Progress Report:

GDAC Firehose Integration With The Genomic Data Commons

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

mnoble@broadinstitute.org

2016_07_01

Outline

- Vision for this work
- Brief review of TCGA GDAC Firehose
- GDC-based runs to date
- Closing remarks

Vision

- While bringing the past to a good resting place:
  - ✓ Finalizing TCGA (i.e. AWGs)

- We are simultaneously working on the future
  - ✓ GDC integration
  - ✓ FireCloud

- We view latter as 2 thrusts of a single goal:
  - ✓ ***GDAN Readiness***
  - ✓ If funded, we **will** be ready to go on day 1
  - ✓ And provide open tools to clarify, scale & democratize
  - ✓ Example: *GDCtools* GIT repo will provide:
    - ✓ Auto-generated Python bindings to GDC api
    - ✓ Simple means by which other centers can mirror GDC
    - ✓ And generate sample freeze lists (loadfiles) (common activity in AWGs)

Timeline

Original

- To be underway by Jan 2016 and “ready” by late May/June
- With 3 major deliverables:
    - Firehose capable of ingesting both legacy & new data from GDC
    - Set of open source Python bindings to the GDC API
    - Set of notes, with possibly additional software artifacts, to help other researchers and data centers adapt to the GDC

Timeline

Original

- To be underway by Jan 2016 and “ready” by late May/June
- With 3 major deliverables:
  - Firehose capable of ingesting both legacy & new data from GDC
  - Set of open source Python bindings to the GDC API
  - Set of notes, with possibly additional software artifacts, to help other researchers and data centers adapt to the GDC

Reality

- Formally began April 2016
- Significant progress already on all major goals
- In light of ***GDAN Readiness*** vision, added 2 more goals:
  - ✓ Release 2 final GDC-based Firehose runs: data, analysis
  - ✓ Encapsulating final snapshot of TCGA data
  - ✓ *Likely mix of GDC & DCC data, to fill holes (e.g. RPPA, Methylation)*

