Update on Genomic Analyses for CPTAC

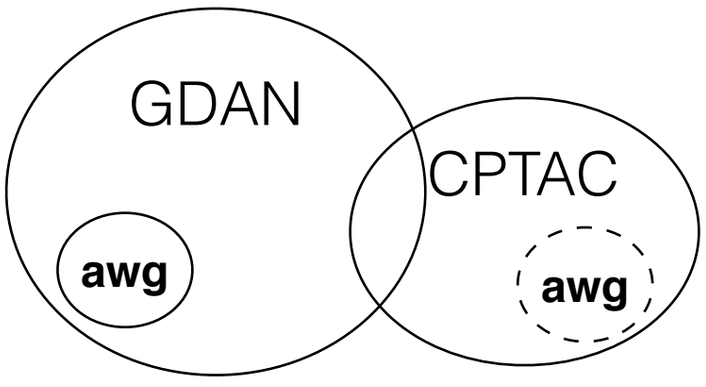
Michael S. Noble
Broad Institute of MIT & Harvard

CPTAC3 NIH Leadership Site Visit to Broad
Cambridge, Massachusetts
2017_05_31

First, since I'm still taking stock of CPTAC, here's a perspective on how I see our role and its potential impact.

Global Proteo-Genomic Cancer Research Community

M.D.s, molecular biologists, professors, computational researchers, SW engineers, clinicians, lab techs, students, pharma, education & outreach



Our consortia are small-ish factions of global community

We serve that community & truth itself

So “it works for me” (on my computer at my institute) is not enough

The lessons we’ve learned have to scale up & be easily undertood & easily accessible to the community

This has potentially enormous force-multiplying global effect

Transition to Firecloud will ultimately help global community more than us

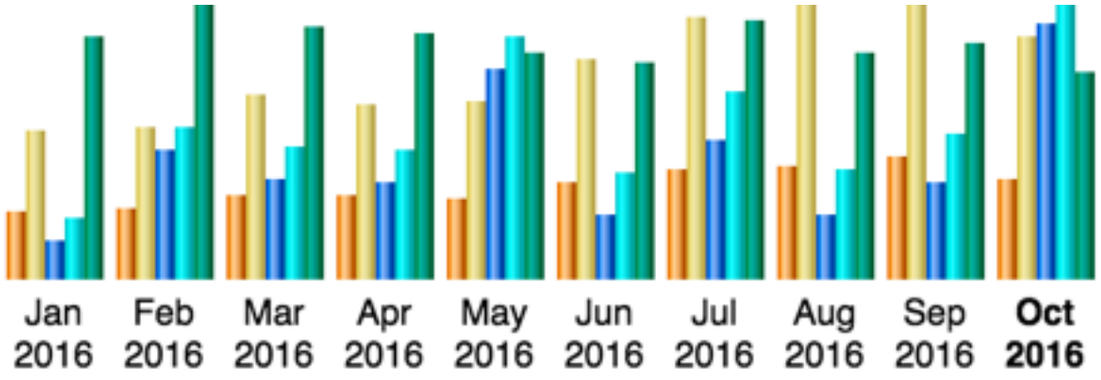
We believe this because …

From *2014 -> Oct 2016* GDAC accessed >12 million times (downloads, pageviews, API calls, etc)

Generating over a PB of traffic
Embedded into multiple external portals around world

Pages	Bandwidth
124,563	3336.66 GB
406,324	2904.64 GB
315,609	4546.87 GB
305,579	2001.54 GB
669,790	24775.02 GB
203,205	169673.87 GB
442,463	535246.48 GB
199,819	24332.77 GB
301,701	56553.57 GB
822,813	23328.92 GB
0	0
0	0
3,791,866	846700.34 GB

