Final Report:

GDAC Firehose Integration With The Genomic Data Commons

M. Noble, T. DeFreitas, D. Heiman
Broad Institute of MIT & Harvard

mnoble@broadinstitute.org
2016_11_18

Outline

- Review S.O.W.
- Firehose: then and now
- Explore latest data & analysis runs
- Back to the Future

For review of Firehose and July 2016 report on GDC integration, see previous talk here.

Statement of Work

1. Develop instance of Firehose that obtains data from GDC via API

2. Prototype a Firehose pipeline capable of ingesting the current version of GDC data for displaying in FireBrowse and test the prototype

3. Develop a plan to make the Firehose pipeline production ready and identify recommended enhancements

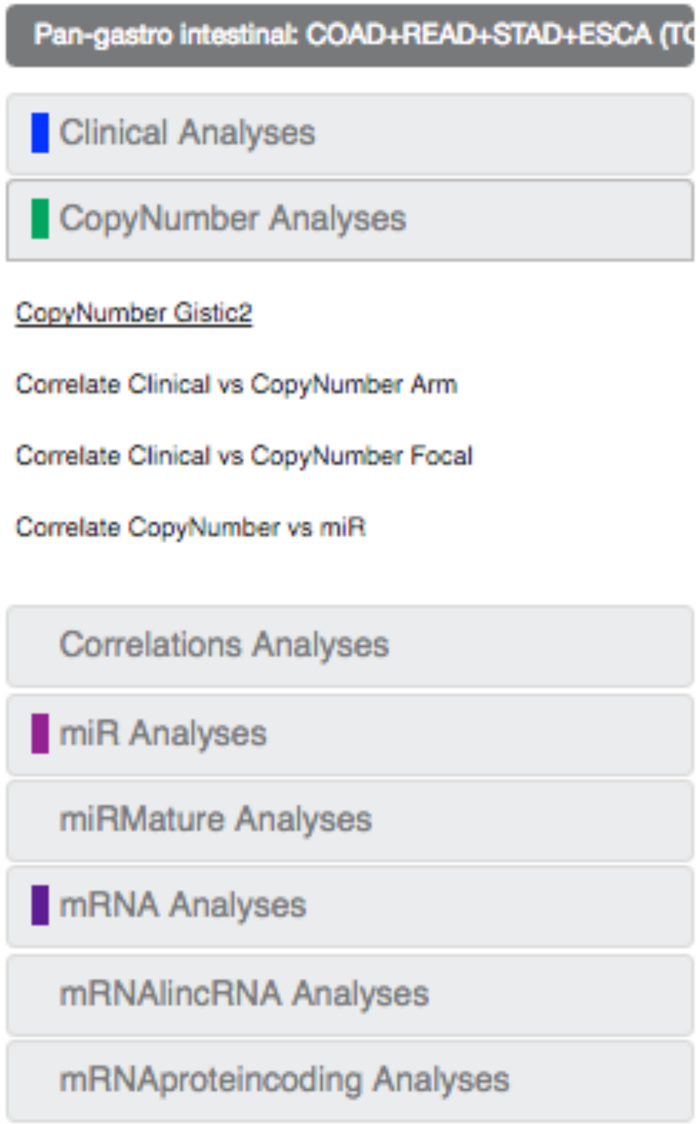
Statement of Work

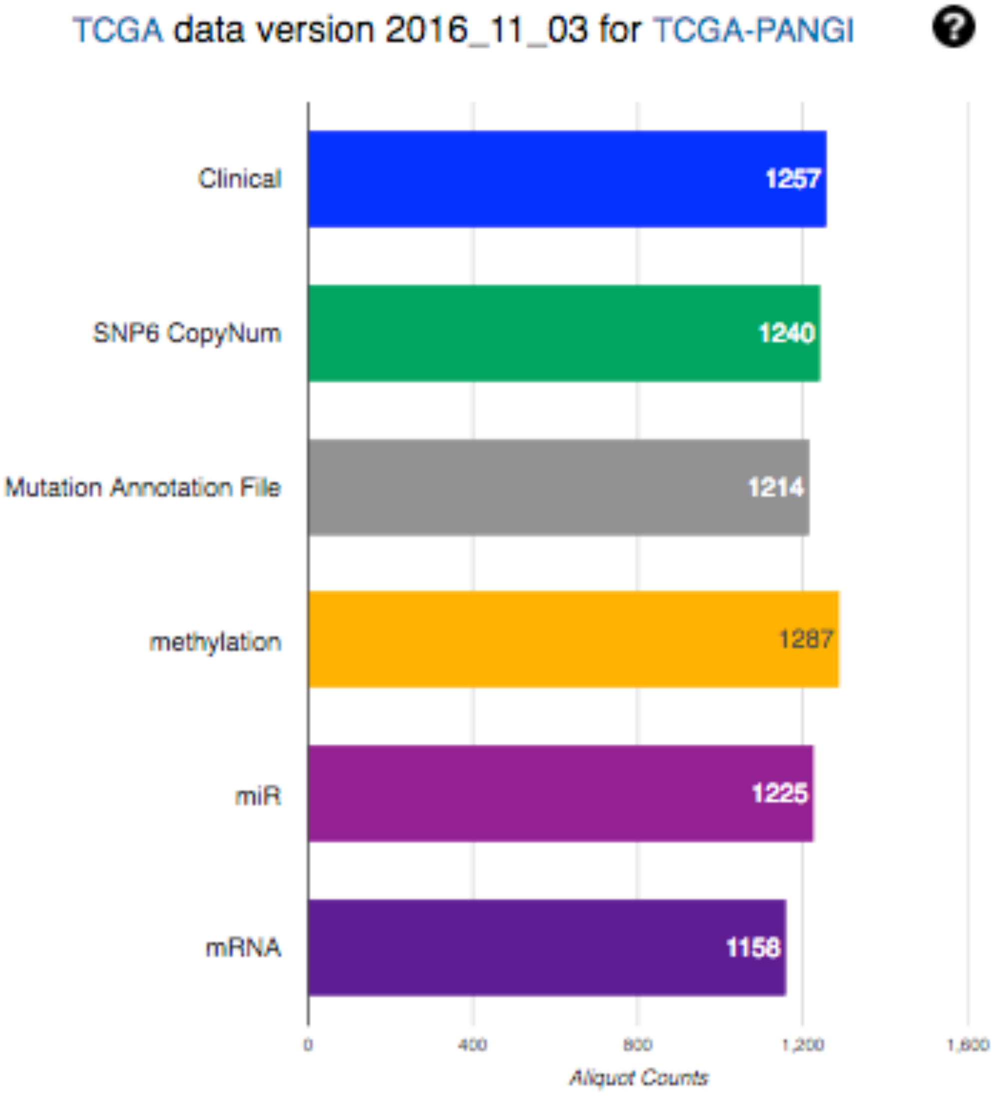
1. Develop instance of Firehose that obtains data from GDC via API
2. Prototype a Firehose pipeline capable of ingesting the current version of GDC data for displaying in FireBrowse and test the prototype
3. Develop a plan to make the Firehose pipeline production ready and identify recommended enhancements