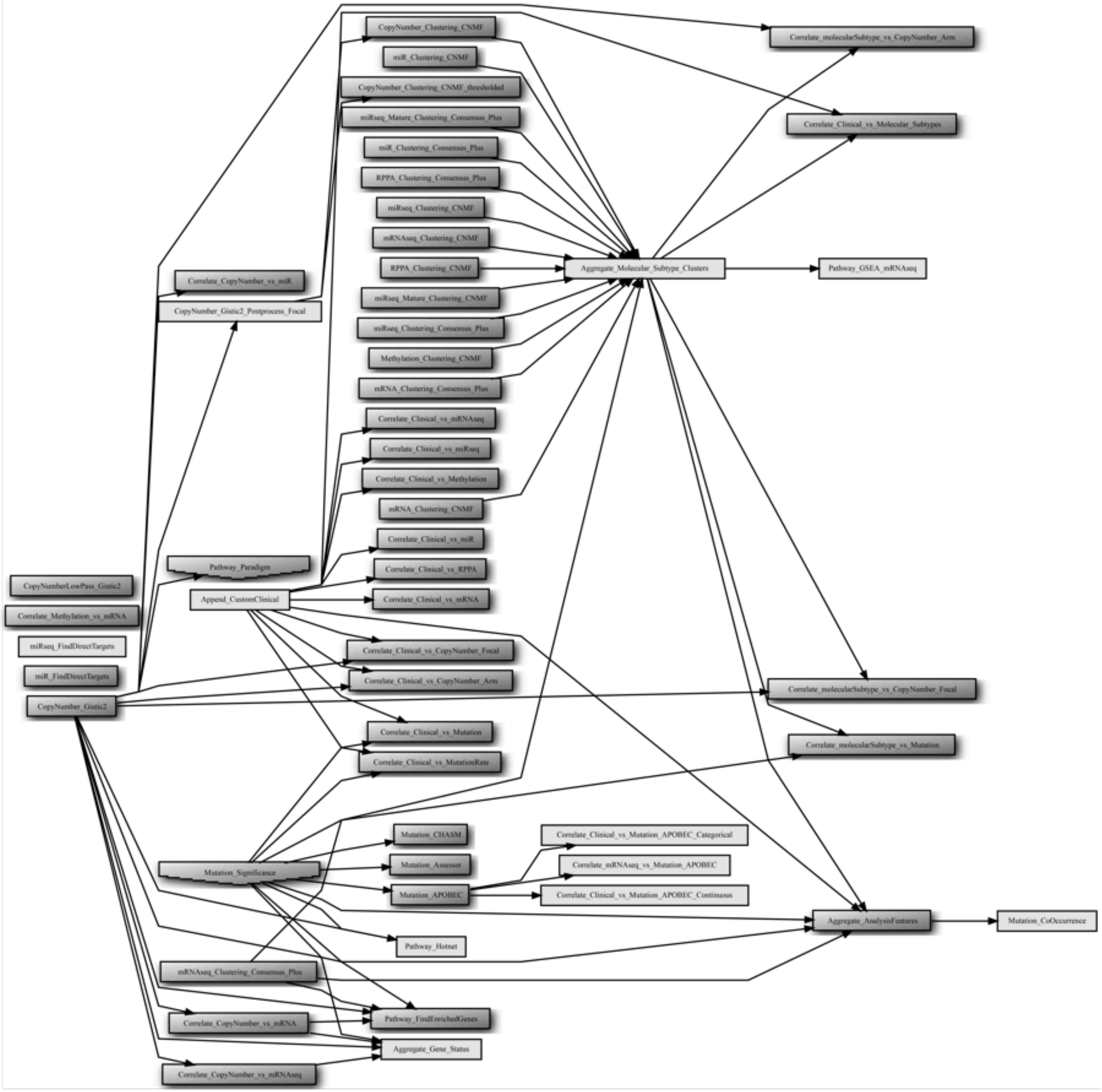
Firehose Review

- Sophisticated job execution manager
- Operated through browser UI and API (even command line)
- So you don't have to manage compute farm (e.g. LSF or SGE)
- At the time, nothing else scaled to TCGA levels **AND** included job avoidance (*way to skip jobs already run*) **AND** organized data in biologically-informed data model (*pairs, sample sets, etc*) **AND** file-system-blind manner (*annotations*)

Instances for multiple projects & labs:

TCGA GDAC (us), TCGA sequencing, GTEX, ICGC, Garraway, Meyerson, Wu …

Allows HPC (compute farm) to be easily exploited by all:
Technical power users (us) AND itinerant M.D.s /P.I.s


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
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
Mutation_CHASM
Mutation_Assessor
Mutation_APOBEC
Mutation_Significance
Correlate_Clinical_vs_Mutation_APOBEC_Categorical
Correlate_mRNAseq_vs_Mutation_APOBEC
Correlate_Clinical_vs_Mutation_APOBEC_Continuous
Aggregate_AnalysisFeatures
Mutation_CoOccurrence
Pathway_Hotnet
mRNAseq_Clustering_Consensus_Plus
Pathway_FindEnrichedGenes
Correlate_CopyNumber_vs_mRNA
Aggregate_Gene_Status
Correlate_CopyNumber_vs_mRNAseq

