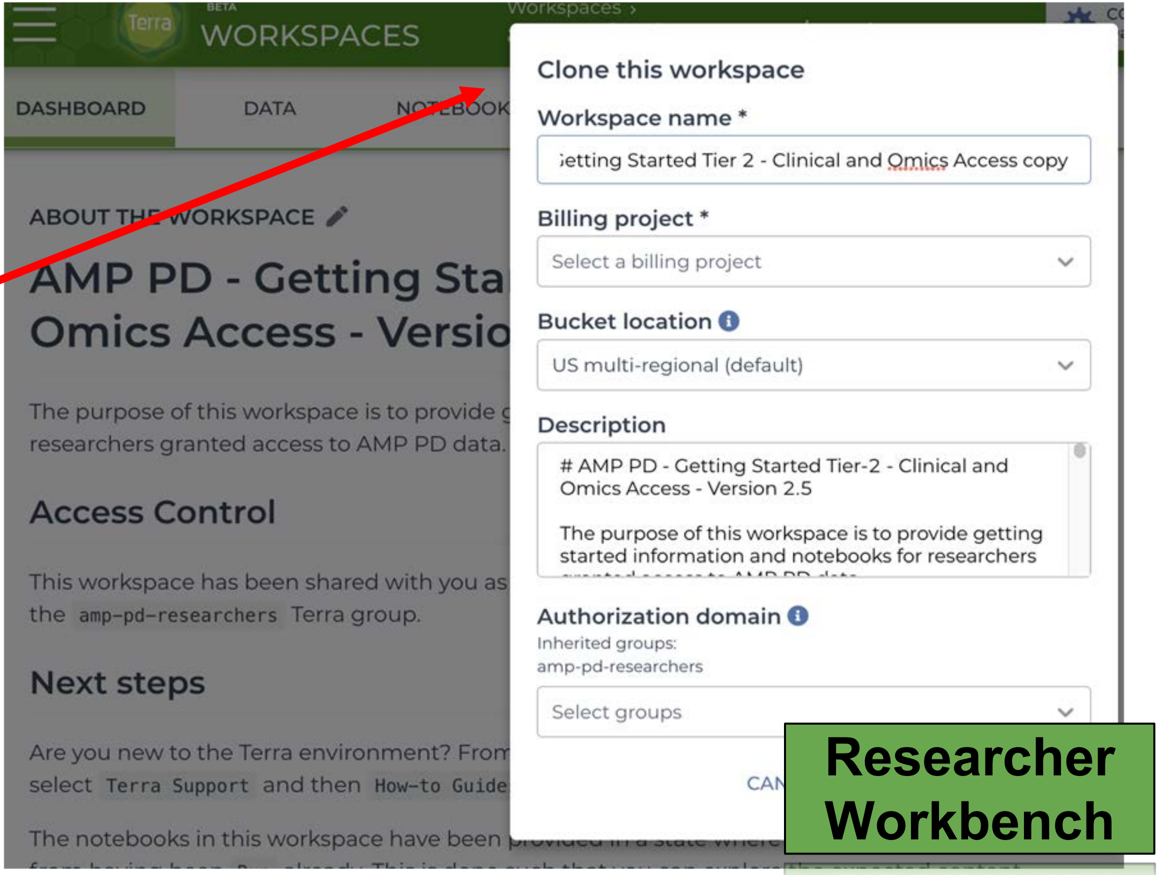
The notebooks in this workspace have been provided in a state where they contain the
having been `Run` already. This is done such that you can explore the expected content
running the notebooks yourself. When you are ready to run the notebooks yourself, yo
own work by making a `Clone` of this workspace to a new workspace from the "snowman" menu in the
upper-right. Note that you'll need to have set up a Terra Billing Project for this.

Notebook	Last edited
Py3 - Clinical - load a CSV from Cloud Storage	Last edited: Dec 28, 2020
Py3 - Clinical - load a table from BigQuery	Last edited: Dec 28, 2020
Py3 - Query a saved Data Explorer cohort	Last edited: Dec 28, 2020
Py3 - RNASeq - Start Here	Last edited: Dec 28, 2020
Py3 - Saving Notebook Results	Last edited: Dec 28, 2020
Py3 - WGS - Query Variants	Last edited: Dec 28, 2020
Py3 - WGS - Start Here	Last edited: Dec 28, 2020
R - Clinical - load a CSV from Cloud Storage	
R - Clinical - load a table from BigQuery	

Researcher Workbench

Workspaces: Read and then Clone





ABOUT THE WORKSPACE

AMP PD - Getting Sta
Omics Access - Versio

The purpose of this workspace is to provide g
researchers granted access to AMP PD data.

Access Control

This workspace has been shared with you as
the `amp-pd-researchers` Terra group.

Next steps

Are you new to the Terra environment? From
select `Terra Support` and then `How-to Guide`

The notebooks in this workspace have been provided in a state where

Clone this workspace

Workspace name *: etting Started Tier 2 - Clinical and Omics Access copy

Billing project *: Select a billing project

Bucket location: US multi-regional (default)

Description:

AMP PD - Getting Started Tier-2 - Clinical and Omics Access - Version 2.5

The purpose of this workspace is to provide getting started information and notebooks for researchers

Authorization domain

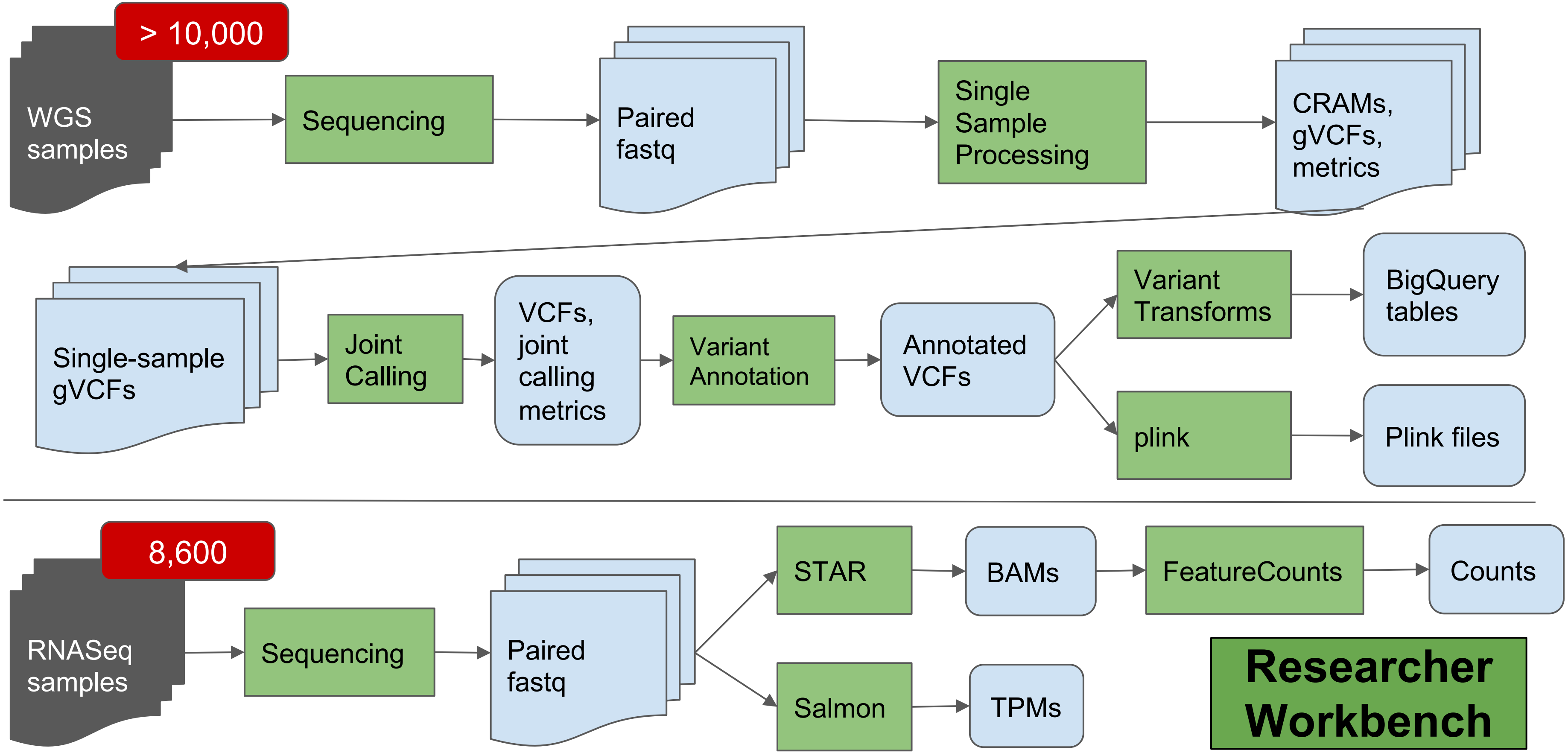
Inherited groups:
amp-pd-researchers

Select groups

CAN

Researcher Workbench

AMP PD omics workflows (Terra / Google Cloud)



