Update: Genomic Analysis Workflow in the (Fire)Cloud

Michael S. Noble
Broad Institute of MIT & Harvard

CPTAC3 Steering Committee Meeting
Bethesda, Maryland
2017_04_07

FIREHOSE
Broad GDAC

Played Several Key roles in TCGA

- very large scale production analysis pipeline
- analytic forest-clearing for researchers & MDs
    - starting state or 2nd opinion for AWGs
- democratization for use beyond TCGA proper
- simplification for everyone
- pushing envelope for rigor @ scale, reproducibility, APIs

Helping TCGA to usher in era of large-scale science, and to serve as model for future ambitious initiatives such as the Genome Data Analysis Network (GDAN)

Not just a dumb, crank-turning automaton

Novel discoveries lurk in Firehose outputs

Example: APOBEC cytidine deaminase(s) are major source of mutations in several cancers

Code developed by Gordenin & Klimczak et al (NIH)

They wanted large-scale testbed, ideally all TCGA

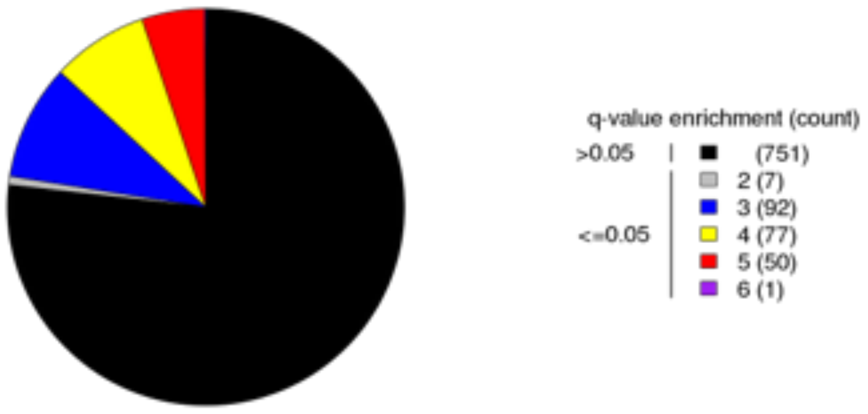
We did not have algorithm expertise

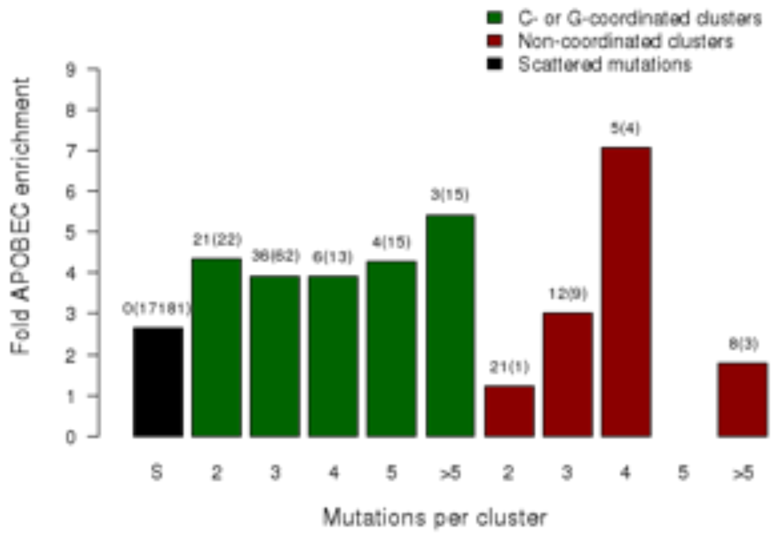
So we collaborated on Firehose install

Leading to numerous publications

Middlebrooks et al, Nat. Gen. 2016 http://dx.doi.org/10.1038/ng.3670

BRCA





However

- The cancer community still lacked a consensus, open-access and fully collaborative solution for *extreme-scale integrative analysis* *
- Within AWGs, discrepancies between informatic systems, data and analyses often had to be reconciled–at significant time cost–to prevent faulty science
- Central data coordination not always nimble as project needs evolved (e.g. aggregate cohorts like COADREAD, GBM+LGG, KIPAN)
- Correlative studies linking molecular findings to patient outcome were limited by the heterogeneous nature of samples (e.g. clinical)
- While online sharing markedly increased during TCGA, results continued to be generated on local compute
- Often with discrepant sample sets and unpublished codes, leaving reproducibility a difficult and largely unsolved problem