Towards **GDAN readiness**, we went well beyond this

Substantial vetting/QC of GDC data and API
Spinning out an open-source **`GDCtools`** repo
Piecewise automated data load into FireCloud
HG38 *stddata workflow* running in FireCloud
HG38 data exposed for download: **`firehose_get`**

Even a new cohort*





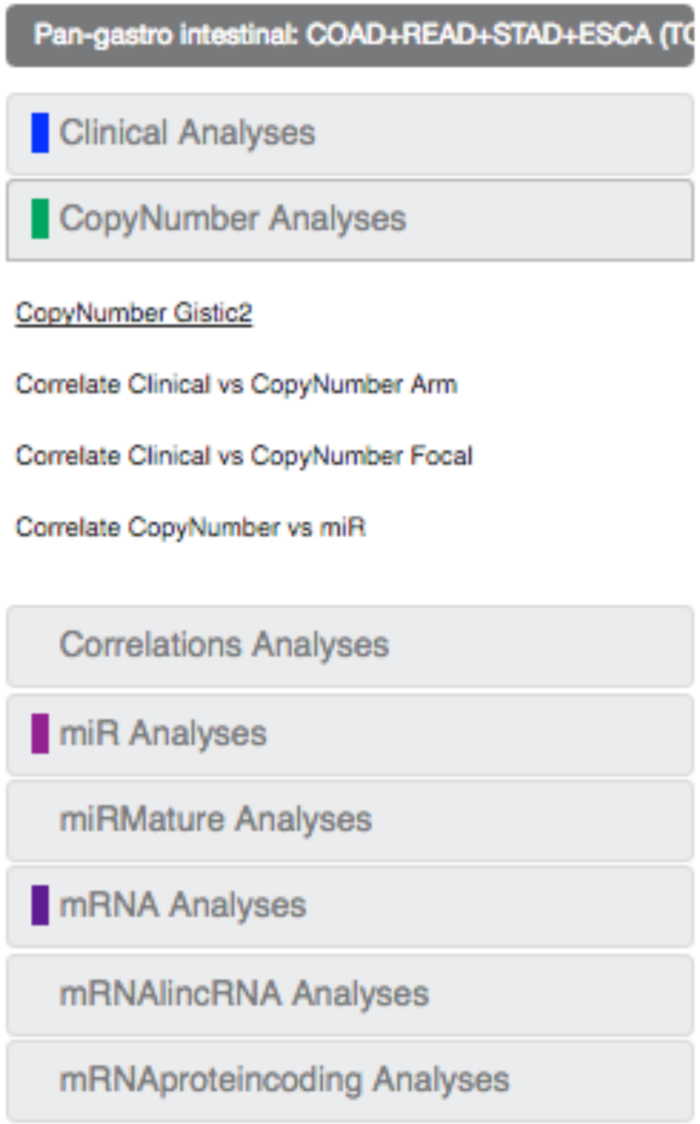
PANGI in gdcbeta FireBrowse

COAD + READ + STAD + ESCA

* Not available in public FireBrowse

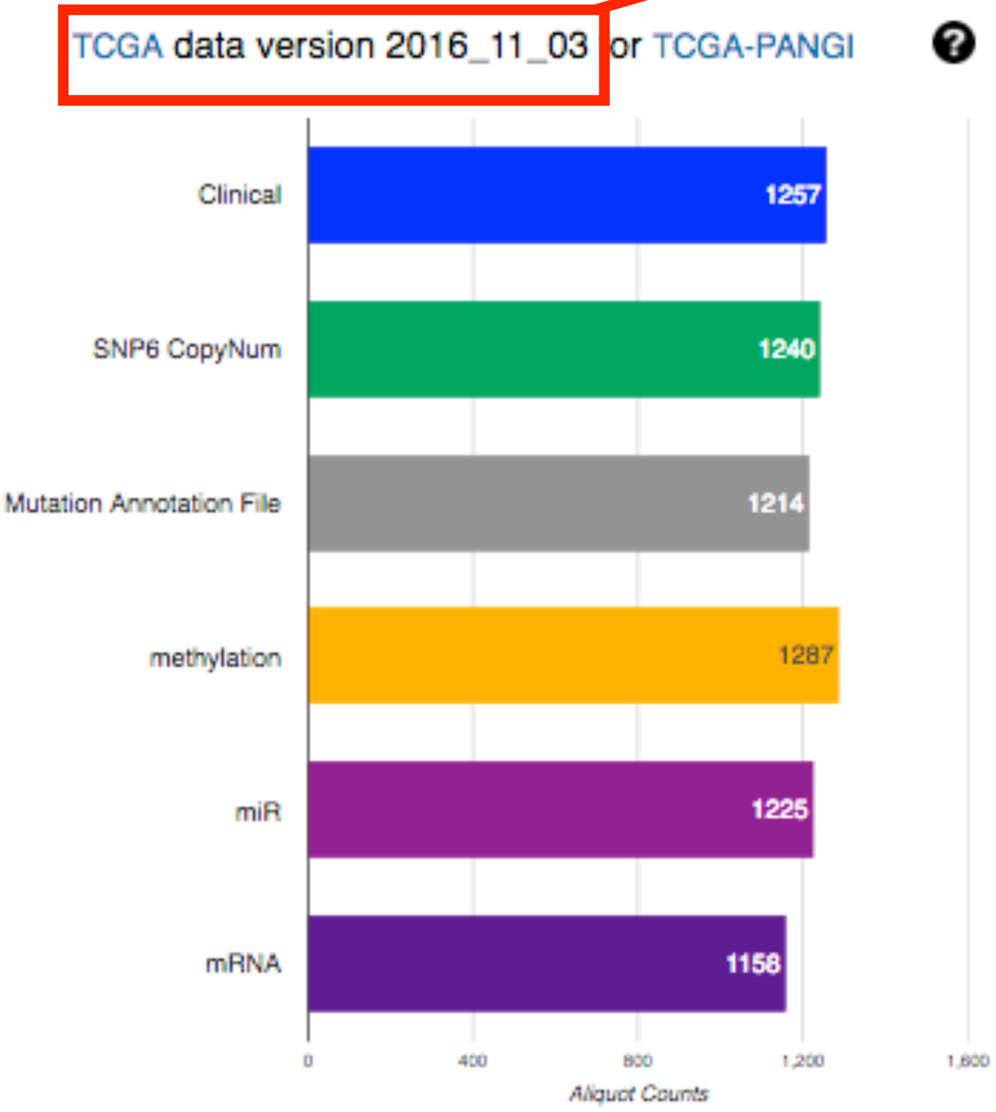
Even a new cohort*

Snapshot of GDC HG38 data



Pan-gastro intestinal: COAD+READ+STAD+ESCA (TC

- 🟦 Clinical Analyses
- 🟩 CopyNumber Analyses
  - CopyNumber Gistic2
  - Correlate Clinical vs CopyNumber Arm
  - Correlate Clinical vs CopyNumber Focal
  - Correlate CopyNumber vs miR
- Correlations Analyses
- 🟪 miR Analyses
- miRMature Analyses
- 🟪 mRNA Analyses
- mRNAlincRNA Analyses
- mRNAproteincoding Analyses