GDAC Analysis Workflow

Run on all TCGA cohorts
>100 tasks per

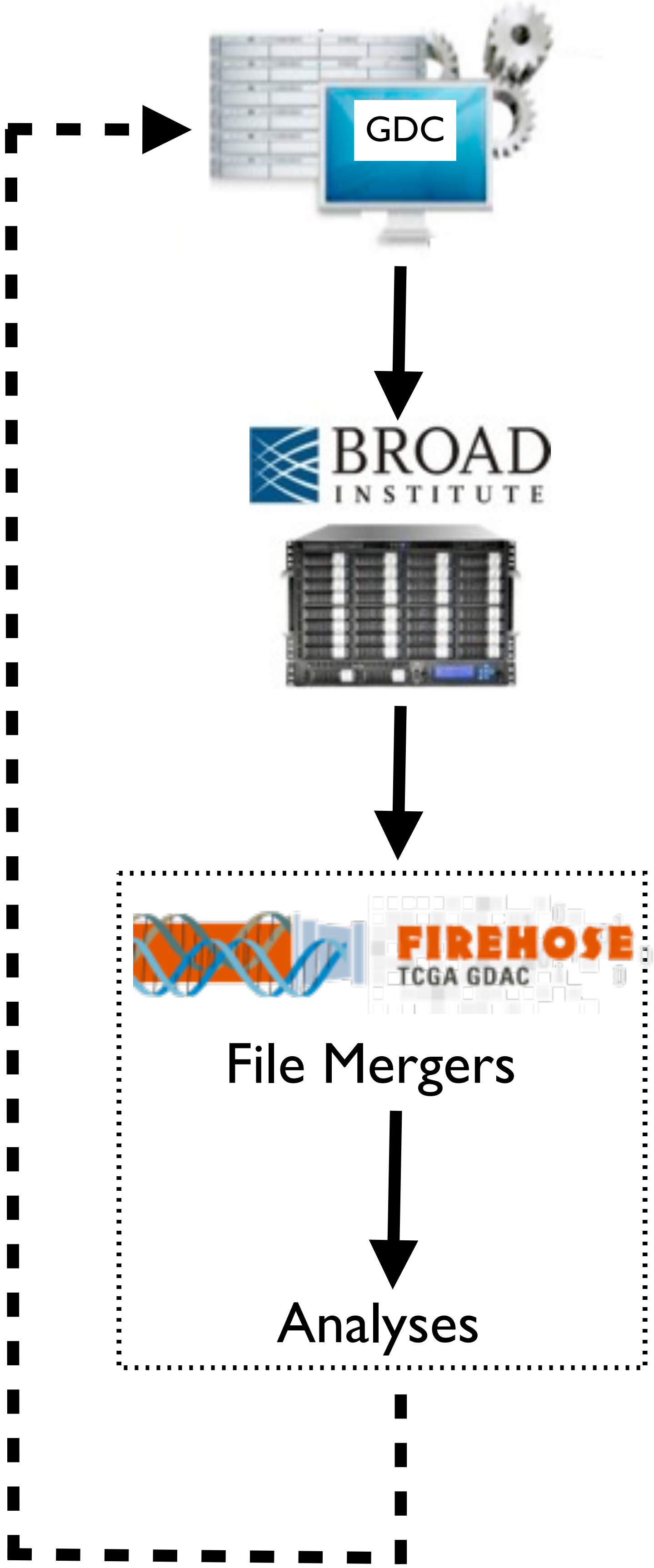
Generates 1500 result reports

Can be initiated by SWEs, CBs, or M.Ds & P.I.s

Plays several key roles

- extreme scale production pipeline
- analytic forest-clearing for researchers & MDs
- democratization for use beyond TCGA proper
- simplification for everyone
- pushing envelope for rigor @ scale, reproducibility, APIs

TCGA GDAC Pipeline Data Flow



File Mergers

Analyses

Sample Data

Automated Mirror To Local Disk

Mirror
(full TCGA ~24 hours)
(diff ~minutes)

Dice

Controlled Ingestion
GO!