> 10,000

WGS samples

Sequencing

Paired fastq

Single Sample Processing

CRAMs, gVCFs, metrics

Single-sample gVCFs

Joint Calling

VCFs, joint calling metrics

Variant Annotation

Annotated VCFs

Variant Transforms

BigQuery tables

plink

Plink files

8,600

RNASeq samples

Sequencing

Paired fastq

STAR

BAMs

FeatureCounts

Counts

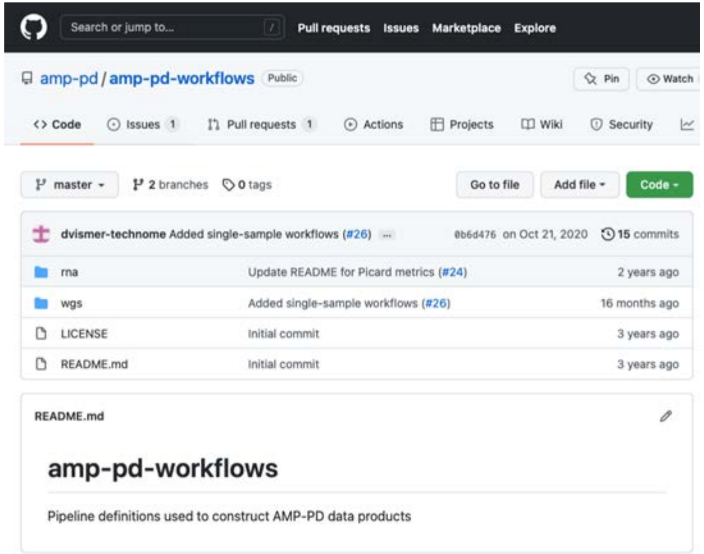
Salmon

TPMs

Researcher Workbench

AMP PD omics workflows

Available in Github



amp-pd / amp-pd-workflows Public

2 branches 0 tags

dvismer-technome Added single-sample workflows (#26) ... 0b6d476 on Oct 21, 2020 15 commits

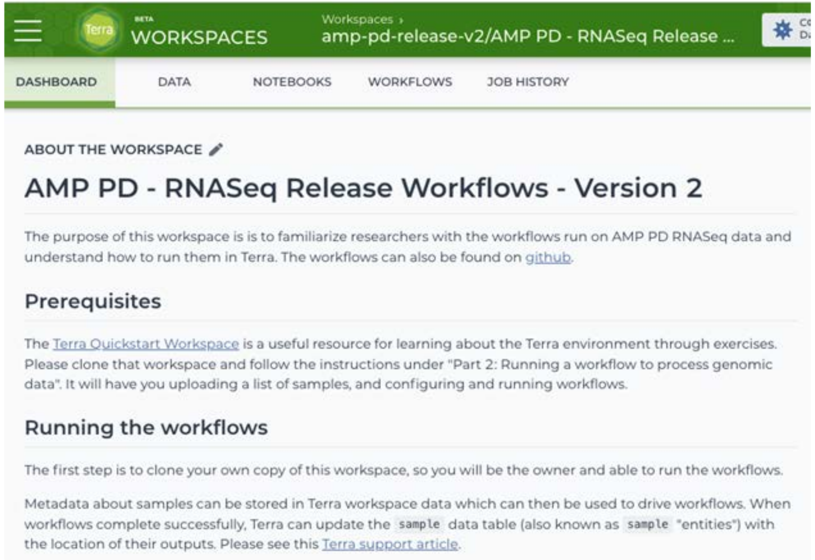
rna	Update README for Picard metrics (#24)	2 years ago
wgs	Added single-sample workflows (#26)	16 months ago
LICENSE	Initial commit	3 years ago
README.md	Initial commit	3 years ago

README.md

amp-pd-workflows

Pipeline definitions used to construct AMP-PD data products

Available in Terra



ABOUT THE WORKSPACE

AMP PD - RNASeq Release Workflows - Version 2

The purpose of this workspace is is to familiarize researchers with the workflows run on AMP PD RNASeq data and understand how to run them in Terra. The workflows can also be found on github.

Prerequisites

The Terra Quickstart Workspace is a useful resource for learning about the Terra environment through exercises. Please clone that workspace and follow the instructions under "Part 2: Running a workflow to process genomic data". It will have you uploading a list of samples, and configuring and running workflows.

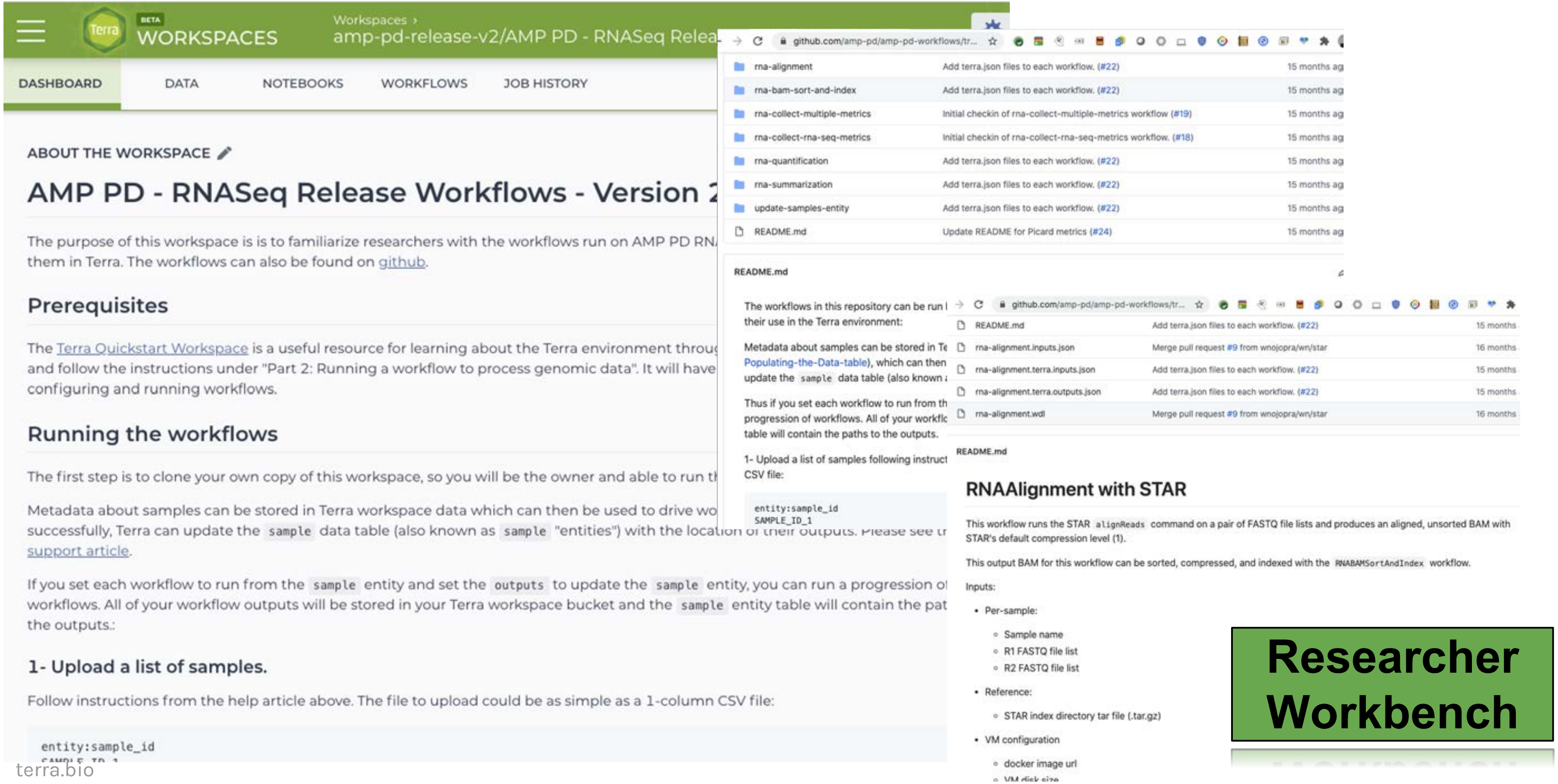
Running the workflows

The first step is to clone your own copy of this workspace, so you will be the owner and able to run the workflows.

Metadata about samples can be stored in Terra workspace data which can then be used to drive workflows. When workflows complete successfully, Terra can update the `sample` data table (also known as `sample` "entities") with the location of their outputs. Please see this Terra support article.

Researcher Workbench

AMP PD RNASeq Workflows workspace



ABOUT THE WORKSPACE

AMP PD - RNASeq Release Workflows - Version 2

The purpose of this workspace is is to familiarize researchers with the workflows run on AMP PD RN
them in Terra. The workflows can also be found on github.

Prerequisites

The Terra Quickstart Workspace is a useful resource for learning about the Terra environment throu
and follow the instructions under "Part 2: Running a workflow to process genomic data". It will have
configuring and running workflows.

Running the workflows

The first step is to clone your own copy of this workspace, so you will be the owner and able to run t

Metadata about samples can be stored in Terra workspace data which can then be used to drive wo
successfully, Terra can update the `sample` data table (also known as `sample` "entities") with the location of their outputs. Please see t
support article.

If you set each workflow to run from the `sample` entity and set the `outputs` to update the `sample` entity, you can run a progression o
workflows. All of your workflow outputs will be stored in your Terra workspace bucket and the `sample` entity table will contain the pat
the outputs.:

1- Upload a list of samples.

Follow instructions from the help article above. The file to upload could be as simple as a 1-column CSV file:

```
entity:sample_id
SAMPLE_ID_1
```

github.com/amp-pd/amp-pd-workflows/tr...

rna-alignment	Add terra.json files to each workflow. (#22)	15 months ag
rna-bam-sort-and-index	Add terra.json files to each workflow. (#22)	15 months ag
rna-collect-multiple-metrics	Initial checkin of rna-collect-multiple-metrics workflow (#19)	15 months ag
rna-collect-rna-seq-metrics	Initial checkin of rna-collect-rna-seq-metrics workflow. (#18)	15 months ag
rna-quantification	Add terra.json files to each workflow. (#22)	15 months ag
rna-summarization	Add terra.json files to each workflow. (#22)	15 months ag
update-samples-entity	Add terra.json files to each workflow. (#22)	15 months ag
README.md	Update README for Picard metrics (#24)	15 months ag

README.md

The workflows in this repository can be run i
their use in the Terra environment:

Metadata about samples can be stored in Te
Populating-the-Data-table), which can then
update the `sample` data table (also known a

Thus if you set each workflow to run from th
progression of workflows. All of your workflo
table will contain the paths to the outputs.