* **Firehose could only be enhanced & operated by Broadies**

Things getting bigger, faster, more complex & integrated

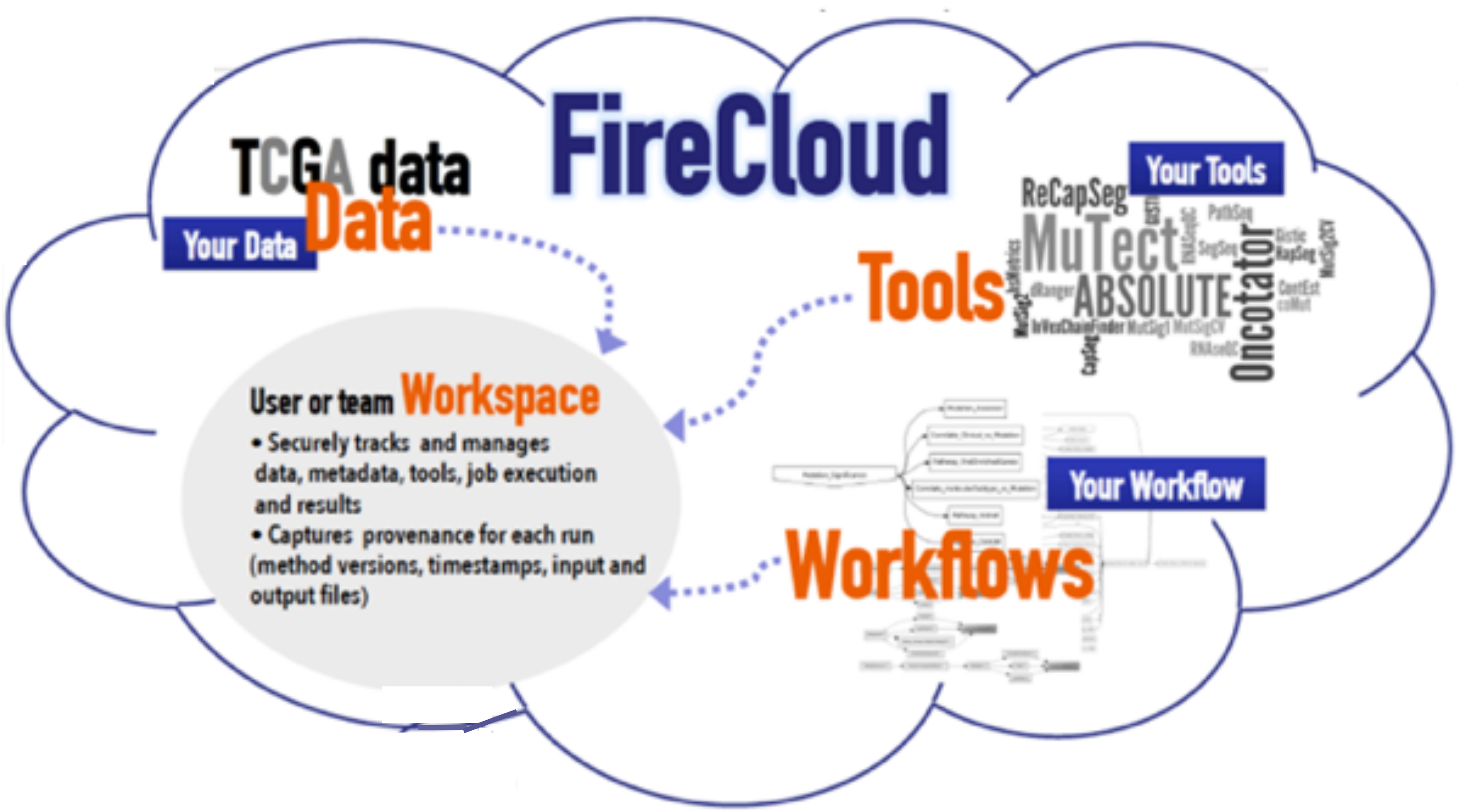
NCI Genome Data Analysis Network (GDAN)

- Collect & analyze ~10K samples in ~2 years (2017–>)
- This took TCGA 6-8 years
- **CPTAC3 & GTEX & Precision Medicine …**
- Should be possible to do JUST IN TIME analyses (as sample trickle incrementally accrues)

There is *no fundamental reason* why these efforts need be conducted as *walled-off silos*, instead of learning & leveraging from & sharing with each other

Stockholm Syndrome: we do not have to be agreeable captives to volume & complexity