Ingest

Pipeline Automated Runs

<24 hours for data run

N days for analysis run

Results

Dicing : part of Firehose secret sauce

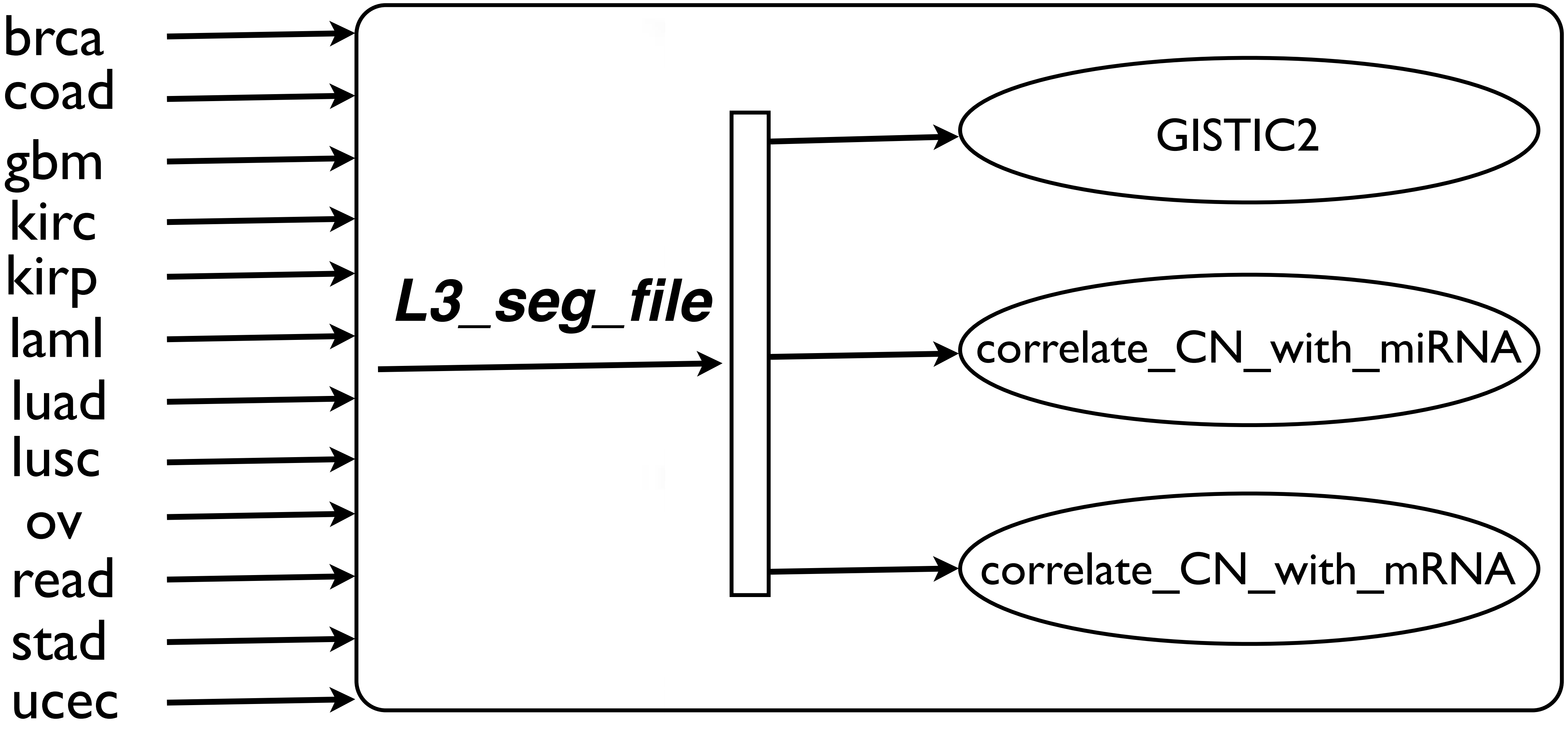
- Scans mirror to interpret & normalize files
  (e.g. in TCGA would make mage-tab from TSV)

- And generate loadfiles
- … large tables of all sample files in cohort
- Used to fill in values of Firehose annotations

- Like mirror, dicing can be ***all*** or ***incremental***

Annotations: more FireHose secret sauce

- Logical identifier for datum: input or output
- Abstracts file system knowledge from algorithms
- Transparent multiplexing across TCGA tumor types



By encapsulating data and algorithm parameters within abstract annotations, instead of only literal values or explicit file system references, Firehose is able to execute analyses codes in data-blind manner across a wide variety of inputs, without modification or onerous bookkeeping for end users. This has proven to be a powerful metaphor for interacting with TCGA data, for example, because once an algorithm is in Firehose it can run on either a single tumor type or all of them with equal ease.

Why talk about loadfiles?

Well, in addition to enabling Firehose

A big part of AWG analysis is establishing sample freeze lists.

And loadfiles are simply sample freeze lists by a different name.

And our *GDCtools* repo will enable this to be done easily by anyone.

GDC-based Firehose pipeline runs

3 so far, data snapshot version given in bold
Harmonized data / API, not legacy

2016_05_25

1 cohort/project (TCGA-ACC)
2 available datatypes (Clinical, Copy Number)