1- Upload a list of samples following instruct
CSV file:

```
entity:sample_id
SAMPLE_ID_1
```

github.com/amp-pd/amp-pd-workflows/tr...

README.md	Add terra.json files to each workflow. (#22)	15 months
rna-alignment.inputs.json	Merge pull request #9 from wnojopra/wn/star	16 months
rna-alignment.terra.inputs.json	Add terra.json files to each workflow. (#22)	15 months
rna-alignment.terra.outputs.json	Add terra.json files to each workflow. (#22)	15 months
rna-alignment.wdl	Merge pull request #9 from wnojopra/wn/star	16 months

README.md

RNAAlignment with STAR

This workflow runs the STAR `alignReads` command on a pair of FASTQ file lists and produces an aligned, unsorted BAM with STAR's default compression level (1).

This output BAM for this workflow can be sorted, compressed, and indexed with the `RNABAMSortAndIndex` workflow.

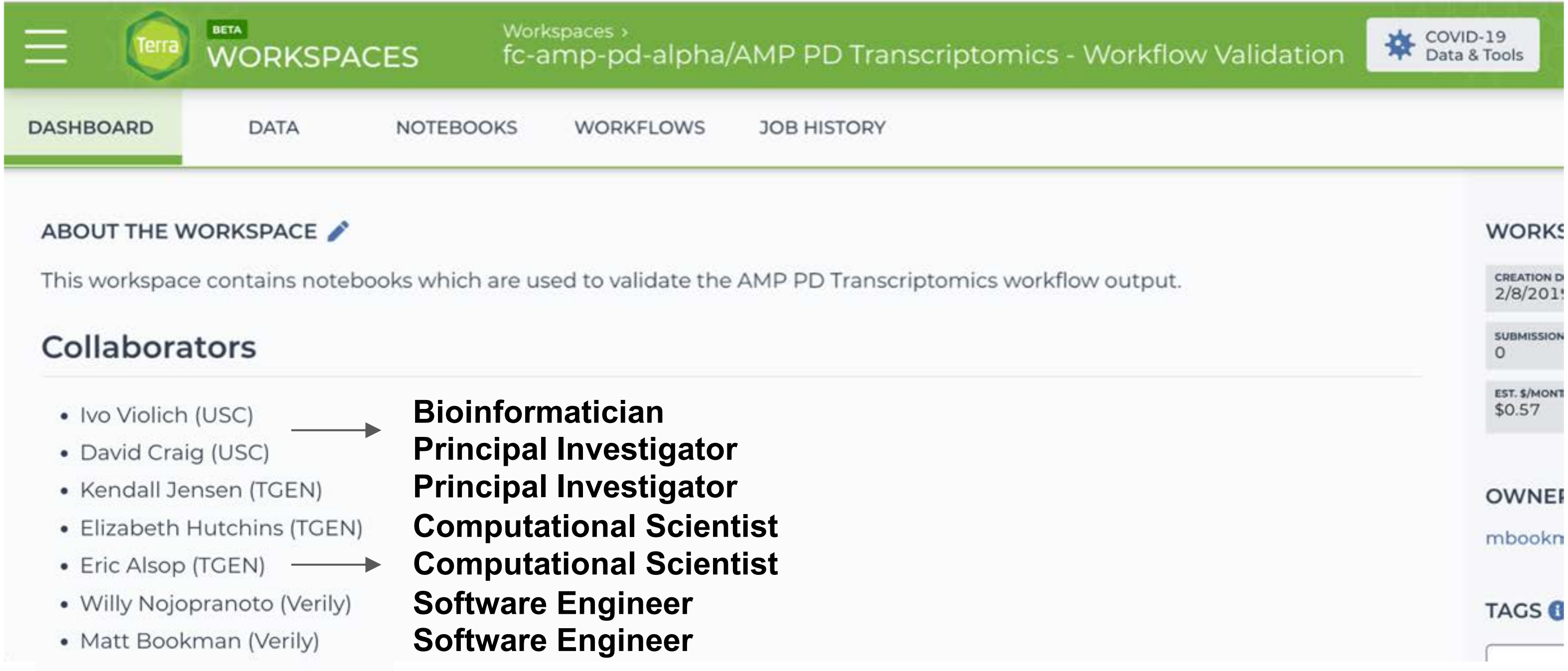
Inputs:

- Per-sample:
  - Sample name
  - R1 FASTQ file list
  - R2 FASTQ file list
- Reference:
  - STAR index directory tar file (.tar.gz)
- VM configuration
  - docker image url
  - VM disk size

Researcher Workbench

terra.bio

Workspace Collaboration (RNASeq QC)



ABOUT THE WORKSPACE

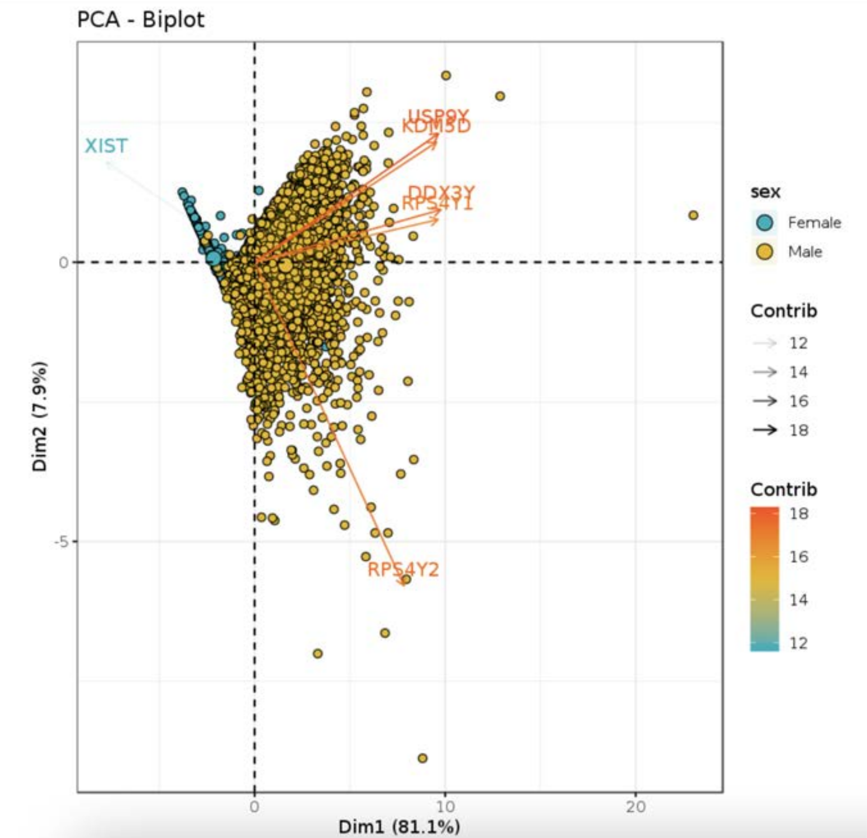
This workspace contains notebooks which are used to validate the AMP PD Transcriptomics workflow output.

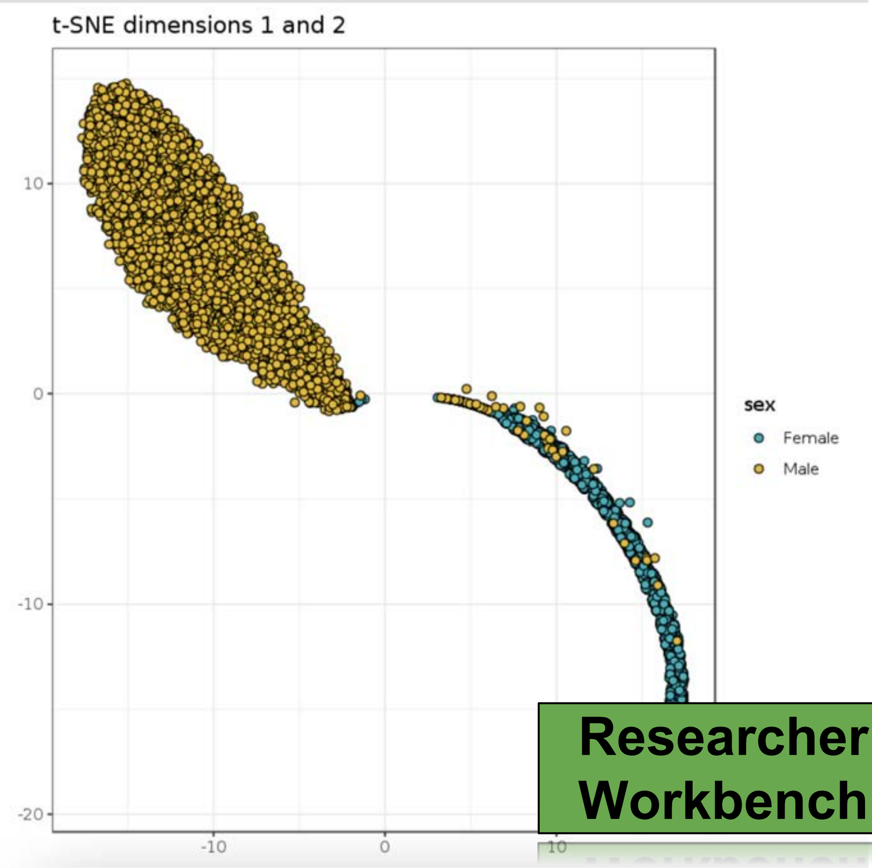
Collaborators

Ivo Violich (USC)	**Bioinformatician**
David Craig (USC)	**Principal Investigator**
Kendall Jensen (TGEN)	**Principal Investigator**
Elizabeth Hutchins (TGEN)	**Computational Scientist**
Eric Alsop (TGEN)	**Computational Scientist**
Willy Nojopranoto (Verily)	**Software Engineer**
Matt Bookman (Verily)	**Software Engineer**

Researcher Workbench

Notebook Collaboration (RNASeq QC)