2016 traffic: Jan thru Oct



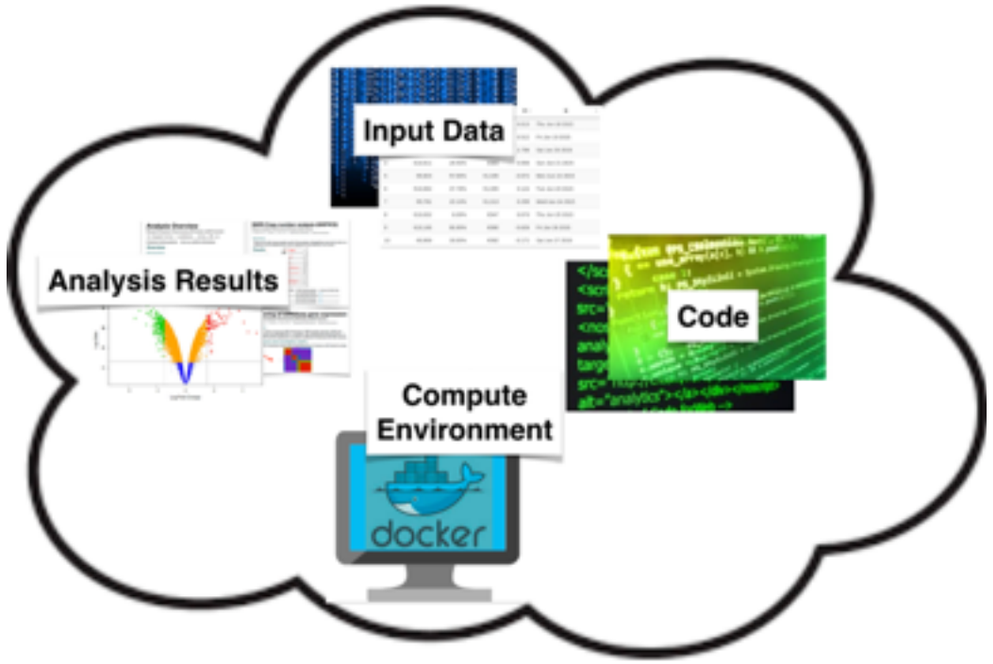
Automated, easily repeatable, production-grade science generates much global interest

Why belabor?
To edify some + maintain troop morale

- **On-prem to cloud is big, long transformation**
- For ~20 years, `ssh` & `bsub` were largely sufficient
  for many to conduct science …
- But with cloud on-prem IT support is waning
  and some science groups ahead of IT on cloud
- So now we each have to become our own
  - *mini-IT department*
  - *cloud VM deployment experts*
  - *and finance admins (e.g. for cloud billing)*
- As noted in April:
  - *pipelines still harder to deploy & debug*
  - *paucity of Institute-wide knowledge & support*

Tactically, we're about 75% through this transition, but it should be well worth the cost once scaled to community

Example: solve reproducibility once & for all



Manuscript analysis workspace

Freeze analysis workspace

Make public @ paper acceptance

Using DOIs as in Firehose runs

Paper cites URL to workspace

Contains <u>everything</u> in paper

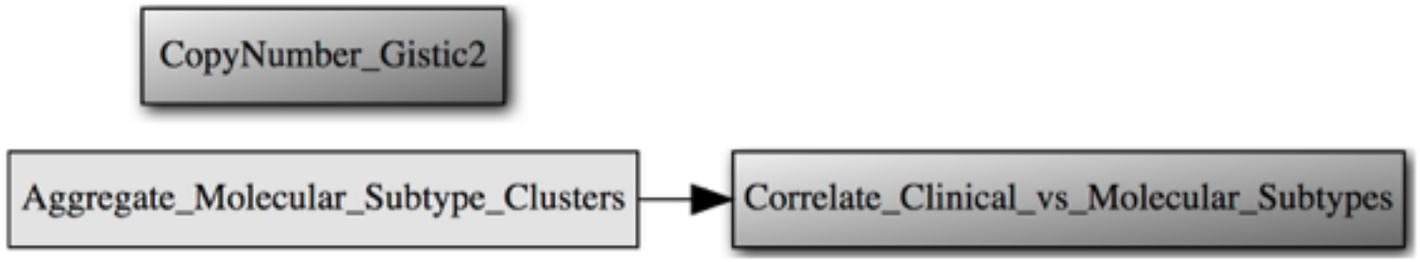
Including virtualized compute env on which analyses were run