TCGA data version 2016_11_03 for TCGA-PANGI

PANGI in gdcbeta FireBrowse

COAD + READ + STAD + ESCA

* Not available in public FireBrowse

Brief Review

FIREHOSE
TCGA GDAC

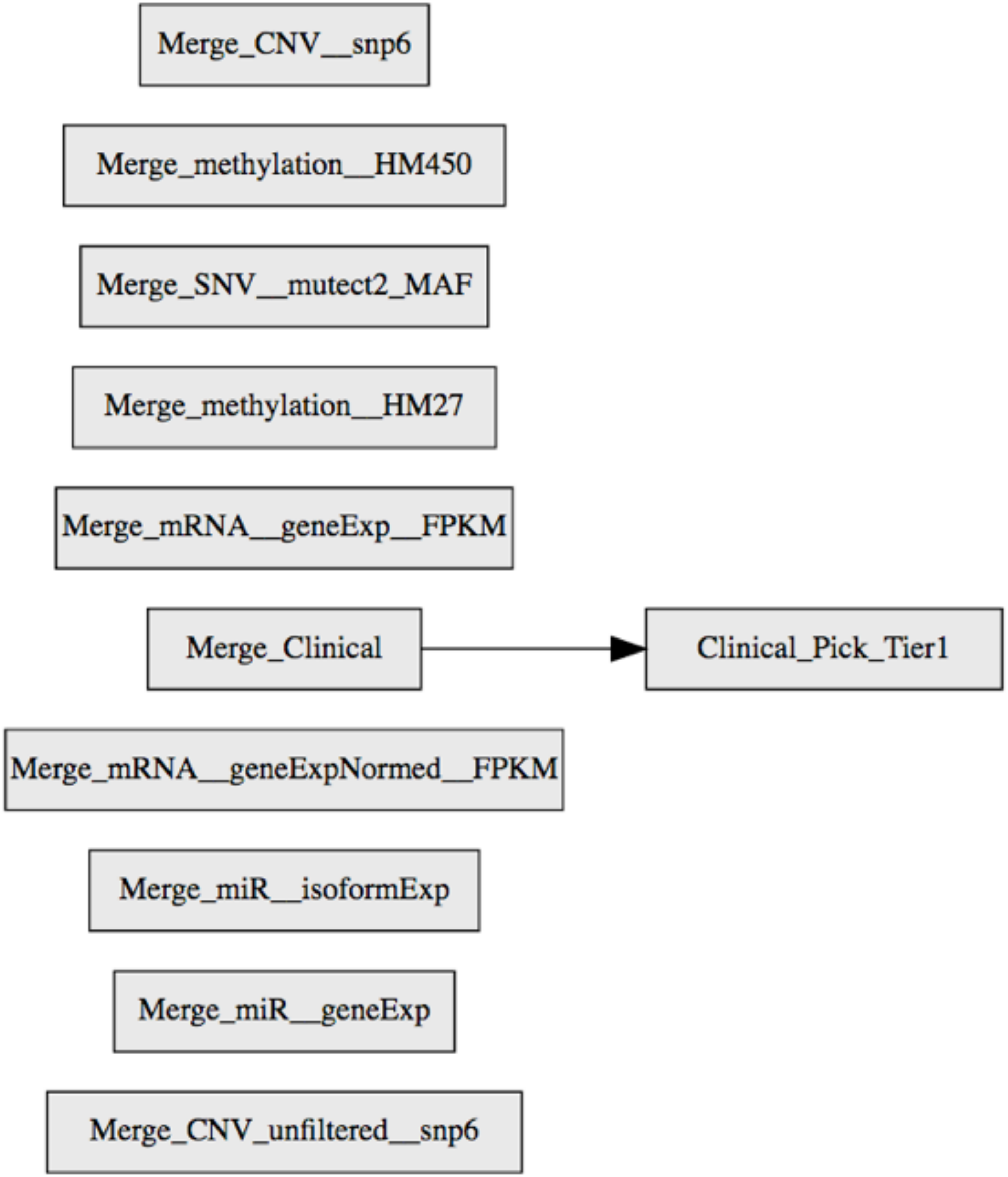
Massive scale production analysis pipeline
Unprecedented throughput & complexity
Delivered through clear & simple tools & visuals

Data : then and now

Legacy
stddata
workflow

86 tasks
(scores of data platforms)

Largely
N—>1
mergers



Merge_CNV__snp6

Merge_methylation__HM450

Merge_SNV__mutect2_MAF

Merge_methylation__HM27

Merge_mRNA__geneExp__FPKM

Merge_Clinical → Clinical_Pick_Tier1

Merge_mRNA__geneExpNormed__FPKM

Merge_miR__isoformExp

Merge_miR__geneExp

Merge_CNV_unfiltered__snp6

HG38 stddata workflow

11 tasks

Still N—>1 mergers

This is all the open-access data served from GDC

A welcome reduction in complexity

```
Merge_snp__genome_wide_snp_6__broad_mit_edu__Level_3__segmented_scna_hg18__seg
Merge_snp__genome_wide_snp_6__broad_mit_edu__Level_3__segmented_scna_hg19__seg
Merge_snp__genome_wide_snp_6__broad_mit_edu__Level_3__segmented_scna_minus_germline_cnv_hg18__seg
Merge_snp__genome_wide_snp_6__broad_mit_edu__Level_3__segmented_scna_minus_germline_cnv_hg19__seg

Merge_snp__human1mduo__hudsonalpha_org__Level_3__segmented_cna__seg
Merge_snp__human1mduo__hudsonalpha_org__Level_3__segmented_cnv__seg
Merge_snp__human1mduo__hudsonalpha_org__Level_3__segmented_loh__seg

Merge_snp__humanhap550__hudsonalpha_org__Level_3__segmented_cna__seg
Merge_snp__humanhap550__hudsonalpha_org__Level_3__segmented_cnv__seg
Merge_snp__humanhap550__hudsonalpha_org__Level_3__segmented_loh__seg
```

Example: 10 forms of legacy open-access CN data
(6 of them only accrued in TCGA pilot)

A welcome reduction in complexity

Merge_snp__genome_wide_snp_6__broad_mit_edu__Level_3__segmented_scna_hg18__seg
Merge_snp__genome_wide_snp_6__broad_mit_edu__Level_3__segmented_scna_hg19__seg
Merge_snp__genome_wide_snp_6__broad_mit_edu__Level_3__segmented_scna_minus_germline_cnv_hg18__seg
Merge_snp__genome_wide_snp_6__broad_mit_edu__Level_3__segmented_scna_minus_germline_cnv_hg19__seg

Merge_snp__human1mduo__hudsonalpha_org__Level_3__segmented_cna__seg
Merge_snp__human1mduo__hudsonalpha_org__Level_3__segmented_cnv__seg
Merge_snp__human1mduo__hudsonalpha_org__Level_3__segmented_loh__seg