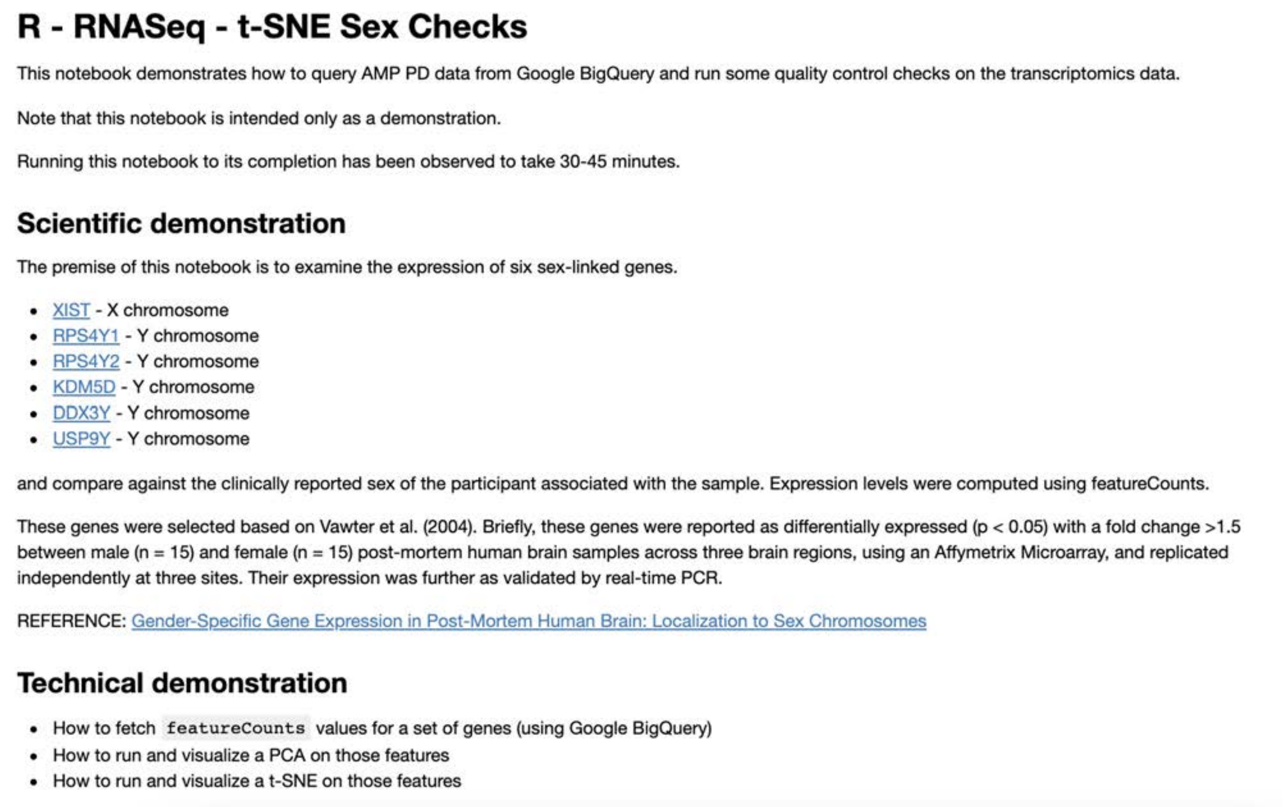




Researcher Workbench

terra.bio

Notebook Collaboration (RNASeq QC)



R - RNASeq - t-SNE Sex Checks

This notebook demonstrates how to query AMP PD data from Google BigQuery and run some quality control checks on the transcriptomics data.

Note that this notebook is intended only as a demonstration.

Running this notebook to its completion has been observed to take 30-45 minutes.

Scientific demonstration

The premise of this notebook is to examine the expression of six sex-linked genes.

- XIST - X chromosome
- RPS4Y1 - Y chromosome
- RPS4Y2 - Y chromosome
- KDM5D - Y chromosome
- DDX3Y - Y chromosome
- USP9Y - Y chromosome

and compare against the clinically reported sex of the participant associated with the sample. Expression levels were computed using featureCounts.

These genes were selected based on Vawter et al. (2004). Briefly, these genes were reported as differentially expressed ($p < 0.05$) with a fold change >1.5 between male ($n = 15$) and female ($n = 15$) post-mortem human brain samples across three brain regions, using an Affymetrix Microarray, and replicated independently at three sites. Their expression was further as validated by real-time PCR.

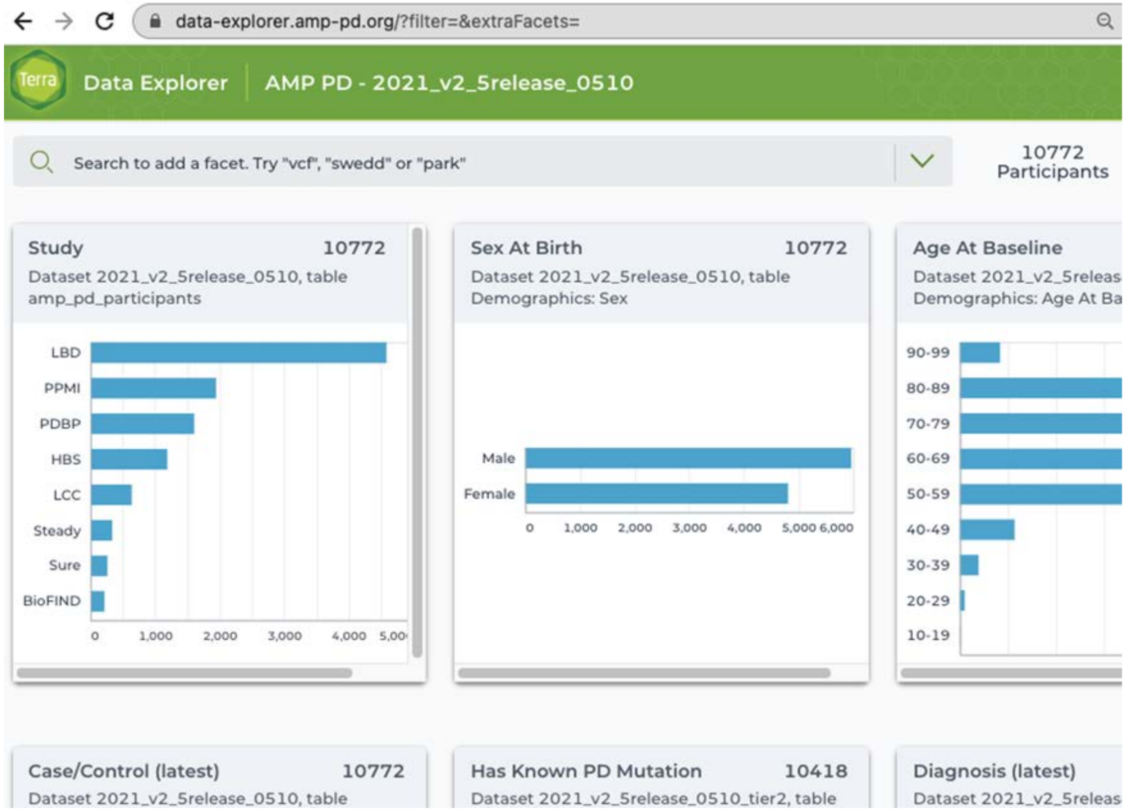
REFERENCE: Gender-Specific Gene Expression in Post-Mortem Human Brain: Localization to Sex Chromosomes

Technical demonstration

- How to fetch `featureCounts` values for a set of genes (using Google BigQuery)
- How to run and visualize a PCA on those features
- How to run and visualize a t-SNE on those features

Researcher Workbench

AMP PD Data Explorer



data-explorer.amp-pd.org/?filter=&extraFacets=

Terra Data Explorer | AMP PD - 2021_v2_5release_0510

Search to add a facet. Try "vcf", "swedd" or "park"

10772 Participants

Study 10772
Dataset 2021_v2_5release_0510, table amp_pd_participants

Sex At Birth 10772
Dataset 2021_v2_5release_0510, table Demographics: Sex

Age At Baseline
Dataset 2021_v2_5releas
Demographics: Age At Ba

Case/Control (latest) 10772
Dataset 2021_v2_5release_0510, table

Has Known PD Mutation 10418
Dataset 2021_v2_5release_0510_tier2, table

Diagnosis (latest)
Dataset 2021_v2_5releas

- Data Explorers are a category of
  - dataset-specific tools
  - general-purpose tools configured for the dataset
- Allows for interactive tools focused on your dataset; can account for cohort and datatype nuances
- Easy to deploy behind dataset-specific access groups
- Different tools can target different personas

Data Explorers

Case Study: Integrative Analysis with AMP PD Transcriptomics Data

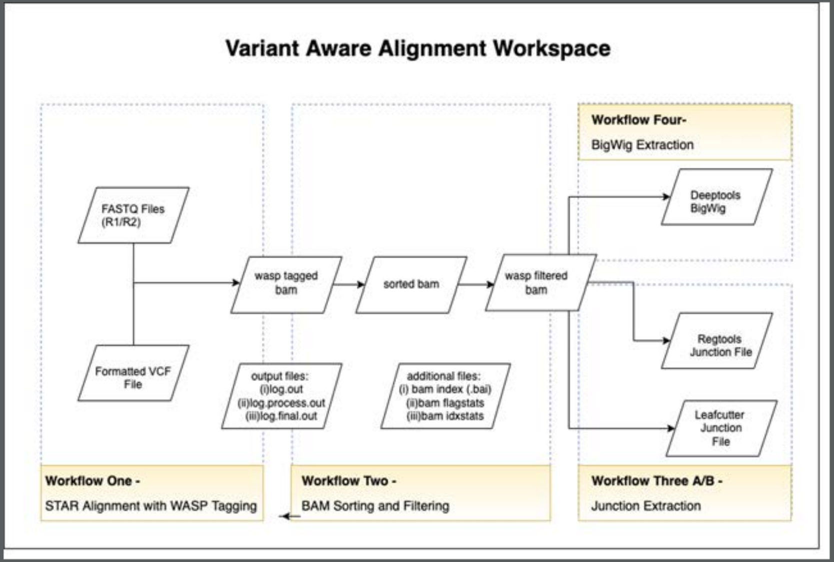
Integrative Analysis and Visualization of AMP-PD Transcriptomic Data

David W. Craig, Ph.D.