Global Infrastructure for Collaborative Extreme-Scale Scientific Analysis

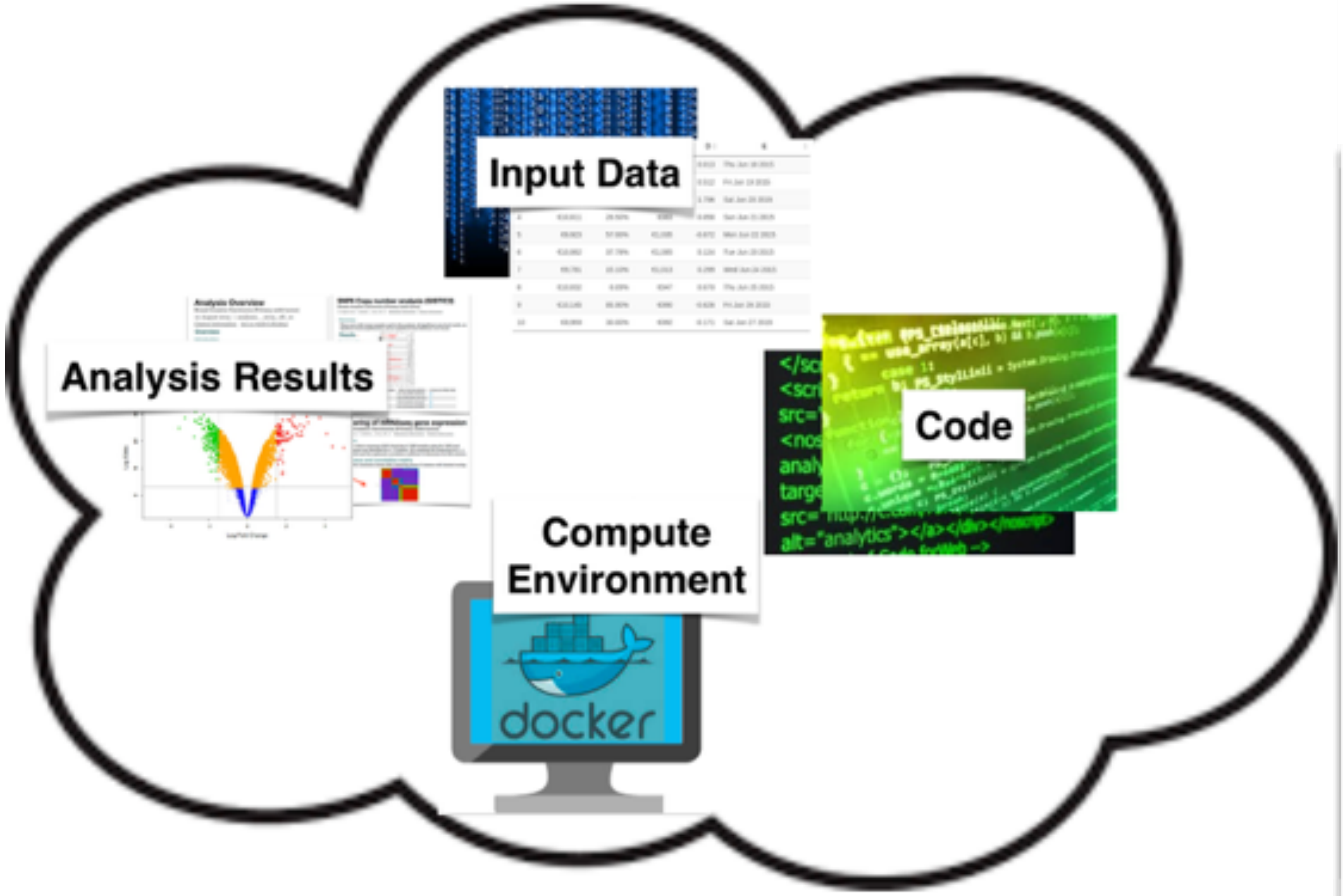


firecloud.org

- Collaborate via shared data, codes, analysis results & compute
- Like Google docs+drive already used to collaborate for papers
- **Trusted digital research assistant:** as samples accrue, perform mundane front line processing & analysis on our behalf, liberating our minds to synthesize the next wave of tools & theories

Ugh: WHY another system????

Extreme scale backend, already to 1000s WGS *(Broad GP)*
Simple URL access, oblivious to location of storage & compute
Pre-loaded with Open and Protected TCGA data
Point click ease, even for MDs, PIs & itinerant users
Eliminate TCGA bottleneck: Firehose for everyone, not just Broad



Manuscript analysis workspace

Slay Reproducibility Dragon

Freeze analysis workspace

Make public @ paper acceptance

Using DOIs as in Firehose runs

Paper cites URL to workspace

Contains everything in paper

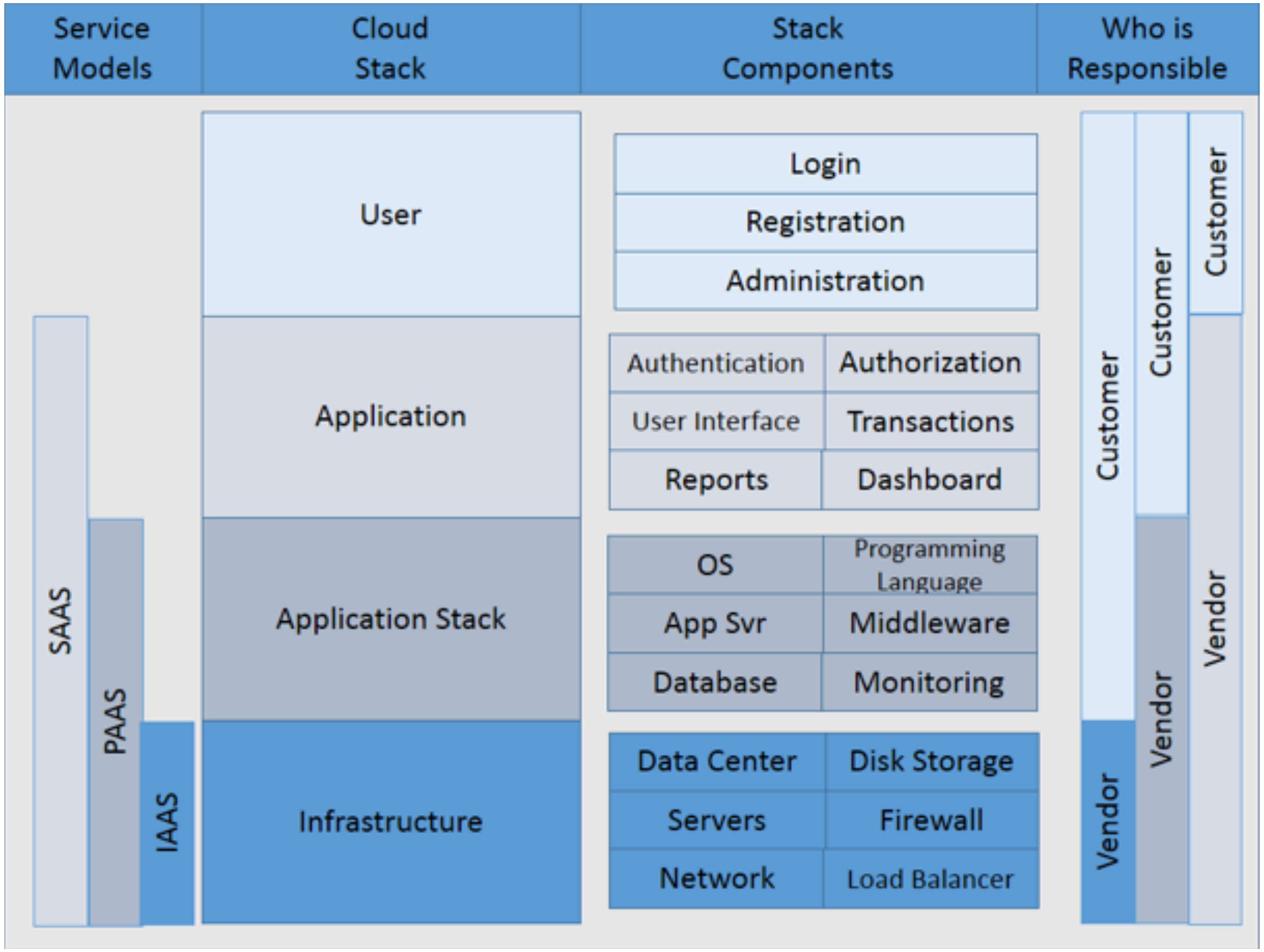
Including virtualized compute env on which analyses were run

With This Approach …

- Zero data need be downloaded to local compute
- Zero code need be installed or executed on local compute
- Instead, manuscript results can be regenerated:
- Directly from browser:
  - ✓ Readers merely clone the workspace (lightweight operation)
  - ✓ Then execute the relevant tasks as desired
  - ✓ Explore follow up hypotheses by customizing params
  - ✓ Even adding new codes or data as desired

Simplest and most complete solution to computational reproducibility yet available