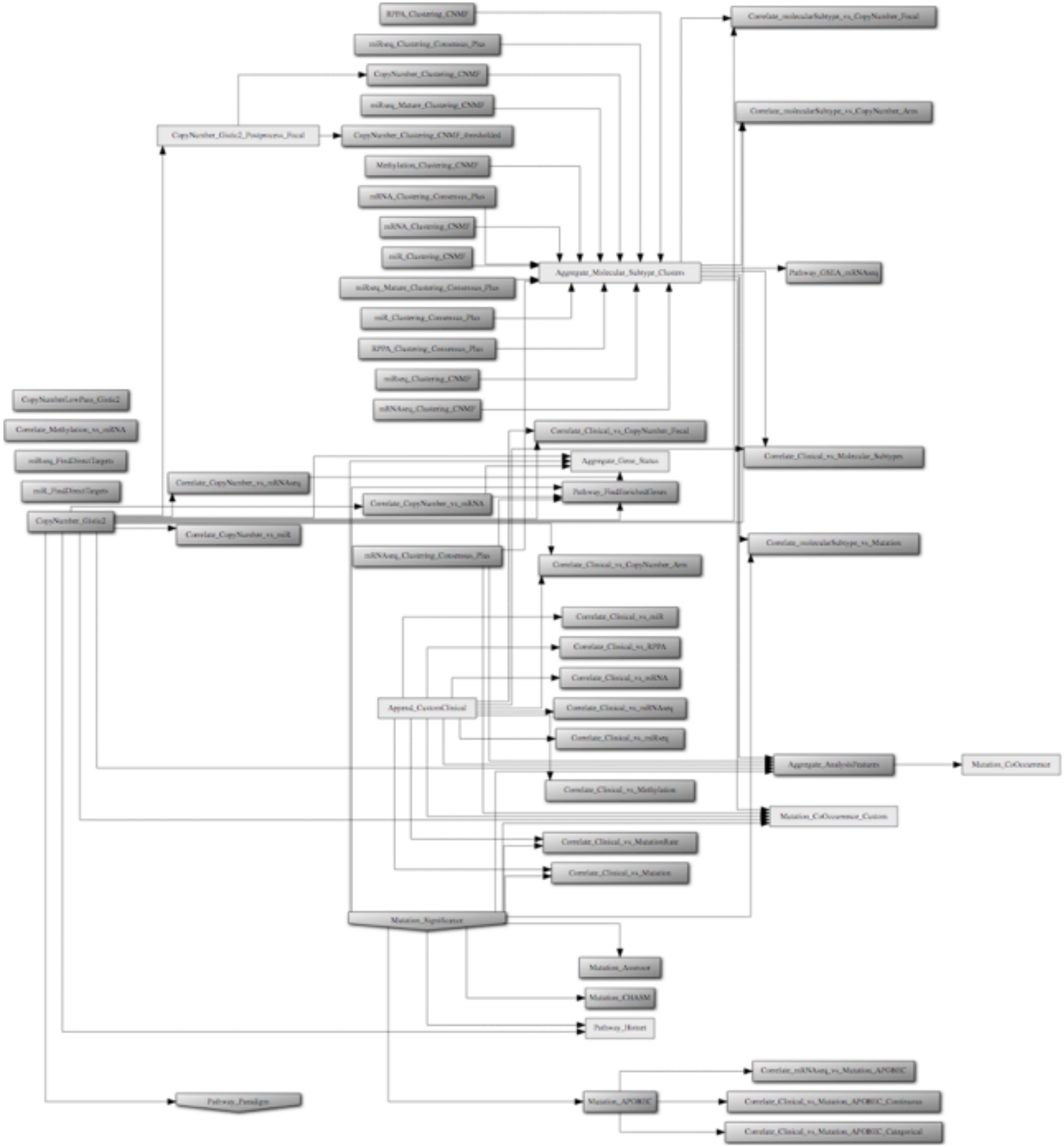
Merge_snp__humanhap550__hudsonalpha_org__Level_3__segmented_cna__seg
Merge_snp__humanhap550__hudsonalpha_org__Level_3__segmented_cnv__seg
Merge_snp__humanhap550__hudsonalpha_org__Level_3__segmented_loh__seg

Example: 10 forms of legacy open-access CN data
(6 of them only accrued in TCGA pilot)

Merge_CNV__snp6
Merge_CNV_unfiltered__snp6

Reduced to just 2 in HG38

Analyses : then and now

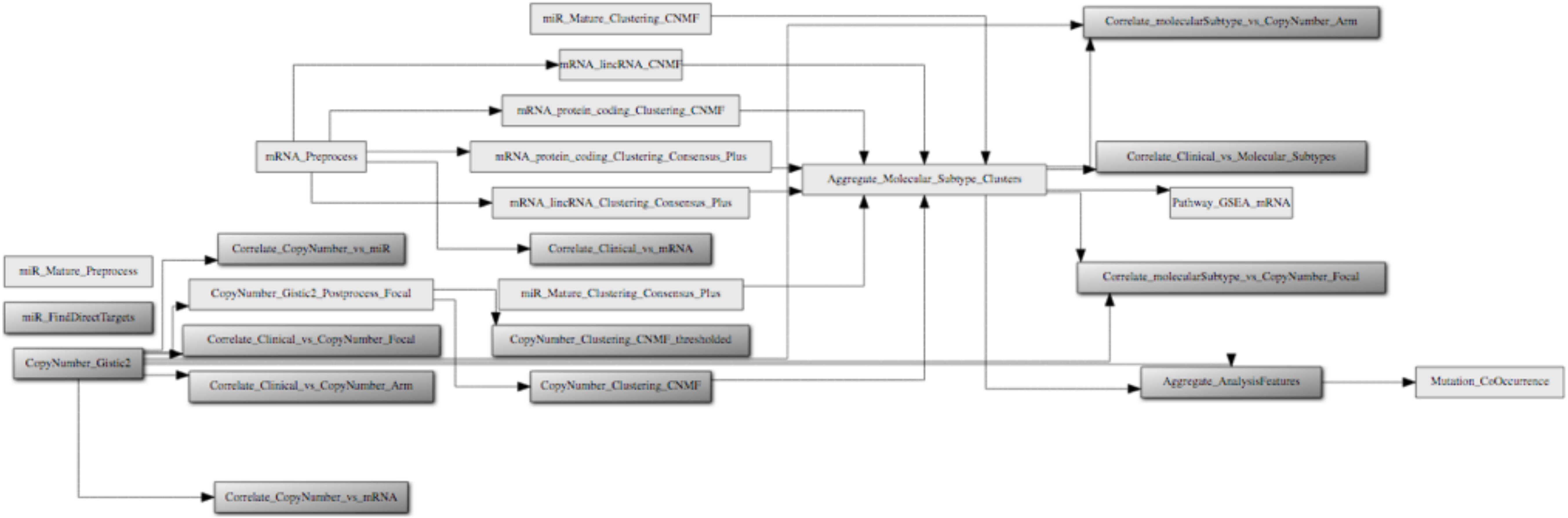


Legacy
TCGA
Analysis
Workflow

~100
coupled
pipelines

10 datatypes

HG38 Analysis Workflow



25 tasks and counting …

Less welcome reduction in complexity

Support for HG38-based analysis is inconsistent / scarce

Data only available for several months (including summer)

Some algorithms simply have not caught up yet

Particularly those sensitive to genomic position

Black-Box: substantial back/forth with algorithm developers

Part of this is also technical debt
Incurred during the rush to publication

HG38 Analysis Summary

Clustering: mRNA, miR, methylation, CN
(both protein-coding & *new lncRNA*)