- 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

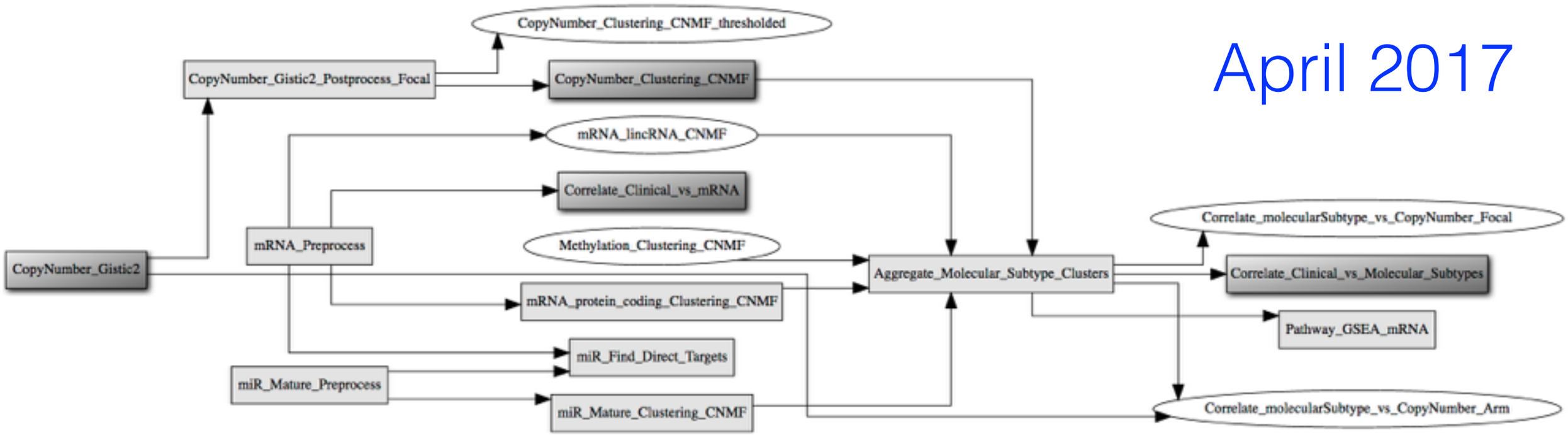
Simplest and most complete solution to computational reproducibility yet available