In principle …



Service Models	Cloud Stack	Stack Components	Who is Responsible		
	User	Login Registration Administration	Customer	Customer	Customer
SAAS	Application	Authentication Authorization User Interface Transactions Reports Dashboard			Vendor
PAAS	Application Stack	OS Programming Language App Svr Middleware Database Monitoring		Vendor	
IAAS	Infrastructure	Data Center Disk Storage Servers Firewall Network Load Balancer	Vendor		

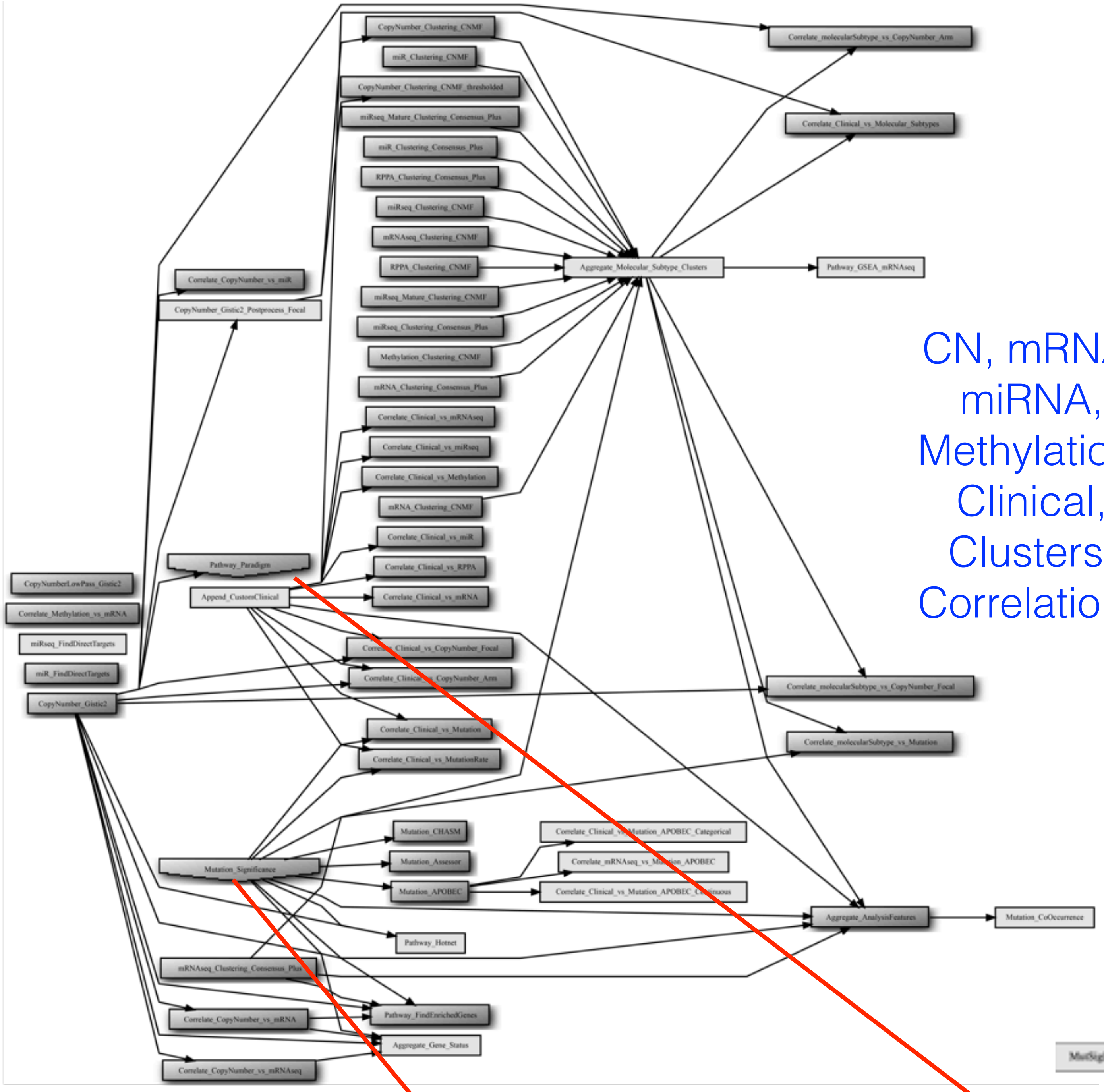
But cloud tech stack **very** thick

Complexity still seeps through

Esp for those who think bi-ology more than techn-ology

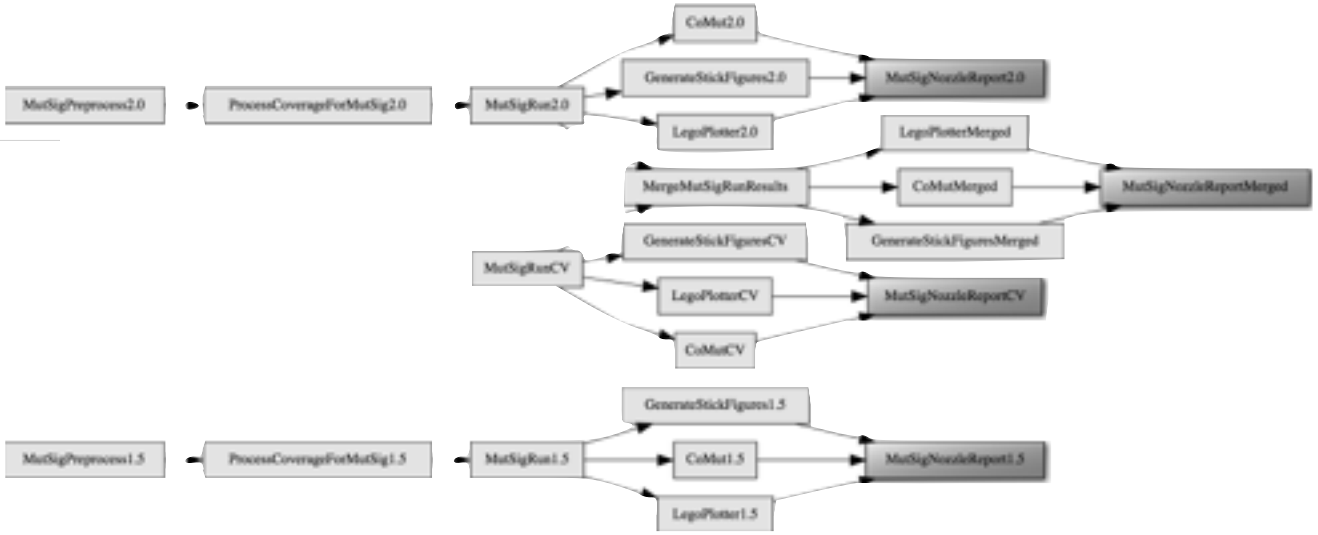
But like Rome, this was not built in a day