Correlations: (all) clinical VS all clusters
clinical vs CN
CN vs miR & mRNA

Significance: CN (GISTIC) under evaluation
SNV (mutation) noticeably absent

Pathways: GSEA in, but not successful yet

iCoMut: code is unaffected, but waiting for integrated results table from complete workflow

FFPE versus Frozen

- Most samples in GDAN will be FFPE (many retrospective)
- With lower quality data than frozen*
- Acceptance criteria lower than TCGA
- Harder to disentangle biological signal from artifacts
- See [Getz 2012 TCGA talk]

This will also impact analyses
Firehose collected, packaged & distributed FFPE
But did not analyze
We will call attention to FFPE sample sets
And analyses performed upon them

---

* But FFPE may have higher purity

How GDCtools fits in

- Aggregation & cleansing is huge part of data science
- Indispensable to modern biomedical research
- Firehose performed this democratizing service in TCGA

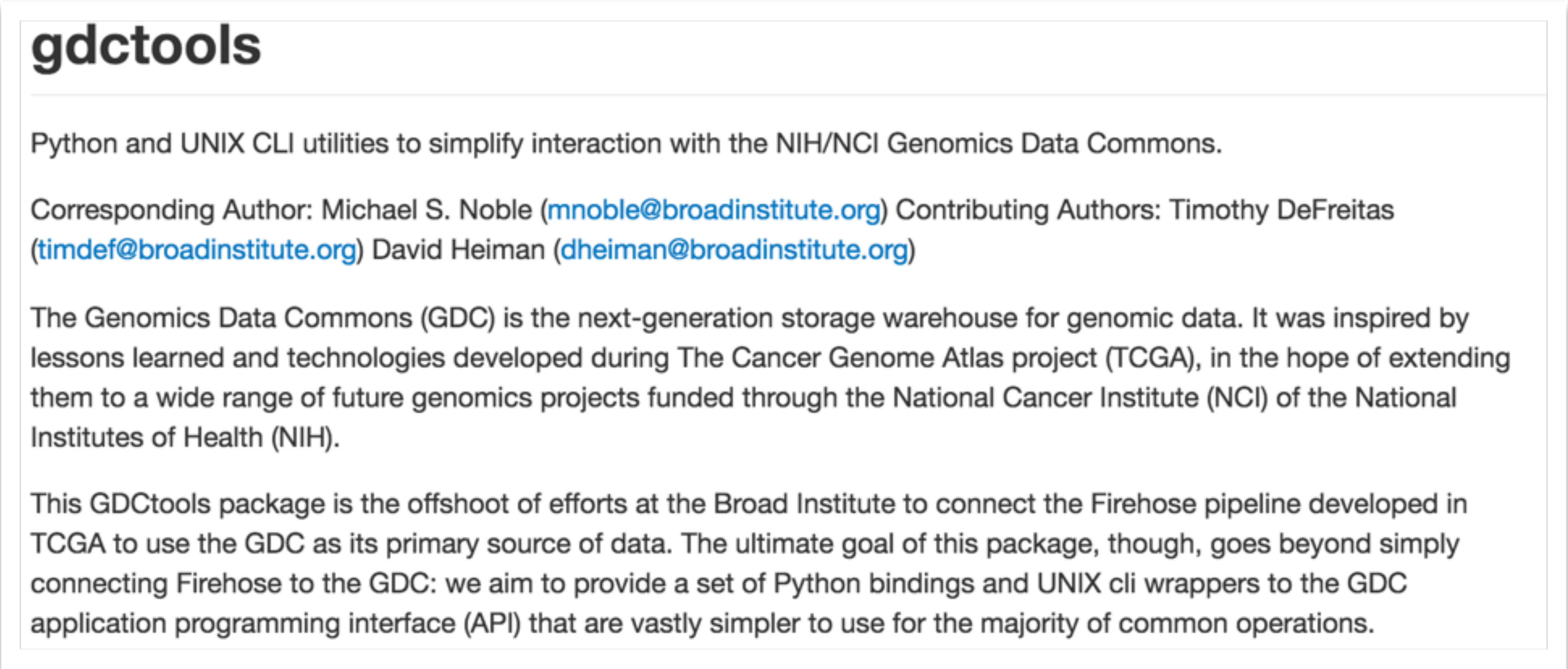
- Aggregation & cleansing is huge part of data science
- Indispensable to modern biomedical research
- Firehose performed this democratizing service in TCGA

GDCtools generalizes it to all GDAN programs
And makes it open-source for everyone

- Aggregation & cleansing is huge part of data science
- Indispensable to modern biomedical research
- Firehose performed this democratizing service in TCGA

GDCtools generalizes it to all GDAN programs
And makes it open-source for everyone

- Allows anyone to immediately program against GDC
- Mirror, dice, freezelist, sample reports … and more
- Begin in just minutes, no need to hire/train staff
- Or learn virtually any of the GDC API



gdctools

Python and UNIX CLI utilities to simplify interaction with the NIH/NCI Genomics Data Commons.

Corresponding Author: Michael S. Noble (mnoble@broadinstitute.org) Contributing Authors: Timothy DeFreitas (timdef@broadinstitute.org) David Heiman (dheiman@broadinstitute.org)

The Genomics Data Commons (GDC) is the next-generation storage warehouse for genomic data. It was inspired by lessons learned and technologies developed during The Cancer Genome Atlas project (TCGA), in the hope of extending them to a wide range of future genomics projects funded through the National Cancer Institute (NCI) of the National Institutes of Health (NIH).

This GDCtools package is the offshoot of efforts at the Broad Institute to connect the Firehose pipeline developed in TCGA to use the GDC as its primary source of data. The ultimate goal of this package, though, goes beyond simply connecting Firehose to the GDC: we aim to provide a set of Python bindings and UNIX cli wrappers to the GDC application programming interface (API) that are vastly simpler to use for the majority of common operations.

https://github.com/broadinstitute/gdctools

Goal: public beta in Q4 2016

Example: gdac_ingestor

Legacy implementation:

- Download TCGA data from DCC
- Annotate, redact, filter replicates
- Generate sample reports & loadfile
- Install to Firehose
- ~5K lines of Python (monolithic, not open)

Example: gdac_ingestor

Legacy implementation:

- Download TCGA data from DCC
- Annotate, redact, filter replicates
- Generate sample reports & loadfile
- Install to Firehose
- ~5K lines of Python (monolithic, not open)

Replaced by
GDCtools +
4 lines BASH

```
gdc_mirror      -config tcga.cfg
gdc_dice        -config tcga.cfg
sample_report   -config tcga.cfg
create_loadfile -config tcga.cfg
```

Already running nightly as cron job

- Mirror & dicing can be ***all*** or ***incremental***
- Highly configurable: even to just 1 case
- Example: to mirror TARGET

```
gdc_mirror     —config target.cfg
```

- Mirror & dicing can be ***all*** or ***incremental***
- Highly configurable: even to just 1 case
- Example: to mirror TARGET