Recent Progress

Porting Genomic Analyses —> FireCloud



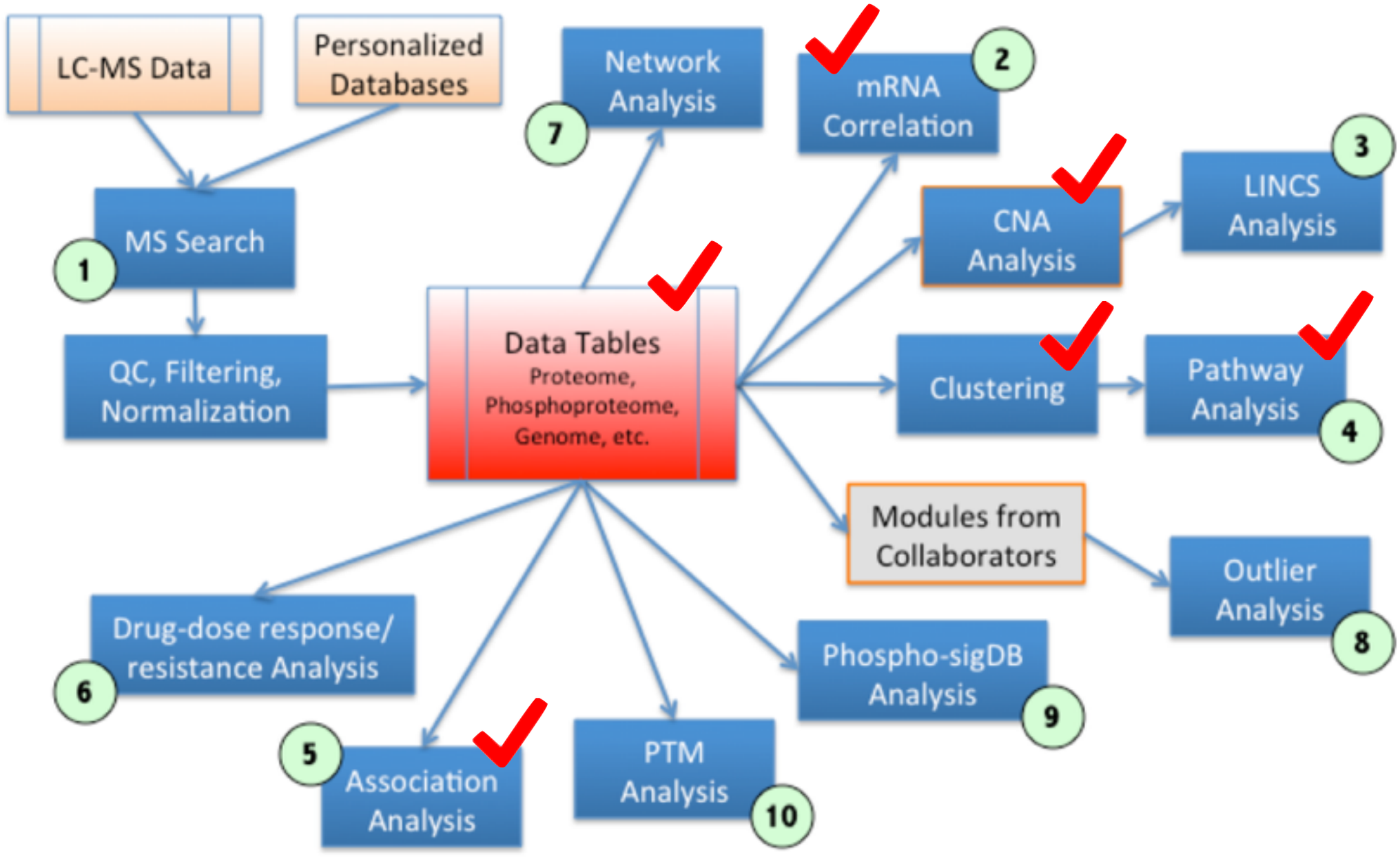
Jan 2017



April 2017

Data: open-access HG38 from GDC (harmonized TCGA)
Pipelines for mRNA, miR, CN, pathways, associations
Roughly 50% coverage of legacy TCGA HG19 pipelines

From Mani's v1 PGDAC document



Red ✓ indicates components covered by the genomic pipeline infrastructure

Genomic Pipeline Updates

Good fraction of effort towards: WGS and HG19—>HG38 migration
Best case: just simple data QC (e.g. miR expression data)
Medium case: algorithm rewiring for new data (mRNA expression)
Worst case: substantial algorithmic vetting & modifications (mutation)

Mutation (SNV) data:

- Active focus of our HG38 QC awg in GDAN (Chaired at Broad)
- Helped instigate GDC data release v6 (May 9)
- Which improved HG38 MAFs to re-capture more HG19 variants
  Using MC3 variant filtering strategies
  And re-processed hundreds of Mutect calls
- GDC data release v7 slated for summer: more MAF
  improvements (OxoG, Panel-of-Normals filters)
- Addressing issues such as in this broken OxoG paper

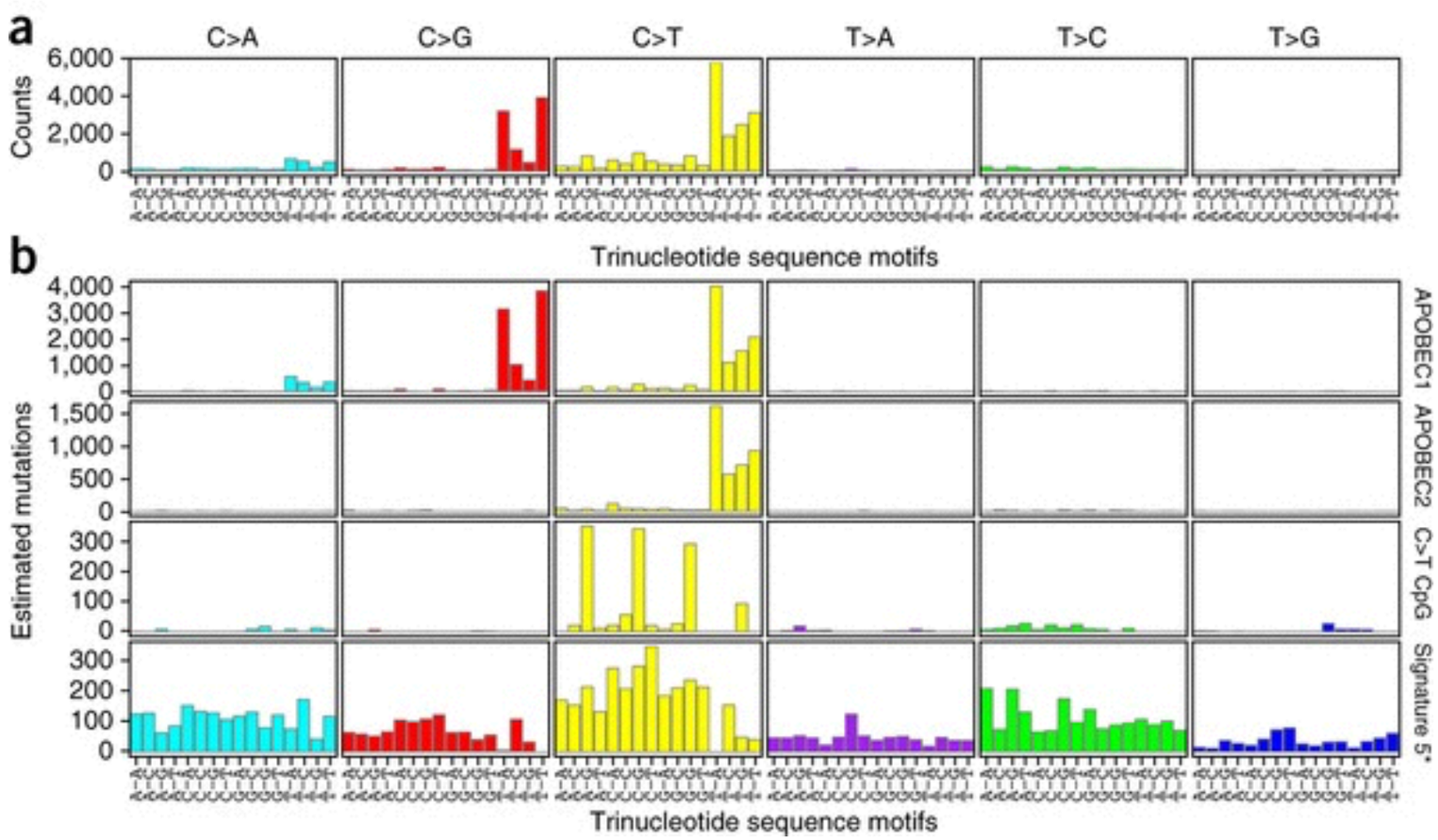
Mutation (SNV) data continued:

- Statistical significance analysis (*MutSig*) stalled at ~80% done
- *P-MACD tool*: Gordenin et al (NIH)
  - Analysis of mutagenesis by APOBEC cytidine deaminases
  - http://www.nature.com/ng/journal/v45/n9/full/ng.2702.html
  - Verified for HG38 (but not in Firecloud yet)
- *SignatureAnalyzer tool*: Jaegil Kim et al (Broad)
  - Verified for HG38
  - Now being installed to FireCloud

Mutational signatures are patterns of base changes associated with specific mutational processes in tumor cells. Our tool uses non-negative matrix factorization (NMF) to discover & characterize signatures across multiple cancers.

Here we show spectrum of base changes identified in the 130 sample TCGA BLCA (bladder cancer) cohort, displayed for mutated pyrimidines and adjacent 3′ and 5′ bases.

4 signatures detected



https://www.nature.com/ng/journal/v48/n6/full/ng.3557.html

Copy Number data:

- New GISTIC HG38 markers file (to adjust for SNP6 liftover)
- Installed to FireCloud
- New test analyses runs on **exome** data conducted
- **Whole genome** GISTIC (marker-less) in on-prem testing

Methylation data:

- Now supported by GDCtools (more on that later)
- Preprocessing pipeline installed to FireCloud
- CNMF clustering pipeline : nearly finished