USC Institute Of
Translational Genomics

Ivo Violich

Ramiyapriya Sivakumar



Variant Aware Alignment Workspace

FASTQ Files (R1/R2)
Formatted VCF File
wasp tagged bam
sorted bam
wasp filtered bam
output files: (i)log.out (ii)log.process.out (iii)log.final.out
additional files: (i) bam index (.bai) (ii)bam flagstats (iii)bam idxstats
Deeptools BigWig
Regtools Junction File
Leafcutter Junction File

Workflow One - STAR Alignment with WASP Tagging
Workflow Two - BAM Sorting and Filtering
Workflow Three A/B - Junction Extraction
Workflow Four- BigWig Extraction

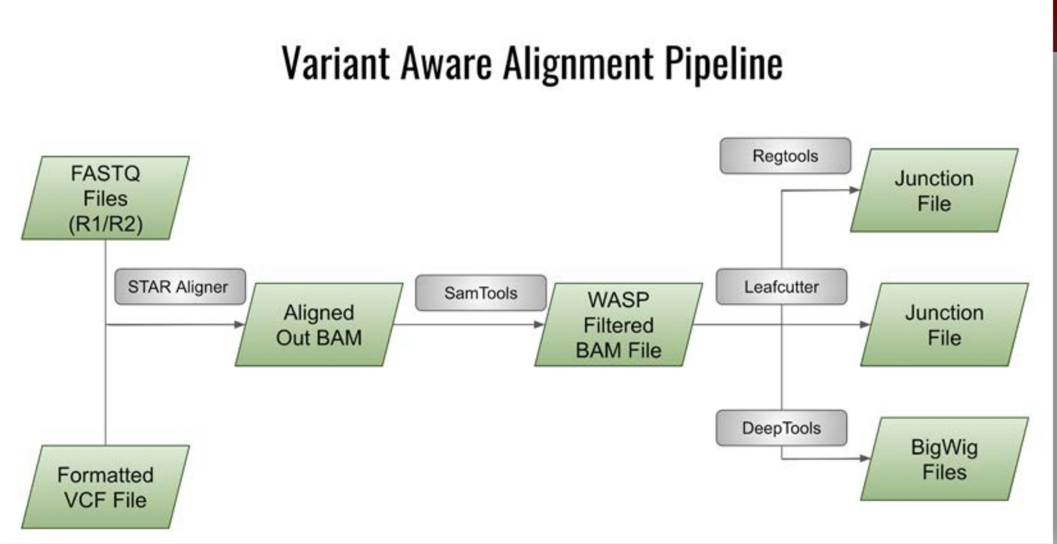
Overview

Goal:

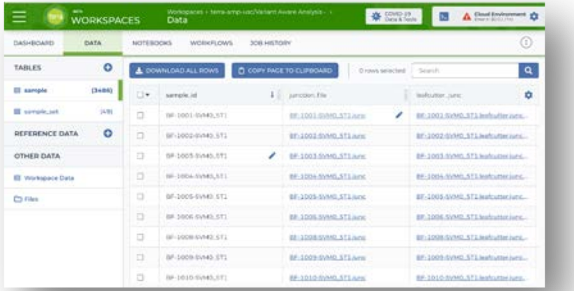
- Full Variant aware alignment w/ VCFs
- Development of a visualization Platform

Analysis Considerations

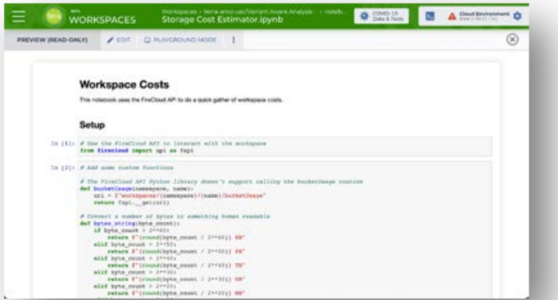
- Full Implementation in Terra
- Samples: 8,357 @ 200M 125bp Reads
- Bases: 208.9 Terabases
- Size: >200 Tbtes
- Files: >25,000 Files



Variant Aware Alignment Pipeline



sample_id	junction_file	leafcutter_junc
BF-1001-SVM0_ST1	BF-1001.SVM0_ST1.junc	BF-1001.SVM0_ST1.leafcutter.junc...
BF-1002-SVM0_ST1	BF-1002.SVM0_ST1.junc	BF-1002.SVM0_ST1.leafcutter.junc...
BF-1003-SVM0_ST1	BF-1003.SVM0_ST1.junc	BF-1003.SVM0_ST1.leafcutter.junc...
BF-1004-SVM0_ST1	BF-1004.SVM0_ST1.junc	BF-1004.SVM0_ST1.leafcutter.junc...
BF-1005-SVM0_ST1	BF-1005.SVM0_ST1.junc	BF-1005.SVM0_ST1.leafcutter.junc...
BF-1006-SVM0_ST1	BF-1006.SVM0_ST1.junc	BF-1006.SVM0_ST1.leafcutter.junc...
BF-1008-SVM0_ST1	BF-1008.SVM0_ST1.junc	BF-1008.SVM0_ST1.leafcutter.junc...
BF-1009-SVM0_ST1	BF-1009.SVM0_ST1.junc	BF-1009.SVM0_ST1.leafcutter.junc...
BF-1010-SVM0_ST1	BF-1010.SVM0_ST1.junc	BF-1010.SVM0_ST1.leafcutter.junc...

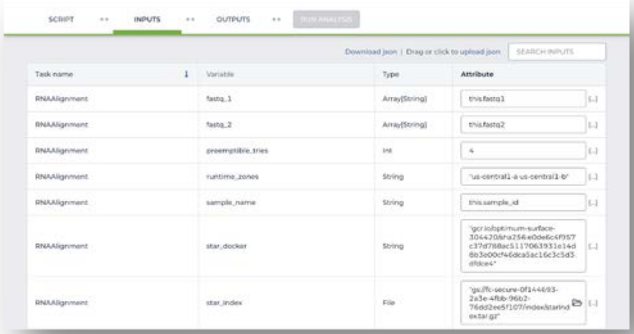


Workspace Costs

This notebook uses the FireCloud API to do a quick gather of workspace costs.

Setup

```
In [1]: # Use the FireCloud API to interact with the workspace
        from firecloud import api as fapi
```

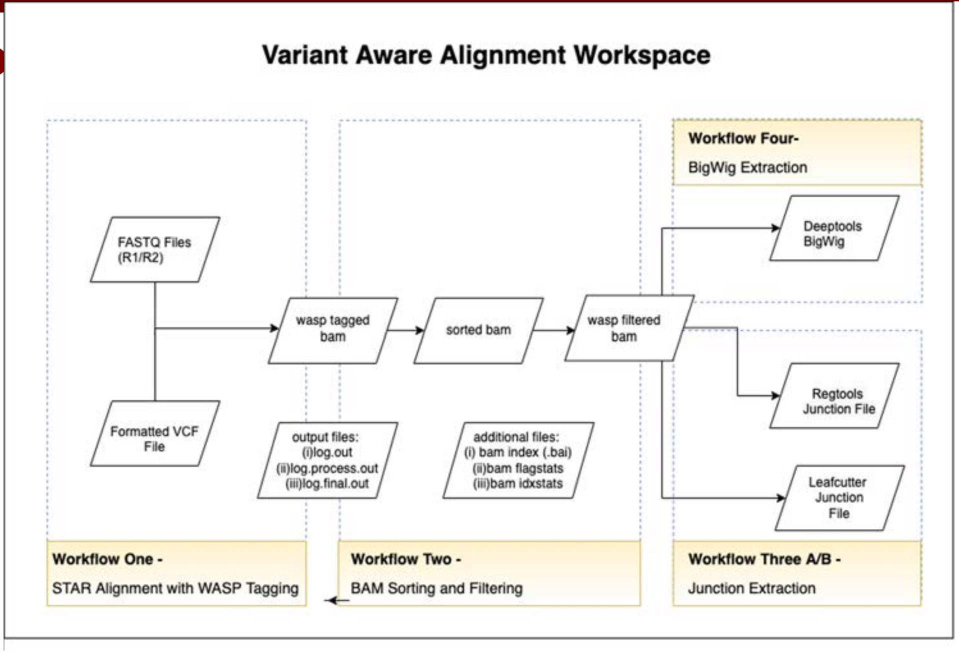


Task name	Variable	Type	Attribute
RNAAlignment	fastq_1	Array[String]	this.fastq1
RNAAlignment	fastq_2	Array[String]	this.fastq2
RNAAlignment	preemptible_tries	Int	4
RNAAlignment	runtime_zones	String	"us-central1-a us-central1-b"
RNAAlignment	sample_name	String	this.sample_id
RNAAlignment	star_docker	String	"gcr.io/optimum-surface-304420/sha256:e0de6c4f957c37d788ac5117063931e14d8b3e00cf46dca5ac16c3c5d3dfdce4"
RNAAlignment	star_index	File	"gs://fc-secure-0f144693-2a3e-4fbb-96b2-76dd2ee5f107/index/starindex.tar.gz"

Focus on Costs

Focused Development/Optimizatio

- Initial 100 samples & Optimization
- Focus on CPU costs
  - Rounds of identifying optimal CPU/storage
  - Pre-emptible CPUs allowed for 80% reduction in costs
- Expanded to larger set
  - About 2-5% failure rate for jobs to run (due to pre-emtible