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

- Example, to Google-bucketize

```
create_loadfile –config tcga.cfg,google.cfg
```

- Mirror & dicing can be ***all*** or ***incremental***
- Highly configurable: even to just 1 case
- Example: to mirror TARGET

```
gdc_mirror    -config target.cfg
```

- Example, to Google-bucketize

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

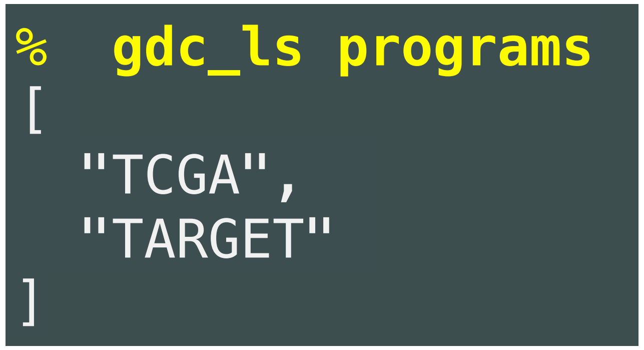
- Minimalist configuration, obeys *union semantics*

```
[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`

- Simple object framework, easy to extend & maintain
- Easy / familiar UNIX look-n-feel for computationalists

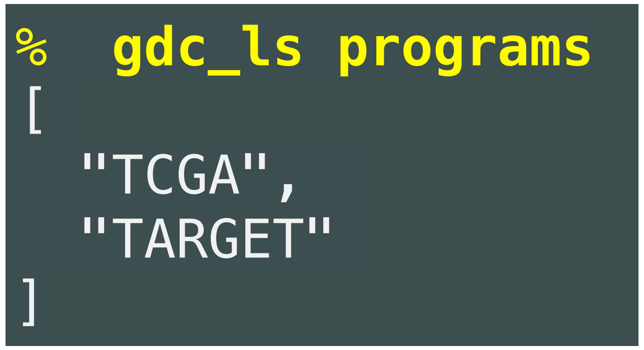
- Simple object framework, easy to extend & maintain
- Easy / familiar UNIX look-n-feel for computationalists



```
%  gdc_ls programs
[
  "TCGA",
  "TARGET"
]
```

What programs have exposed data?

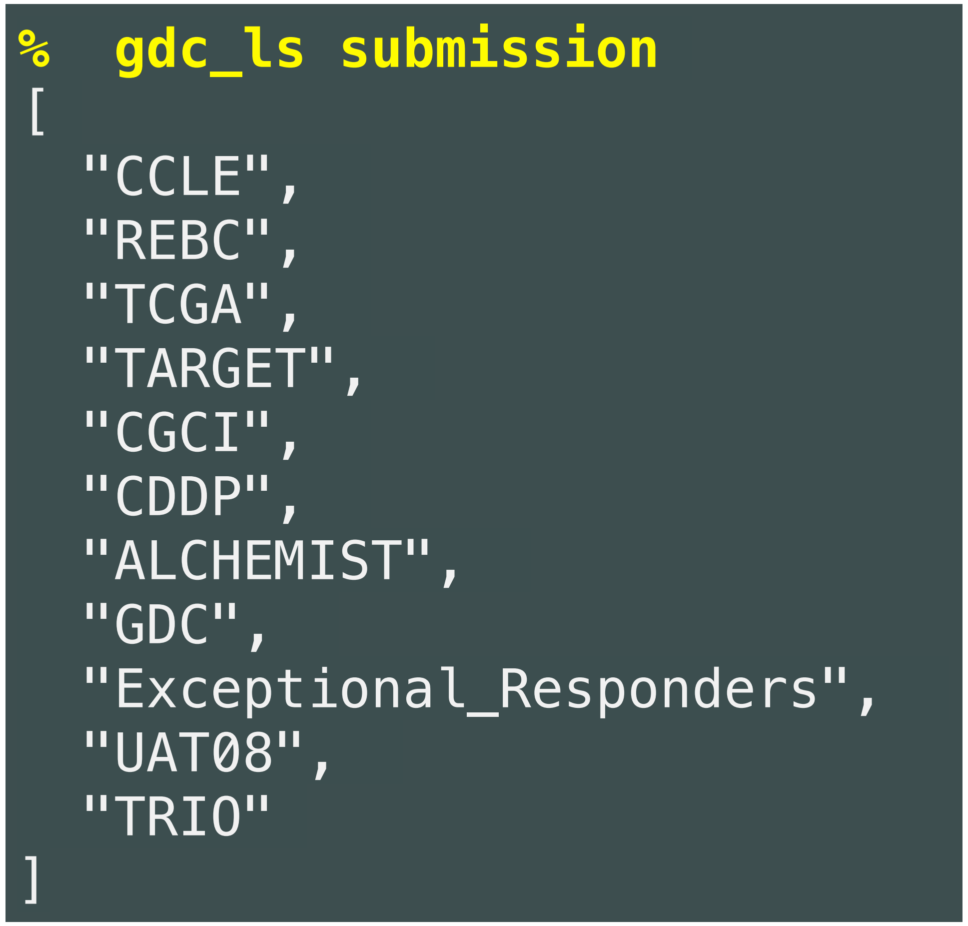
- Simple object framework, easy to extend & maintain
- Easy / familiar UNIX look-n-feel for computationalists



```
%  gdc_ls programs
[
  "TCGA",
  "TARGET"
]
```

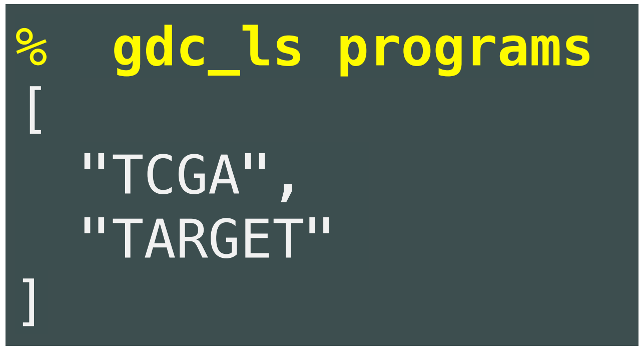
What programs have exposed data?

What programs have submitted data?