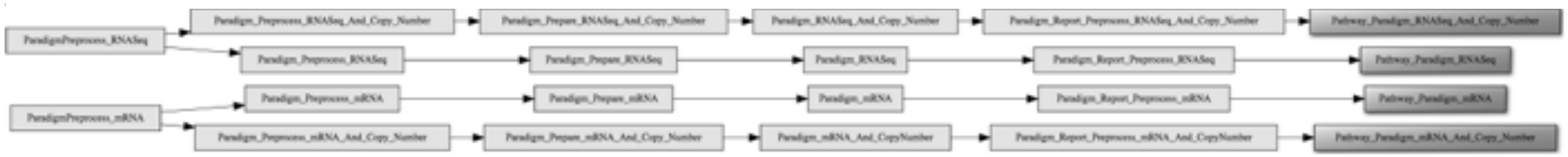
July 2016 Firehose Workflow


CopyNumber_Clustering_CNMF
miR_Clustering_CNMF
CopyNumber_Clustering_CNMF_thresholded
miRseq_Mature_Clustering_Consensus_Plus
miR_Clustering_Consensus_Plus
RPPA_Clustering_Consensus_Plus
miRseq_Clustering_CNMF
mRNAseq_Clustering_CNMF
RPPA_Clustering_CNMF
miRseq_Mature_Clustering_CNMF
miRseq_Clustering_Consensus_Plus
Methylation_Clustering_CNMF
mRNA_Clustering_Consensus_Plus
Correlate_Clinical_vs_mRNAseq
Correlate_Clinical_vs_miRseq
Correlate_Clinical_vs_Methylation
mRNA_Clustering_CNMF
Correlate_Clinical_vs_miR
Correlate_Clinical_vs_RPPA
Correlate_Clinical_vs_mRNA
Correlate_CopyNumber_vs_miR
CopyNumber_Gistic2_Postprocess_Focal
Pathway_Paradigm
Append_CustomClinical
CopyNumberLowPass_Gistic2
Correlate_Methylation_vs_mRNA
miRseq_FindDirectTargets
miR_FindDirectTargets
CopyNumber_Gistic2
Correlate_Clinical_vs_CopyNumber_Focal
Correlate_Clinical_vs_CopyNumber_Arm
Correlate_Clinical_vs_Mutation
Correlate_Clinical_vs_MutationRate
Mutation_Significance
Mutation_CHASM
Mutation_Assessor
Mutation_APOBEC
Pathway_Hotnet
mRNAseq_Clustering_Consensus_Plus
Correlate_CopyNumber_vs_mRNA
Pathway_FindEnrichedGenes
Aggregate_Gene_Status
Correlate_CopyNumber_vs_mRNAseq
Correlate_molecularSubtype_vs_CopyNumber_Arm
Correlate_Clinical_vs_Molecular_Subtypes
Aggregate_Molecular_Subtype_Clusters
Pathway_GSEA_mRNAseq
Correlate_molecularSubtype_vs_CopyNumber_Focal
Correlate_molecularSubtype_vs_Mutation
Correlate_Clinical_vs_Mutation_APOBEC_Categorical
Correlate_mRNAseq_vs_Mutation_APOBEC
Correlate_Clinical_vs_Mutation_APOBEC_Continuous
Aggregate_AnalysisFeatures
Mutation_CoOccurrence