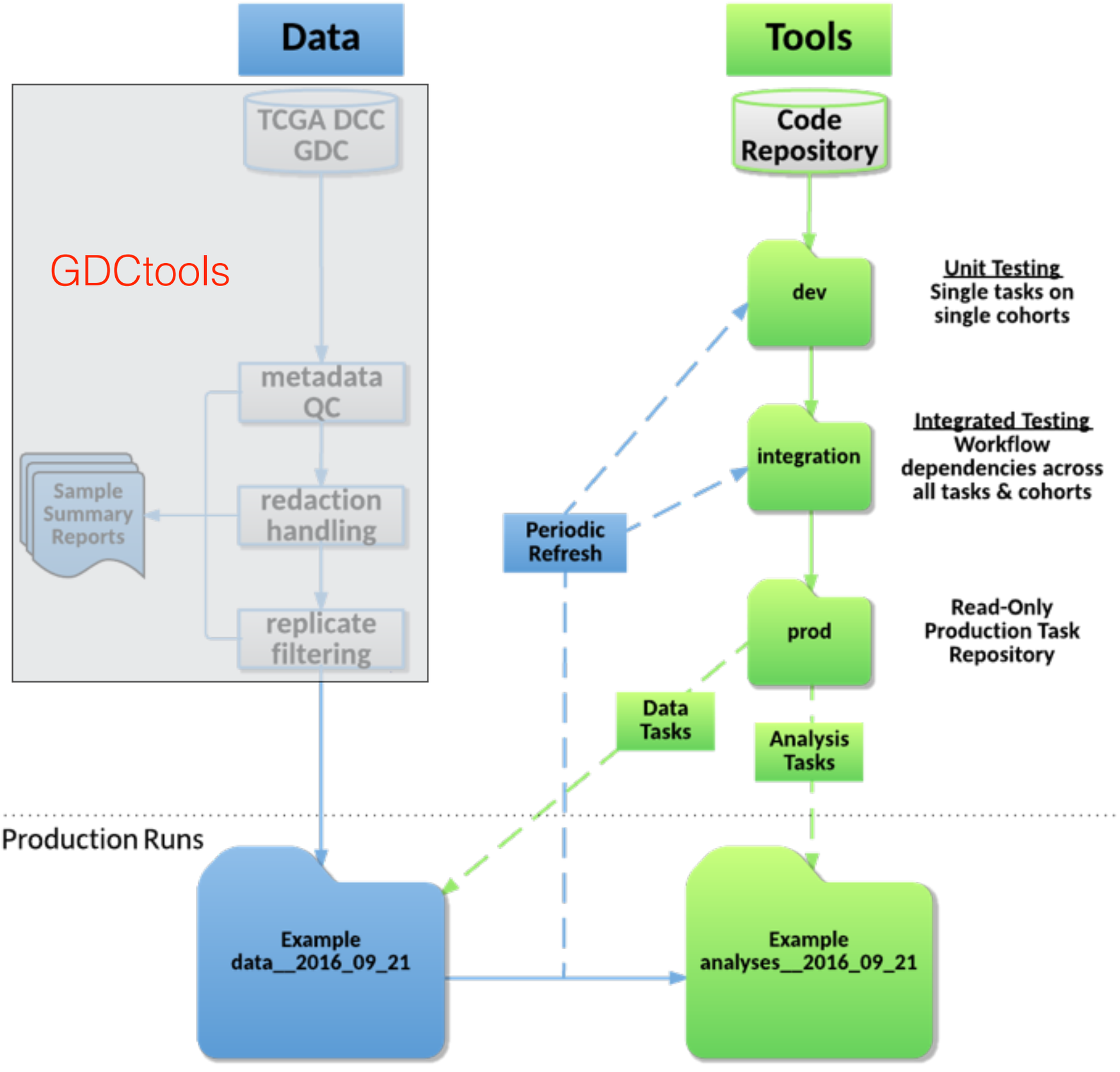
Clinical data:

- loaded to FireCloud, but associations need major refactor
- “technical debt” R code from legacy TCGA

Online results from Nov HG38 run:

http://gdac.broadinstitute.org/runs/analyses__2016_11_03/reports/

Infrastructure Updates



Infrastructure is defined as tooling which enables end-to-end genomic pipeline operation for GDAN & CPTAC (similar to TCGA)

Includes

FireCloud
GDCtools
FISS
GDAC scripts
FireBrowse
iCoMut

GDCtools

- Enable one to quickly use & program against GDC
- Via open-source, config-file driven Python & UNIX utilities
- Begin in just minutes, no need to hire/train staff

```
linux%  git clone https://github.com/broadinstitute/gdctools
linux%  make test
```

- Or learn virtually any of the GDC portal or API

By wrapping the GDC RESTful API (which was written by programmers, for programmers) in a set of **domain-aware tools**, GDCtools **lets more users interact more quickly with the GDC**, in memes familiar to them—as biomedical researchers & informaticians—rather than as web or database programmers.