```
%  gdc_ls submission
[
  "CCLE",
  "REBC",
  "TCGA",
  "TARGET",
  "CGCI",
  "CDDP",
  "ALCHEMIST",
  "GDC",
  "Exceptional_Responders",
  "UAT08",
  "TRIO"
]
```

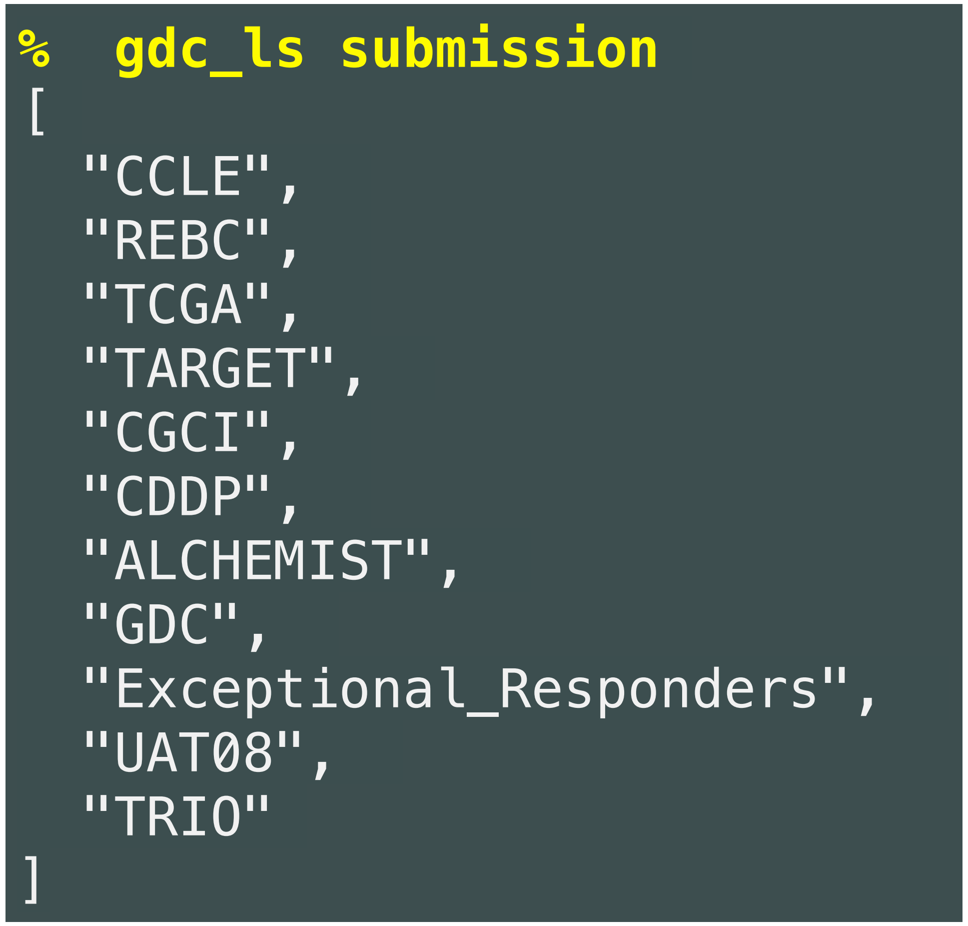
- Simple object framework, easy to extend & maintain
- Easy / familiar UNIX look-n-feel for computationalists



```
%  gdc_ls programs
[
  "TCGA",
  "TARGET"
]
```

What programs have exposed data?

What programs have submitted data?



```
%  gdc_ls submission
[
  "CCLE",
  "REBC",
  "TCGA",
  "TARGET",
  "CGCI",
  "CDDP",
  "ALCHEMIST",
  "GDC",
  "Exceptional_Responders",
  "UAT08",
  "TRIO"
]
```

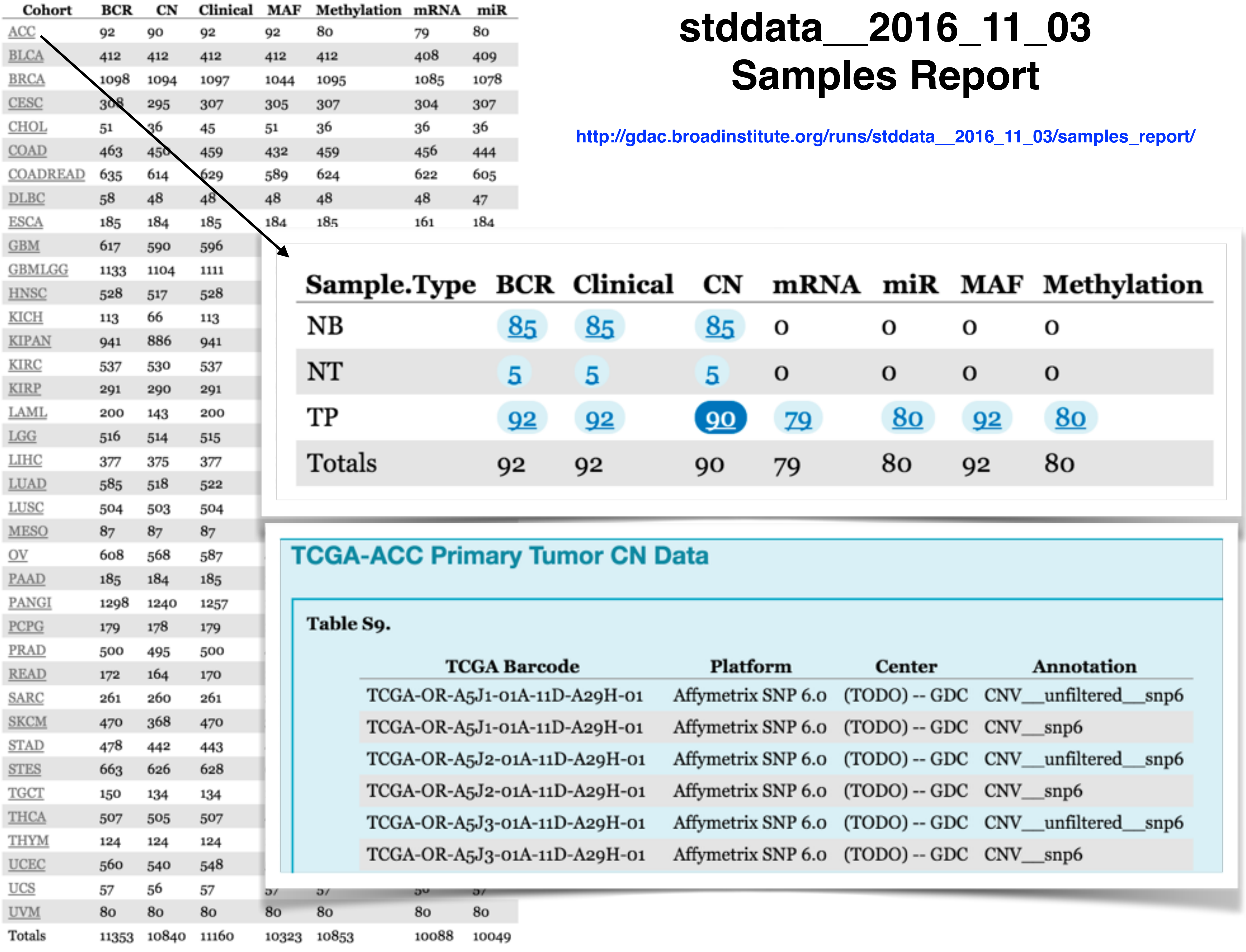
Auto-generated Python bindings: coming soon

Walk through of Artifacts

Credentials: gdcBeta (geeDseaB)

stddata__2016_11_03
Samples Report

http://gdac.broadinstitute.org/runs/stddata__2016_11_03/samples_report/



Cohort	BCR	CN	Clinical	MAF	Methylation	mRNA	miR
ACC	92	90	92	92	80	79	80
BLCA	412	412	412	412	412	408	409
BRCA	1098	1094	1097	1044	1095	1085	1078
CESC	308	295	307	305	307	304	307
CHOL	51	36	45	51	36	36	36
COAD	463	456	459	432	459	456	444
COADREAD	635	614	629	589	624	622	605
DLBC	58	48	48	48	48	48	47
ESCA	185	184	185	184	185	161	184
GBM	617	590	596				
GBMLGG	1133	1104	1111				
HNSC	528	517	528				
KICH	113	66	113				
KIPAN	941	886	941				
KIRC	537	530	537				
KIRP	291	290	291				
LAML	200	143	200				
LGG	516	514	515				
LIHC	377	375	377				
LUAD	585	518	522				
LUSC	504	503	504				
MESO	87	87	87				
OV	608	568	587				
PAAD	185	184	185				
PANGI	1298	1240	1257				
PCPG	179	178	179				
PRAD	500	495	500				
READ	172	164	170				
SARC	261	260	261				
SKCM	470	368	470				
STAD	478	442	443				
STES	663	626	628				
TGCT	150	134	134				
THCA	507	505	507				
THYM	124	124	124				
UCEC	560	540	548				
UCS	57	56	57				
UVM	80	80	80	80	80	80	80
Totals	11353	10840	11160	10323	10853	10088	10049

Sample.Type	BCR	Clinical	CN	mRNA	miR	MAF	Methylation
NB	85	85	85	0	0	0	0
NT	5	5	5	0	0	0	0
TP	92	92	90	79	80	92	80
Totals	92	92	90	79	80	92	80

TCGA-ACC Primary Tumor CN Data

Table S9.

TCGA Barcode	Platform	Center	Annotation
TCGA-OR-A5J1-01A-11D-A29H-01	Affymetrix SNP 6.0	(TODO) -- GDC	CNV__unfiltered__snp6
TCGA-OR-A5J1-01A-11D-A29H-01	Affymetrix SNP 6.0	(TODO) -- GDC	CNV__snp6
TCGA-OR-A5J2-01A-11D-A29H-01	Affymetrix SNP 6.0	(TODO) -- GDC	CNV__unfiltered__snp6
TCGA-OR-A5J2-01A-11D-A29H-01	Affymetrix SNP 6.0	(TODO) -- GDC	CNV__snp6
TCGA-OR-A5J3-01A-11D-A29H-01	Affymetrix SNP 6.0	(TODO) -- GDC	CNV__unfiltered__snp6
TCGA-OR-A5J3-01A-11D-A29H-01	Affymetrix SNP 6.0	(TODO) -- GDC	CNV__snp6