CN, mRNA, miRNA, Methylation, Clinical, Clusters, Correlations

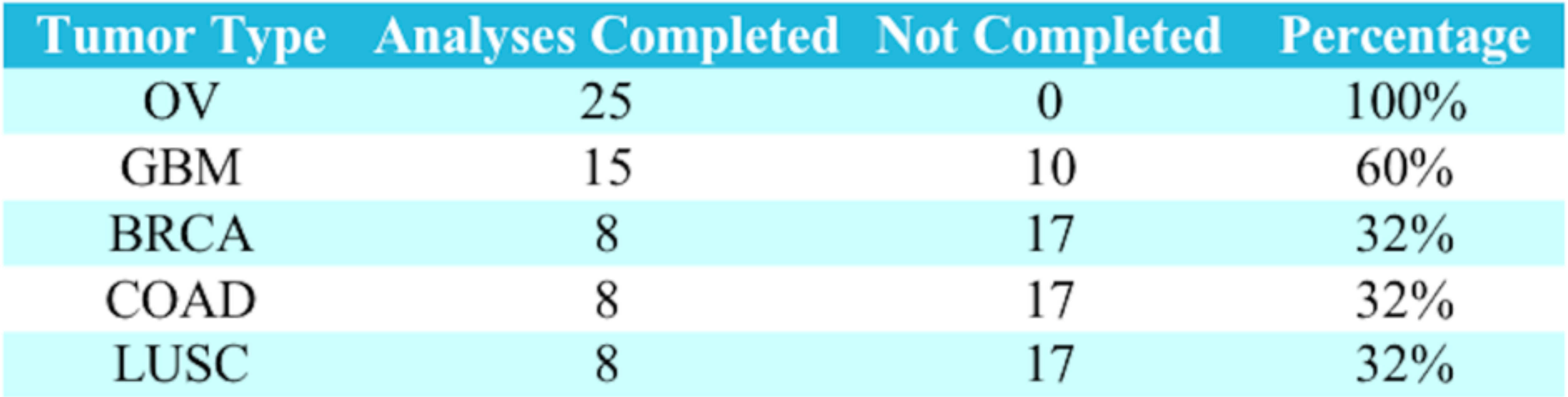
Mutation Significance




ParadigmPreprocess_RNASeq
Paradigm_Preprocess_RNASeq_And_Copy_Number
Paradigm_Prepare_RNASeq_And_Copy_Number
Paradigm_RNASeq_And_Copy_Number
Paradigm_Report_Preprocess_RNASeq_And_Copy_Number
Pathway_Paradigm_RNASeq_And_Copy_Number
Paradigm_Preprocess_RNASeq
Paradigm_Prepare_RNASeq
Paradigm_RNASeq
Paradigm_Report_Preprocess_RNASeq
Pathway_Paradigm_RNASeq
ParadigmPreprocess_mRNA
Paradigm_Preprocess_mRNA
Paradigm_Prepare_mRNA
Paradigm_mRNA
Paradigm_Report_Preprocess_mRNA
Pathway_Paradigm_mRNA
Paradigm_Preprocess_mRNA_And_Copy_Number
Paradigm_Prepare_mRNA_And_Copy_Number
Paradigm_mRNA_And_CopyNumber
Paradigm_Report_Preprocess_mRNA_And_CopyNumber
Pathway_Paradigm_mRNA_And_Copy_Number

Pathways

Broad GDAC Firehose Dec 2010 Run



Tumor Type	Analyses Completed	Not Completed	Percentage
OV	25	0	100%
GBM	15	10	60%
BRCA	8	17	32%
COAD	8	17	32%
LUSC	8	17	32%

Analysis Status for 5 Most Populous Tumor Cohorts
Few pipelines, run on few cohorts, mostly failures
~ 1/10th of ~84K data aliquots were accrued at this point

http://gdac.broadinstitute.org/runs/analyses__2010_12_23/

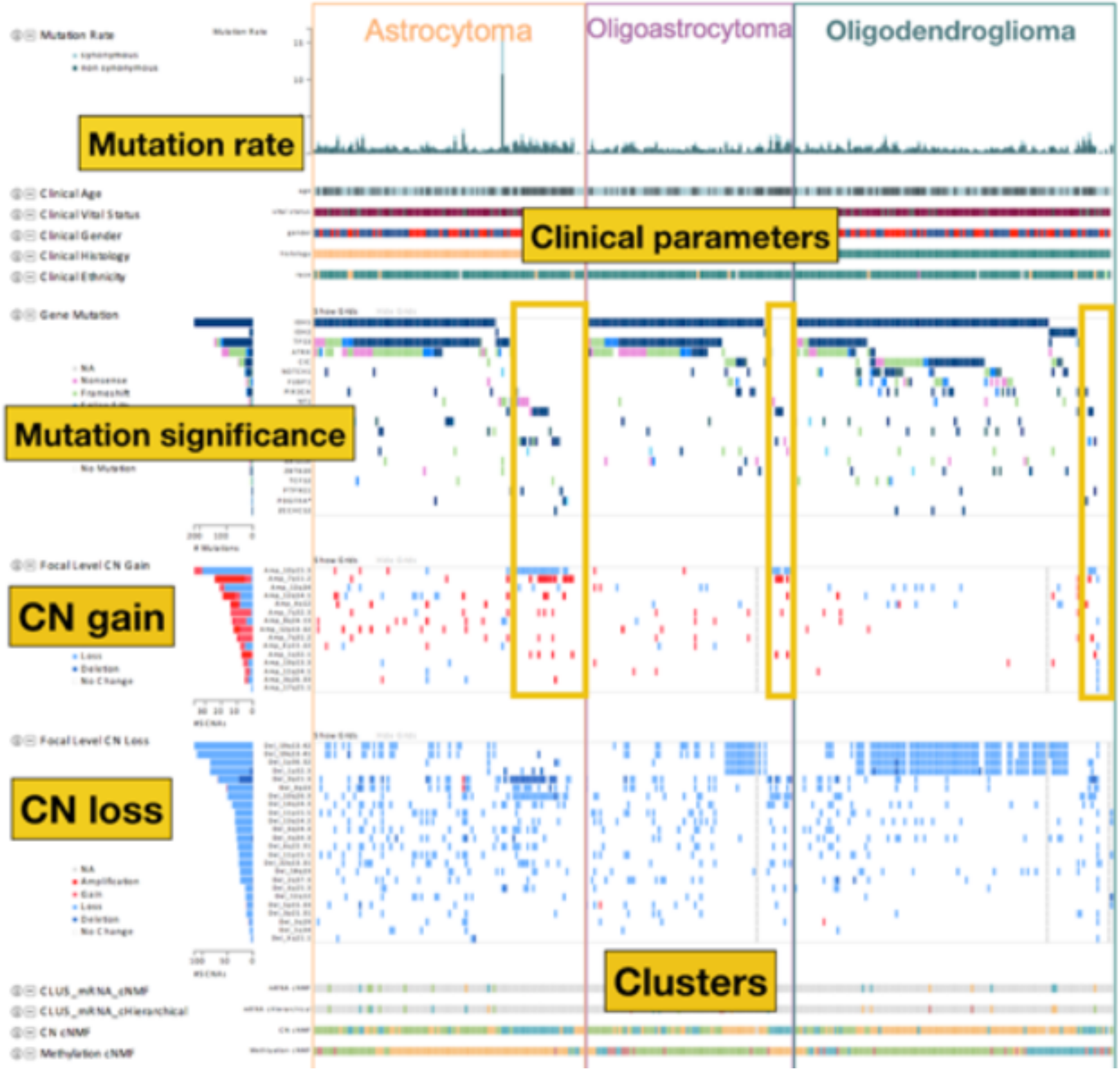
But Grew To: 62 GDAC runs/year, on live data stream