File Sizes

- FASTQ: 4 GB ~ 33 Tbytes
- BAM: 4 Gb ~ 33 Tbytes
- Tables: 1-2 Tbytes


Cost Estimate for Variant Aware Alignment

Variant Aware Alignment Run Cost Estimate	Cost Per Sample	Number of Samples	Estimated Cost
Workflow One - Converting BAM to FASTQ	$0.42	8356	$3,509.52
Workflow Two - Variant Aware STAR Alignment	$0.34	8356	$2,841.04
Workflow Three - Filtering BAM for WASP Tags	$0.11	8356	$919.16
Workflow Four - Converting BAM to JUNC with Leafcutter	$0.04	8356	$334.24

Storage

- ~70 Tb @ 0.020 Gb/Mo is $1400 Mo Standard

Notes

- Egres
  - (data b/t cold-line & offsite could be significant)

Visualization Platform

Allow Gene Level Viewing / Downloading

- Allow gene/region browsing integrating the multi-omic data/analysis & clinical harmonization
- Content Appropriate To Tier 1/Tier 2 access

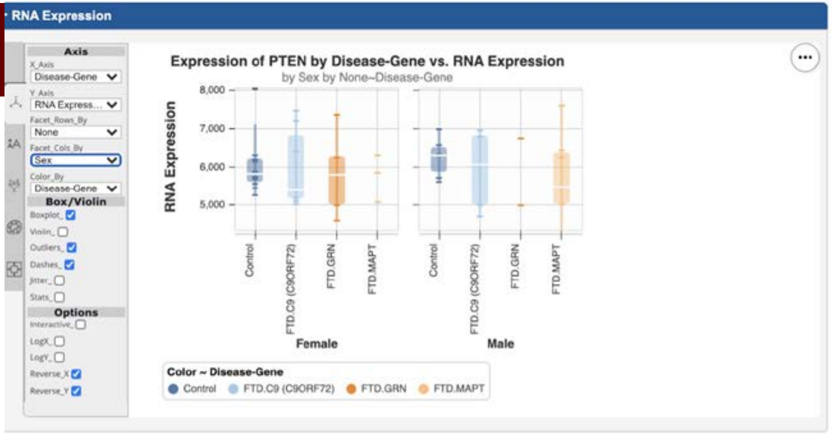
Component 1: Graph Builder (Gene/Transcript Level)

- Dynamic graph building w/ 10+ Types of Joined Data by Gene/Transcript/SNP

Component 2: Genome Browser (Region Level)

- Multiple Tracks/Interactivity/Indiviudal & Summary Level Data

Integration With Other (FOUNDIN-PD)


RNA Expression

Axis

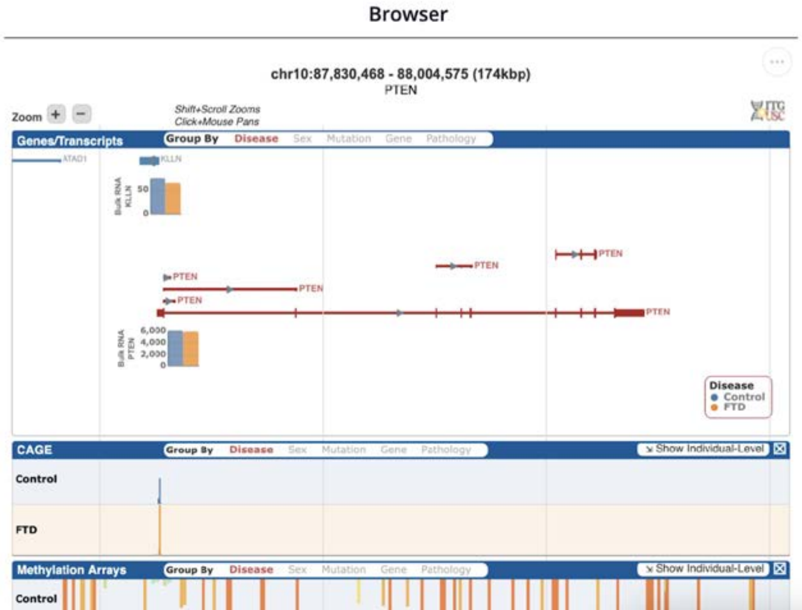
X_Axis: Disease-Gene
Y_Axis: RNA Express...
Facet_Rows_By: None
Facet_Cols_By: Sex
Color_By: Disease-Gene

Box/Violin

Boxplot_ ☑
Violin_ ☐
Outliers_ ☑
Dashes_ ☑
Jitter_ ☐
Stats_ ☐

Options

Interactive_ ☐
LogX_ ☐
LogY_ ☐
Reverse_X ☑
Reverse_Y ☑

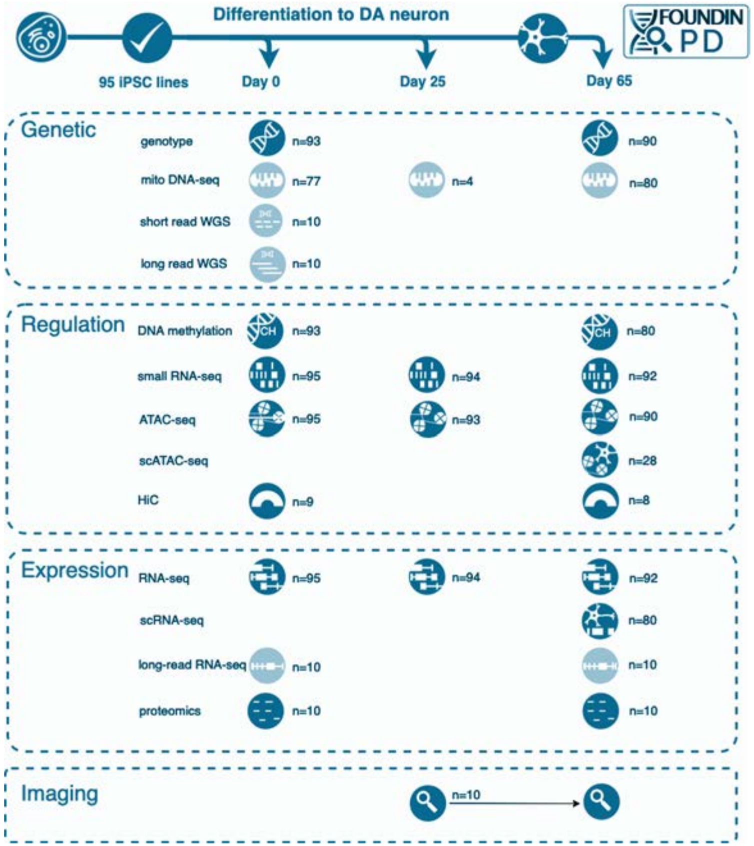


INTEGRATION W/ OTHER PROJECTS

FOUNDIN PD

- AMP-PD
  - Whole-blood RNA
  - Whole-blood DNA
  - >4,000 individuals

- FOUNDIND PD
  - 96 IPSCs
  - Overlapping Donors
  - Multiomics



Differentiation to DA neuron

		95 iPSC lines Day 0	Day 25	Day 65
Genetic	genotype	n=93		n=90
	mito DNA-seq	n=77	n=4	n=80
	short read WGS	n=10		
	long read WGS	n=10		
Regulation	DNA methylation	n=93		n=80
	small RNA-seq	n=95	n=94	n=92
	ATAC-seq	n=95	n=93	n=90
	scATAC-seq			n=28
	HiC	n=9		n=8
Expression	RNA-seq	n=95	n=94	n=92
	scRNA-seq			n=80
	long-read RNA-seq	n=10		n=10
	proteomics	n=10		n=10
Imaging			n=10 →	

Team

GLADSTONE INSTITUTES

Keck Medicine of USC

tgen AN AFFILIATE OF City of Hope

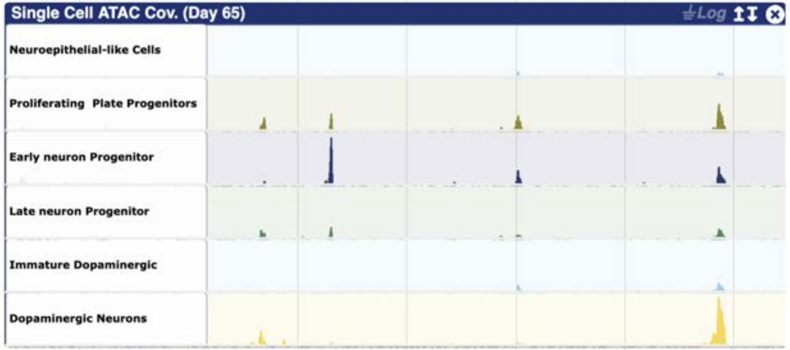
DZNE Deutsches Zentrum für Neurodegenerative Erkrankungen in der Helmholtz-Gemeinschaft