External/community contributions to GIThub repo
Constantly evolving: 40 commits since April CPTAC

- It's well understood that MUCH EFFORT in data-driven science is
- Aggregation, cleansing, sample counting & tracking, reports
- Esp for consortium-scale datasets
- Such as TCGA: 33 cohorts, 11.5K patients, 85K data aliquots
- Firehose performed this democratizing service in TCGA

  In ~5K lines of Python, BUT *internal/monolithic (not open)*

GDCtools aims to generalize this, to all data at GDC

And make it open-source for everyone

Largely replaced by GDCtools + 4 lines BASH

```
gdc_mirror      -config tcga.cfg
gdc_dice        -config tcga.cfg
gdc_loadfile    -config tcga.cfg
gdc_report      -config tcga.cfg
```

This is essentially our nightly cron job

- Easily download & process ***all*** or ***subset(s)***
- Highly configurable: even to just 1 case
- Example: to mirror all of TARGET

```
gdc_mirror     –config target.cfg
```

- Example: create pan-gastro-intestinal cohort

```
[aggregates]
TCGA-PANGI: TCGA-COAD,TCGA-READ,TCGA-STAD,TCGA-ESCA
```

✓ Zero coding, just 1 line in config file

✓ Then run `gdc_loadfile` tool for 2 mins

✓ Then load & analyze en-masse in FireCloud

- Example: to put in Google cloud buckets

```
gdc_loadfile  --config tcga.cfg,google.cfg
```

- This is how we specify loads to FireCloud

```
[loadfiles]
DIR: %(ROOT_DIR)s/loadfiles/google
FILE_PREFIX: gs://broad-institute-gdac/gdc/dice
FORMAT: firecloud
```

Entire content of `google.cfg`
Simply replaces `[loadfiles]` directive from `tcga.cfg`

Takeaway: when CPTAC genomic data exposed at GDC, we will have it at Broad within hours

FISS : The (Fi)reCloud (S)ervice (S)elector

Programmatic interface to FireCloud (FC), providing a set of low- and high-level Python bindings to the FireCloud API, as well as UNIX CLI bindings.

This captures a large majority of use cases, and is *derived from legacy FISS developed to operate GDAC FireHose pipeline in TCGA: so feature set is clear.*

Like GDCtools, aims to translate RESTful web API into services that resonate more closely with majority of expected FC users--in memes familiar to them as biomed researchers & informaticians rather than database or web programmers.

Used in FC pipeline operations for GDAN, GTEX, CPTAC
And interactive, exploratory work (e.g. online Python notebooks)
3 releases (including public) and 62 commits since April CPTAC

FISS is part of the “secret” of how we did as many as 62 runs per year in TCGA, with as many as 1500 analyses on 85K samples

https://github.com/broadinstitute/fiss

Example: start/stop/edit/query workflows

```
linux%  fiss -l flow_
    flow_copy
    flow_diff
    flow_edit
    flow_exists
    flow_export
    flow_import
    flow_list
    flow_show
    flow_start
    flow_status
    flow_stop
```

Legacy Firehose

```
linux%  fissfc -l flow_
    flow_acl
    flow_delete
    flow_list
    flow_new
    flow_set_acl
    flow_start
```

Firecloud

Demonstrates the aforementioned lag in feature set for cloud

Kick off an analysis run

```
linux%  fiss -F analyses_start

analyses_start ()
{
    Usage 2 "<space> <sample_set>
    Initiate the GDAC Analyses workflow, for given <sample_set> in <space>";
    Priority=Deadline;
    flow_start $1 $2 Analyses
}
```

Monitor the run

```
linux%  fiss -F analyses_status

analyses_status ()
{
    Usage 2 "<space> <sample_set>
    Get status of GDAC Analyses workflow, for given <sample_set> in <space>";
    flow_status -tsv $1 $2 Analyses
}
```

Other infrastructure work: packaging

Before public release our data & analysis results are packaged by pipelines into archives