runs executed >1500 pipelines on ~84K aliquots

Compressing ~50TB of heterogeneous input to ~10GB results

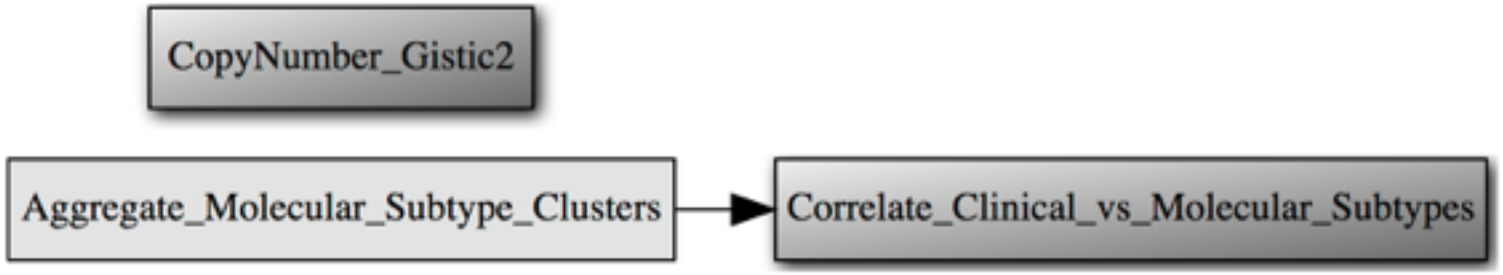
Entire multimodal analyses summarized in a single figure

Interact with and manipulate figures, directly from online publication

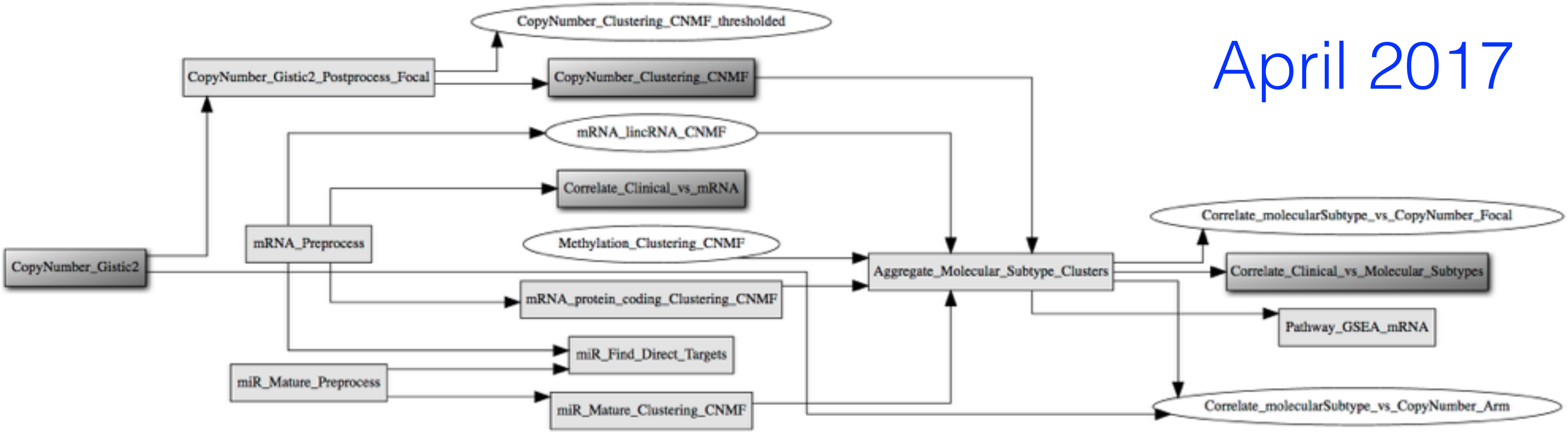


firebrowse.org/iCoMut/?cohort=LGG

Porting Genomic Analyses —> FireCloud


CopyNumber_Gistic2
Aggregate_Molecular_Subtype_Clusters
Correlate_Clinical_vs_Molecular_Subtypes