```
%    firehose_get  -auth gdcBeta:geeDseaB   data   2016_11_03
```

```
Validating run selection against Broad Institute website ...
You've asked to download archives for the following disease cohorts

     TCGA-ACC TCGA-BLCA TCGA-BRCA TCGA-CESC TCGA-CHOL ...

from the stddata__2016_11_03 Firehose run. If this is correct,
shall we continue with download? (y|yes|n|no) [no]?

Attempting to retrieve data for Broad GDAC run stddata__2016_11_03 ...
--2016-11-17 14:21:21--  https://gdac.broadinstitute.org/runs/stddata__2016_11_03/data/TCGA-ACC/20161103/
gdac.broadinstitute.org_TCGA-ACC-NB.Clinical_Pick_Tier1.Level_4.2016110300.0.0.tar.gz
Saving to: `stddata__2016_11_03/TCGA-ACC/20161103/gdac.broadinstitute.org_TCGA-ACC-
NB.Clinical_Pick_Tier1.Level_4.2016110300.0.0.tar.gz'
--2016-11-17 14:21:21--  https://gdac.broadinstitute.org/runs/stddata__2016_11_03/data/TCGA-ACC/20161103/
gdac.broadinstitute.org_TCGA-ACC-NB.Clinical_Pick_Tier1.Level_4.2016110300.0.0.tar.gz.md5
Saving to: `stddata__2016_11_03/TCGA-ACC/20161103/gdac.broadinstitute.org_TCGA-ACC-
NB.Clinical_Pick_Tier1.Level_4.2016110300.0.0.tar.gz.md5'
--2016-11-17 14:21:22--  https://gdac.broadinstitute.org/runs/stddata__2016_11_03/data/TCGA-ACC/20161103/
gdac.broadinstitute.org_TCGA-ACC-NB.Clinical_Pick_Tier1.aux.2016110300.0.0.tar.gz

...
```

firehose_get v0.4.8 may be obtained here

Demo gdcbeta FireBrowse

Analysis Plan : In Brief

Massive corpus of Firehose HG38 results
Needs substantial review prior to public release
Want to help?

Continue iterating with algorithm developers
Integrate when ready:

- Mutation assessor (MSKCC) Dec 2016
- MutSig: in test by Dec 2016
- Oncotator: possibly Dec 2016, but more likely Q1 2017
- APOBEC (Gordenin et al, NIH): Dec 2016

Pay down tech debt: Best-practices migration of legacy pipelines
To cloud, to reduce storage / compute Cloud costs

Aim is for first public release : late Q1 2017

Towards Global Infrastructure for Collaborative High-throughput Cancer Genomics Analysis

Motivation: The TCGA is an international model for cancer genome projects.

But still lacks a **consensus, open-access, fully collaborative and reproducible system for extreme-scale integrative analysis**.

Work is done by **Analysis Working Groups (AWGs)** operating in cycles (and branches) of freezing/thawing data, analyses, figures, reports and biological findings, until convergence on scientific paper(s).

To maximally advance **scientific knowledge** and its **dissemination**, we need **a single place that combines data, tools, results, discoveries and compute environment.**

FireCloud can be for collaborative science
What Google drive/docs is for collaborative writing

With The Following Aims

Aim 1 - Global Infrastructure for Collaborative Extreme-Scale Cancer Analysis
Broad-internal Firehose → **FireCloud with data from GDC, for Standard runs, AWGs and GDACs.** A collaborative open virtual shared infrastructure (Reproducibility/DOIs) that combines data, tools, results, **and compute environment** in one place!

Aim 2 - Operation of Standard Workflows at Scale
Lead a collaborative effort to define and regularly operate the **GDAN Standard Workflow** on data freezes (as in Firehose). Serve as a first pass analysis for AWGs.

Aim 3 - Rapid Continuous Evolution of Standard Workflows
We will integrate new tools for QC, multi-samples, clinical, etc.. GDACs and Developers can develop, test, benchmark and demonstrate usefulness of tools (self service). The GDAN will promote tools to become part of the Standard GDAN workflow.

Aim 4 – Improved Capabilities for Scientific Exploration, Clinical Diagnostics, Reproducibility and Scientific Discoveries
Generation of automated text and graphical reports and constantly updating and growing a **scientific knowledge base** to easily identify new scientific discoveries (“what’s new?”)

(Funded for GDAN)

From Previous Talk: GDC & Cloud

GDAN will combine GDC and cloud-based analysis
But why copy data from GDC to local compute?
We can save money, time, and reduce confusion IF:

- Instead of ONLY supporting JSON download of data
- GDC also loaded data into cloud storage
- And exposed data via bucket-ized URIs
- So that algorithms in cloud-based analysis systems
- Don't have to copy data from GDC, but rather just reference it in available cloud storage
- Avoiding double-copies and additional costs (double-pay)

Fin