For downloading en-masse

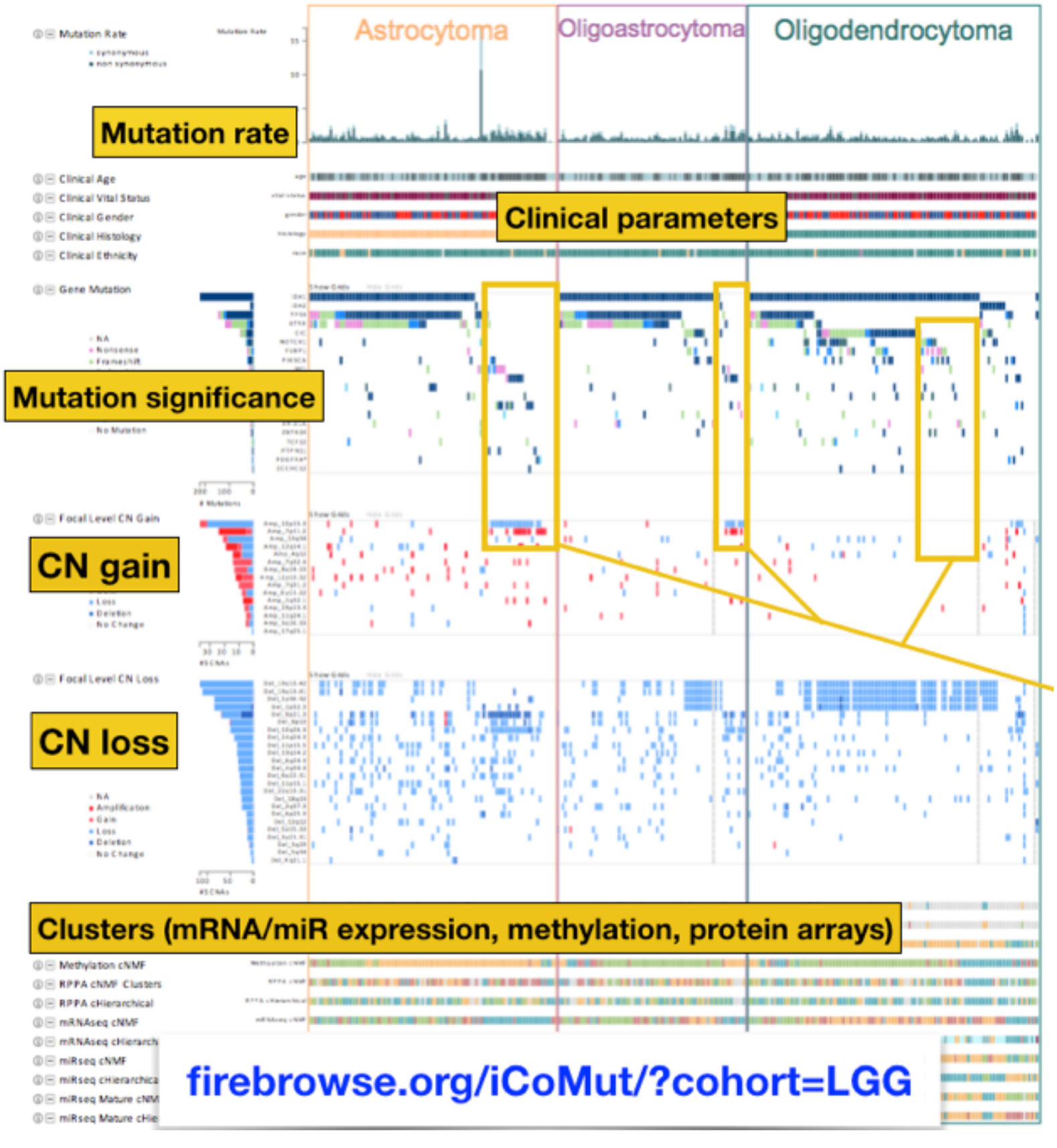
```
firehose_get : retrieve open-access results of Broad Institute TCGA GDAC runs
Version: 0.3.3 (Author: Michael S. Noble)

Usage: firehose_get [flags]  RunType  Date  [tumor_type, ... ]
```

firehose_get

```
BLCA BRCA CESC COADREAD DLBC GBM HNSC KIRC KIRP LAML LGG LIHC
LNNH LUAD LUSC OV PAAD PRAD SKCM STAD THCA UCEC PANCANCER
```

- Download all or parts
- Of any run since 2012
- **Open & password access**
- Select by run type & date
- Subselect by disease type
- Or analysis type:
- See what runs we did
- Or what analyses in each

Or online inspection at GDAC portal & FireBrowse



Packaging can be more difficult in cloud (non-local storage, non-shared VMs)

But we have devised a way of automatically adding to FireCloud tasks & extracting (currently in test)

How to use FireCloud: free to register, read access to public spaces
Creating a workspace or running code requires billing project

Initial Experimentation Also Free: Courtesy of NCI

Link: Request Free FireCloud Credits

Tier1:	$300	
Tier2:	$1000	(after Tier1 used)
Tier3:	up to $10K	(after Tier2 used)

First-come, first-served. Expire in September 2017

FireCloud team @ Broad happy to set up workshop for CPTAC

Acknowledgements

David Heiman
Kane Hadley
Sam Meier
Karsten Krug
Gordon Saksena
Hailei Zhang
Jaegil Kim
Tim DeFreitas
Julian Hess
Vicky Horst

The front line comp bios & software engineers

PI: Gad Getz, D. R. Mani

Fin