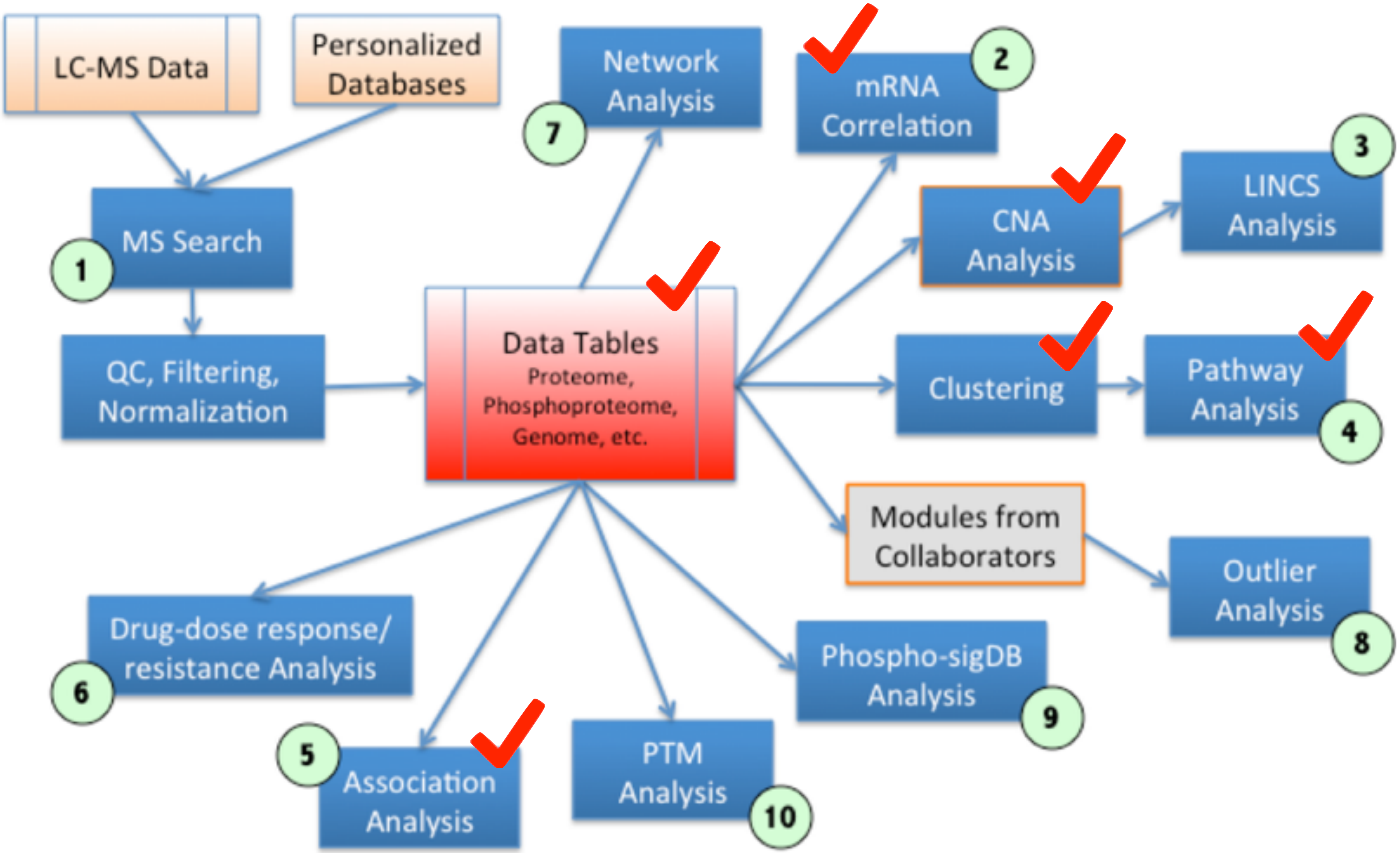
Jan 2017


CopyNumber_Clustering_CNMF_thresholded
CopyNumber_Gistic2_Postprocess_Focal
CopyNumber_Clustering_CNMF
mRNA_lincRNA_CNMF
Correlate_Clinical_vs_mRNA
Correlate_molecularSubtype_vs_CopyNumber_Focal
mRNA_Preprocess
Methylation_Clustering_CNMF
CopyNumber_Gistic2
Aggregate_Molecular_Subtype_Clusters
Correlate_Clinical_vs_Molecular_Subtypes
mRNA_protein_coding_Clustering_CNMF
Pathway_GSEA_mRNA
miR_Find_Direct_Targets
miR_Mature_Preprocess
miR_Mature_Clustering_CNMF
Correlate_molecularSubtype_vs_CopyNumber_Arm

April 2017

Data: all open-access HG38 from GDC

From Mani's v1 PGDAC Timeline



In the next 3-5 months, we expect components #1-6 to be available and functional.

Open-Source Ecosystem of Tooling has emerged

FISS: load & control FC via API & simple Python/UNIX CLI

Firecloud dev toolkit : more easily create Google VMs, interact with Google storage, dockerize codes, etc

GDCtools

- Simplify search & retrieval of data from GDC
- Directly from UNIX command line / Python
- Automate tasks common to data-driven science
- Mirror *legacy* or *harmonized* data
- Sample reports

https://github.com/broadinstitute/fiss

https://github.com/broadinstitute/firecloud_developer_toolkit

https://github.com/broadinstitute/gdctools

These are gradually streamlining use of FireCloud, e.g.

When CPTAC genomic data at GDC, we will have it at Broad within hours (maybe even in FireCloud)

But as we heard from Nathan Edwards yesterday:

Develop locally, iteratively experiment
Then deploy to production on cloud
For running at scale
Connect to other tools & data in ecosystem
Collaborate globally, publish

Proteomic Data & Analyses: More Needed

P-CDAP further along: how to leverage?
Better Aspera auto-download from DCC
More experiments with MSGF docker (PNNL)
BioContainers: leverage and contribute

FireCloud Pipeline Differentiators

Seamlessly combine proteogenomic data & analyses in one platform
HUGE corpus of TCGA protected data already in FC, no copying
Automated connection to GDC … not sure P-CDAP?
API-driven analyses & data querying (e.g FireBrowse, GDCtools)
Persistent Sandbox (workspace): DOI attached to publication,
Push-button reproducibility Exploratory visualization

But as Marcin Cieslik noted: schedule realities strongly influence the chosen path

Registering for FireCloud is free. This gives read-only access to public spaces.

*But you need a FireCloud Billing Project to **create** a new workspace.*

Two ways to gain access to a FireCloud Billing Project:

1. **Existing billing project** OWNERS can authorize you to an existing **FireCloud Billing Project.**

2. You create your own **FireCloud Billing Project,** by first setting up a Google Billing Account.

FireCloud team @ Broad happy to set up workshop for CPTAC

Will reach out when construction noise quiets a little more

Acknowledgements

PI: **Gad Getz,** D. R. Mani

Broad Institute

Daniel DiCara
David Heiman
Harindra Arachchi
Hailei Zhang
Juok Cho
Jaegil Kim
Gordon Saksena
Douglas Voet
William Mallard
Michael Lawrence
Petar Stojanov
Lihua Zou
Chip Stewart
Scott Frazer
Pei Lin
Kristian Cibulskis
Lee Lichtenstein
Aaron McKenna
Andrey Sivachenko
Carrie Sougnez
Lee Lichtenstein
Steven Schumacher
Raktim Sinha

Belfer/DFCI/MDACC

Juinhua Zhang
Spring Liu
Sachet Shukla
Terrence Wu

IGV & GenePattern teams @ Broad

Jill Mesirov
Michael Reich
Peter Carr
Marc-Danie Nazaire
Jim Robinson
Helga Thorvaldsdottir

Broad Institute Leadership: Todd Golub, Eric Lander

Harvard Medical School

Matthew Meyerson
Andrew Cherniack
Juliann Chmielecki
Rameen Beroukhim
Scott Carter

Peter Park
Nils Gehlenborg
Semin Lee
Richard Park

In Particular

David Heiman
Kane Hadley
Sam Meier
Karsten Krug
Hailei Zhang
Jaegil Kim
Tim DeFreitas
D. R. Mani

The front line computational biologists
and software engineers.

Fin