NIH National Institute on Aging *Intramural Research Program*

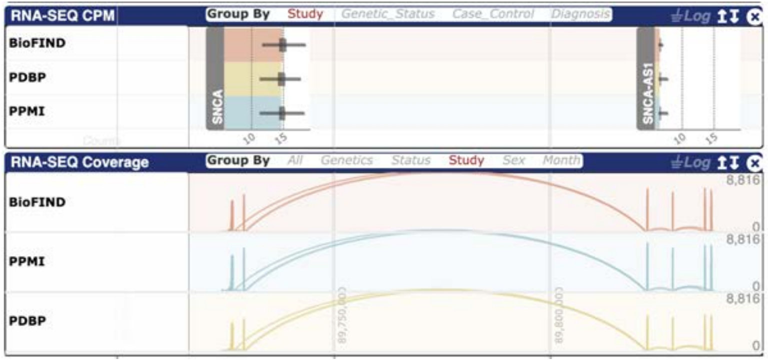
THE MICHAEL J. FOX FOUNDATION FOR PARKINSON'S RESEARCH

INTEGRATION W/ OTHER PROJECTS

AMP-PD



Foundinpd.org

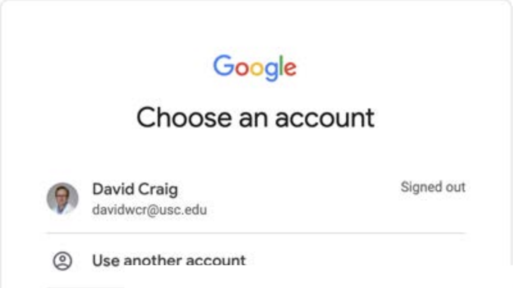


Browser

Tracks Allow Individual Integration

IDENTIFICATION: *IDENTITY AWARE MANAGEMENT*

Single Sign On



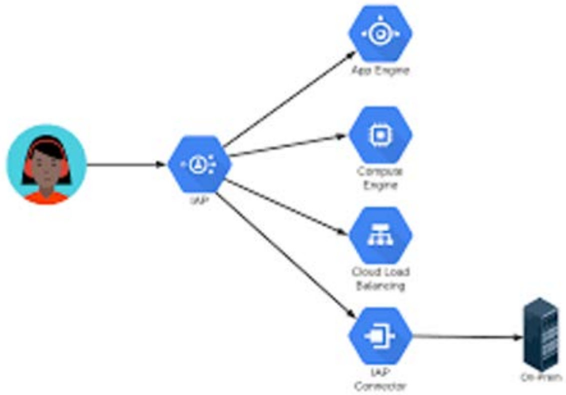
Google

Choose an account

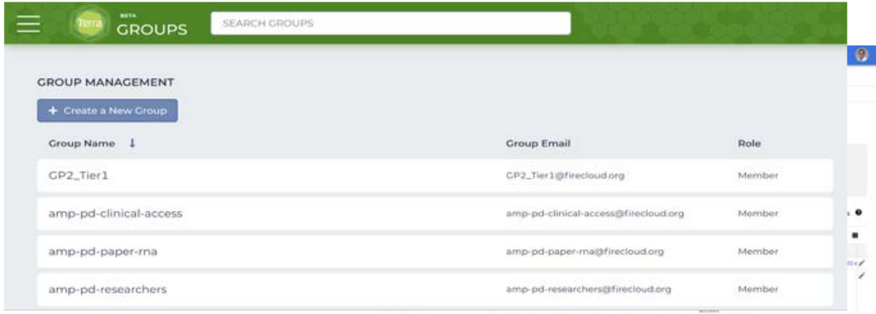
David Craig
davidwcr@usc.edu
Signed out

Use another account

Rem



Groups Management

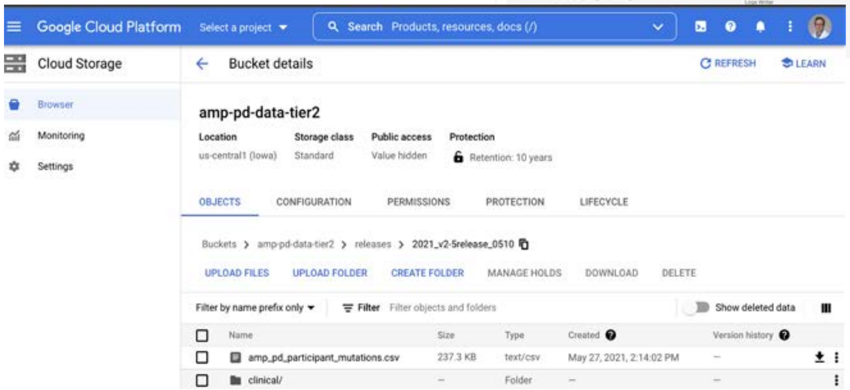


GROUP MANAGEMENT

+ Create a New Group

Group Name ↓	Group Email	Role
GP2_Tier1	GP2_Tier1@firecloud.org	Member
amp-pd-clinical-access	amp-pd-clinical-access@firecloud.org	Member
amp-pd-paper-rna	amp-pd-paper-rna@firecloud.org	Member
amp-pd-researchers	amp-pd-researchers@firecloud.org	Member

Buckets



Bucket details

amp-pd-data-tier2

Location	Storage class	Public access	Protection
us-central1 (Iowa)	Standard	Value hidden	Retention: 10 years

Buckets > amp-pd-data-tier2 > releases > 2021_v2-5release_0510

Filter by name prefix only ▾ Filter Filter objects and folders Show deleted data

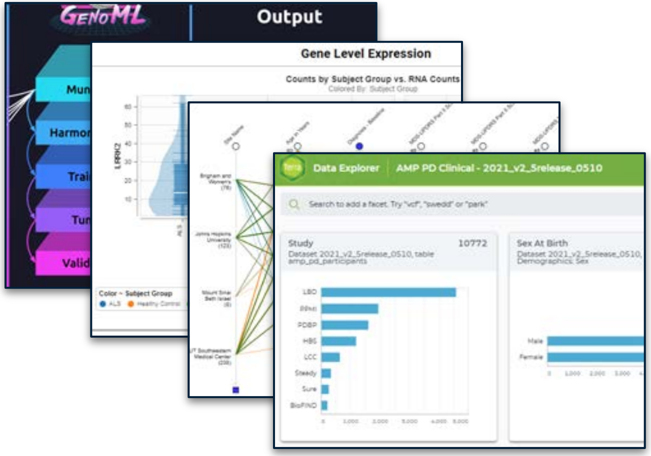
☐	Name	Size	Type	Created	Version history
☐	amp_pd_participant_mutations.csv	237.3 KB	text/csv	May 27, 2021, 2:14:02 PM	–
☐	clinical/	–	Folder	–	–

Enabling Community Contributions: Post-launch activities

Different Use Cases and User Profiles

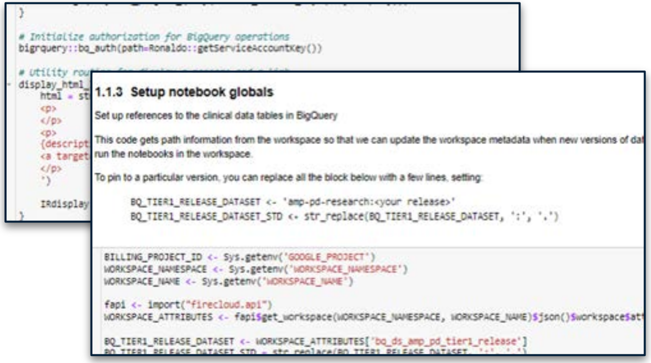
Data Explorers and Visualization Tools

- AMP PD Data Explorer
- Gene Explorer
- GenoML Model Explorer
- Parallel Coordinates Clinical Explorer
- Extensions via Workspaces



Workspaces and Notebooks

- Basic Mechanics
- Proteomics, Genomics, Transcriptomics Analyses
- Multiple Omics Analyses
- Data Production Workflows
- Data Production QC Analyses



```
}

# Initialize authorization for BigQuery operations
bigrquery::bq_auth(path=Ronaldo::getServiceAccountKey())

# Utility rou
display_html_
    html = st
    <p>
    </p>
    <p>
    {descript
    <a target
    </p>
    ')

    IRdisplay
}
```

1.1.3 Setup notebook globals

Set up references to the clinical data tables in BigQuery

This code gets path information from the workspace so that we can update the workspace metadata when new versions of dat
run the notebooks in the workspace.

To pin to a particular version, you can replace all the block below with a few lines, setting:

```
    BQ_TIER1_RELEASE_DATASET <- 'amp-pd-research:<your release>'
    BQ_TIER1_RELEASE_DATASET_STD <- str_replace(BQ_TIER1_RELEASE_DATASET, ':', '.')
```

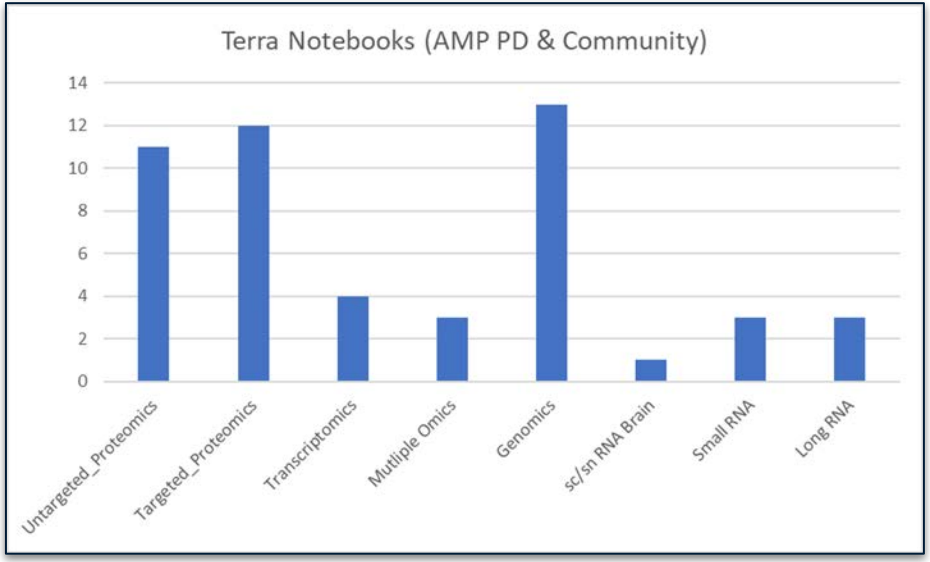
```
BILLING_PROJECT_ID <- Sys.getenv('GOOGLE_PROJECT')
WORKSPACE_NAMESPACE <- Sys.getenv('WORKSPACE_NAMESPACE')
WORKSPACE_NAME <- Sys.getenv('WORKSPACE_NAME')

fapi <- import("firecloud.api")
WORKSPACE_ATTRIBUTES <- fapi$get_workspace(WORKSPACE_NAMESPACE, WORKSPACE_NAME)$json()$workspace$att

BQ_TIER1_RELEASE_DATASET <- WORKSPACE_ATTRIBUTES['bq_ds_amp_pd_tier1_release']
```

Building the AMP PD Community

Transitioning Data Production to Resource Ecosystem

50+ Workspaces Underway

- AMP PD Official Workspaces
- Partners and Collaborators
- Funded Grantees
- User Community
  - Vendors
  - Analysts



Encouraging Contributions

- Outreach
- Training and Webinars
- Grants
- Subsidized Workspaces

User Workspace → Subsidized Data → AMP PD Official

Enabling Collaboration

Key Components of the AMP PD Resource Ecosystem

Encouraging Contributions

- Well Organized Data
- Flexible Infrastructure
- Simple Access Management
- Getting-Started Examples
- Compelling Analyses
- Active Community

Community Building

Outreach, training, grants, and subsidies

Terra Platform and Explorers

Workspaces, workflows, notebooks and visualizations

Google Cloud Infrastructure

Core infrastructure for storage and compute

Research Data

Organized data harmonized across datasets

It Takes a Village

Steering Committee:
- Deb Babcock (NIH/NINDS) – Co-Chair
- Pablo Sardi (Sanofi) – Co-Chair
- Billy Dunn (FDA)
- David Glazer (Verily)
- Ekemini Riley (ASAP)
- Ellen Gadbois (NIH/OD)
- Ellen Wann (NIH/OD)
- Gerald (Dave) Podskalny (FDA)
- Gopi Ganji (GSK)
- Guhan Nagappan (GSK)
- Leslie Shinobu (BMS)
- Lyn Jakemen, (NINDS)
- Pat Bellgowan (NINDS)
- Robert Moccia (Pfizer)
- Sonya Dumanis (ASAP)
- Shameek Biswas (BMS)
- Suzana Petanceska (NIA)
- Todd Sherer (MJFF)
- Walter Koroshetz (NIH/NINDS)
- William Marks (Verily)

Ad-hoc/Guest
- Barry Landin (FNIH Contractor)
- Dave Vismer (FNIH Contractor)
- Irit Rapley (BMS)
- Matt Bookman (Verily)
- Richard Hargreaves (BMS)

Metabolomics Working Group
- Deb Babcock (NINDS) – Co-Chair
- Lilu Guo (Sanofi) – Co-Chair
- Barry Landin (FNIH Contractor)
- Chrissa Dwyer (GSK)
- Eline Appelmans (FNIH)
- Kevin Jensen (Celgene/BMS)
- Marcus Bantscheff (GSK)
- Mark Frasier (MJFF)
- Suzana Petanceska (NIA)

Ad-hoc
- Rima Kaddurah-Daouk (Duke)
- Michael Schwarzchild (Harvard)

Data Usage and Policies Working Group
- Alyssa Reimer (MJFF) – Co-Chair
- Deb Babcock (NINDS) – Co-Chair
- Claire Wegel (IU/GP2)
- Clemens Scherzer (BWH)
- Ekemini Riley (ASAP)
- Eline Appelmans (FNIH)
- Matt Bookman (Verily)
- Yuliya Kuras (Harvard)
- Sonja Scholz (NIA – LBD)
- Dave Alonso (MJFF – LCC)

Outreach and Portal Working Group
- Patrick Bellgowan(NINDS) – Co-Chair
- Leslie Kirsch (MJFF) – Co-Chair
- Suzanne Beschamps (FNIH)
- Carl Wonders (NINDS)
- Chris Swanson-Fischer (NINDS)
- David Alonso (MJFF)
- Deb Babcock (NINDS)
- Eline Appelmans (FNIH)
- Kathleen Parkes (Verily)
- Kristine Treece (NINDS Contractor)
- Sumit Dey (QMUL/GP2)

Ad-hoc
- Matt Bookman (Verily)
- Barbara Marebwa (MJFF)
- Amy Adams (NINDS)
- Barry Landin (FNIH Contractor)
- Katherine Thompson (FNIH)
- Nadia Douaji (NINDS)
- Ryan Stewart-Frederick (FNIH Contractor)

RNA Seq Brain Working Group
- Deb Babcock (NINDS) – Co-Chair
- Kevin Jensen (Celgene/BMS)– Co-Chair
- Barry Landin (FNIH Contractor)
- Bradford Casey (MJFF)
- Charissa Dwyer (GSK)
- Eline Appelmans (FNIH)
- David Craig (USC)
- Dinesh Kumar (Sanofi)
- Suzana Petanceska (NIA)

Data Analyses Working Group
- Bradford Casey (MJFF) – Co-Chair
- Hirotaka Iwaki (NIA) – Co-Chair
- Alyssa Reimer (MJFF)
- Andrew Sharp (MSSM)
- Andy Singleton (NIA/LNG)
- Christoph Bussler (Verily)
- Clemens Scherzer (BWH)
- David Alonso (MJFF)
- David Craig (USC)
- David Knowles (NYGenome)
- David Vismer (FNIH Contractor)
- Deb Babcock (NINDS)
- Eline Appelmans (FNIH)
- Elizabeth Hutchins (TGEN)
- Erin Teeple (Sanofi)
- Hampton Leonard (NIA/NINDS)
- Ivo Violich (USC)
- Jie Yuan (BWH)
- Kendall Jenssen (TGEN)
- Mary Makarious (NIA/NINDS)
- Matt Bookman (Verily)
- Meaghan Cogswell (Sanofi)
- Mike Nalls (NIA/NINDS)
- Ninad Amondikar (MJFF)
- Ramsey Magana (GSK)
- Ramiya Sivakumar (USC)
- Ricardo Vialle (MSSM)
- Rob Woodruff (GSK)
- Ruifeng Hu (BWH)
- Samantha Hutten (MJFF)
- Shameek Biswas (BMS)
- Srini Shankara (Sanofi)
- Towfique Raj (MSSM)
- Xianjun Dong (BWH)

Clinical Data Harmonization Working Group
- Alyssa Reimer (MJFF) – Co-Chair
- Deb Babcock (NINDS) – Co-Chair
- Barry Landin (FNIH Contractor)
- Clemens Scherzer (BWH)
- David Pulford (GSK)
- David Vismer (FNIH Contractor)
- Eline Appelmans (FNIH)
- Erin Teeple (Sanofi)
- Mary Makarious (NIA/NINDS)
- Matt Bookman (Verily)
- Nicole Glazer (Celgene/BMS)
- Robert Moccia (Pfizer)
- Yuliya Kuras (Harvard)

WGS and TS Subgroup
- Andy Singleton (NIA/LNG) – WGS Co-Chair
- Shameek Biswas (Celgene/BMS) – WGS Co-Chair
- Kendall Jensen (TGen) – Co-chair
- Deb Babcock (NINDS) – Co-Chair (Temp)
- Ashutosh Pandey (GSK)
- Barry Landin (FNIH Contractor)
- Bradford Casey (MJFF)
- Christine Swanson-Fischer (NINDS)
- Christoph Bussler (Verily)
- Daniel Seaton (GSK)
- David Craig (USC)
- David Pulford (GSK)
- David Vismer (FNIH Contractor)
- Dena Hernandez (NIA)
- Eline Appelmans (FNIH)
- Elizabeth Hutchins (TGen)
- Guhan Nagappan (GSK)
- Hampton Leonard (NIA)
- Hirotaka Iwaki (NIA)
- Ivo Violich (USC)
- Jerrod Schwartz (Verily)
- Jie Yuan (BWH)
- Lyn Jakeman (NINDS)
- Mark Frasier (MJFF)
- Mary Makarious (NIA/NINDS)
- Matt Bookman (Verily)
- Matt Edwards (Verily)
- Meaghan Cogswell (Sanofi)
- Mike Nalls (NIA)
- Ramsey Magana (GSK)
- Robert Moccia (Pfizer)
- Robert Woodruff (GSK)
- Ruifeng Hu (BWH)
- Shawn Levy (Hudson Alpha)
- Sonja Scholz (NIA)
- Srini Shankara (Sanofi)
- Willy Nojopranoto (Verily)
- Xianjun Dong (BWH)

Access Compliance Team
Admin Support Team
Operations Team
Publication Subcommittee
and all the alumni and external experts that have helped us along the way

Proteomics Working Group
- Deb Babcock (NINDS) – Co-Chair
- Samantha Hutten (MJFF) – Co-Chair
- Amilcar Flores-Morales (Sanofi)
- Andy Christoforou (BMS)
- Bailin Zhang (Sanofi)
- Barry Landin (FNIH Contractor)
- Bradford Casey (MJFF)
- Christine Swanson-Fischer (NINDS)
- Eline Appelmans (FNIH)
- Howard Schulman (KOL)
- Lyn Jakeman (NINDS)
- Marcus Bantscheff (GSK)
- Pablo Sardi (Sanofi)

Umbrella Data Working Group
- Deb Babcock (NINDS) – Co-Chair
- Matt Bookman (Verily) – Co-Chair
- Alyssa Reimer (MJFF)
- Andy Singleton (NIA/LNG)
- Arthur Toga (USC)
- Bailin Zhang (Sanofi)
- Barry Landin (FNIH Contractor)
- Bradford Casey (MJFF)
- Codrin Lungu (NINDS)
- David Craig (USC)
- Eline Appelmans (FNIH)
- Erin Teeple (Sanofi)
- Juan Botia (KCL)
- Lungu Codrin (NINDS)
- Mike Nalls (NIA, LNG)
- Pablo Sardi (Sanofi)
- Ritu Kapur (Verily)
- Robert Moccia (Pfizer)
- Shameek Biswas (BMS)
- Willy Nojopranoto (Verily)

Special thanks to all cohort contributors & study participants

FNIH

AMP PD is made possible by the Public and Private Partners and members who have worked tirelessly to make this valuable resource available to the scientific community.