2016_05_27

33 cohorts (all of core TCGA cohorts)
still only 2 datatypes

2016_06_29

38 cohorts: 33 core TCGA + 5 aggregates

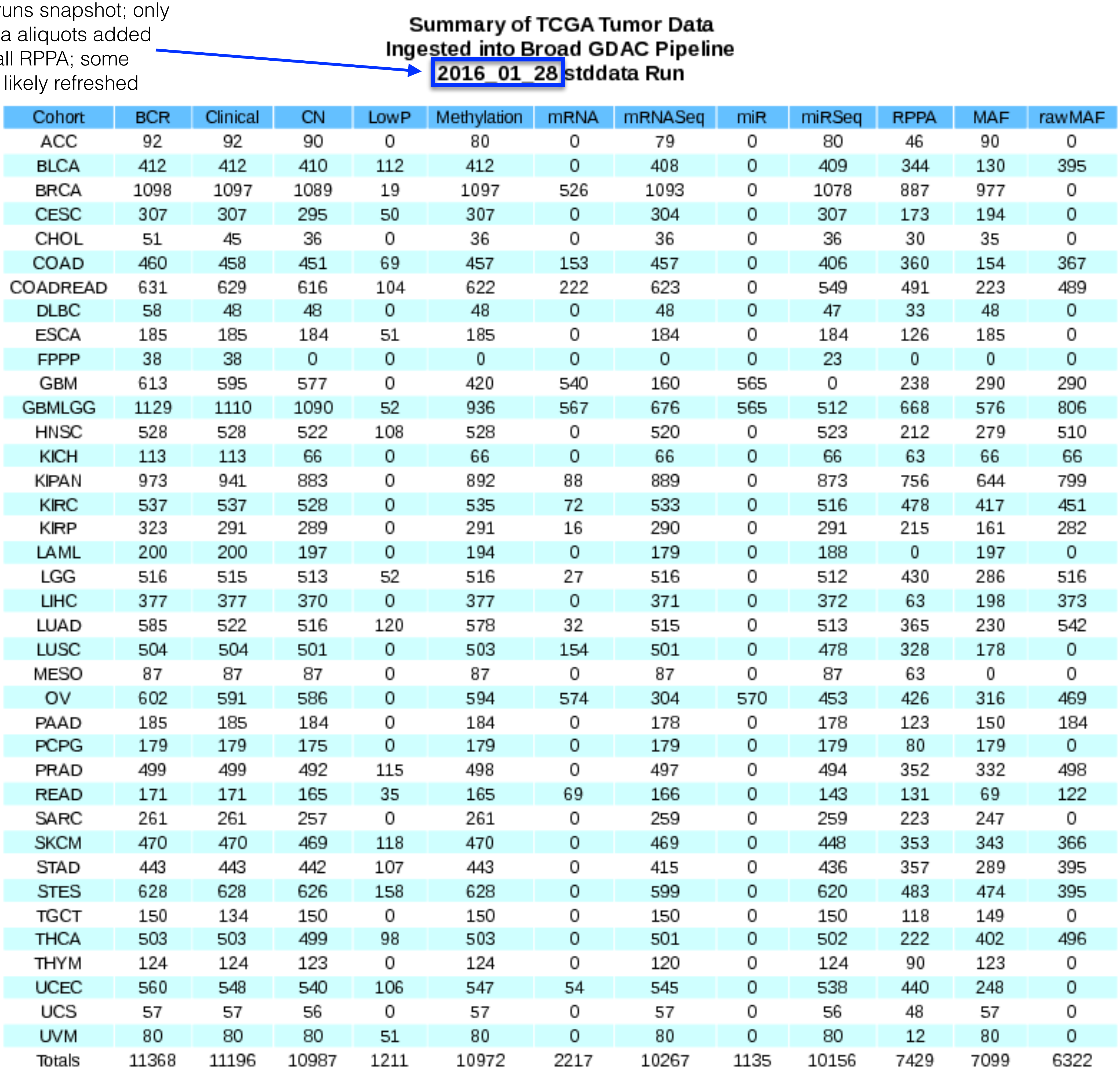
COADREAD:	COAD,	READ		
GBMLGG:	GBM,	LGG		
KIPAN:	KICH,	KIRC,	KIRP	
STES:	STAD,	ESCA		
PANGI:	COAD,	READ,	STAD,	ESCA

4 datatypes: Clinical, CN, miRSeq, mRNASeq

Summary of TCGA Tumor Data
Ingested into Broad GDAC Pipeline
2016_01_28 stddata Run

Cohort	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNASeq	miR	miRSeq	RPPA	MAF	rawMAF
ACC	92	92	90	0	80	0	79	0	80	46	90	0
BLCA	412	412	410	112	412	0	408	0	409	344	130	395
BRCA	1098	1097	1089	19	1097	526	1093	0	1078	887	977	0
CESC	307	307	295	50	307	0	304	0	307	173	194	0
CHOL	51	45	36	0	36	0	36	0	36	30	35	0
COAD	460	458	451	69	457	153	457	0	406	360	154	367
COADREAD	631	629	616	104	622	222	623	0	549	491	223	489
DLBC	58	48	48	0	48	0	48	0	47	33	48	0
ESCA	185	185	184	51	185	0	184	0	184	126	185	0
FPPP	38	38	0	0	0	0	0	0	23	0	0	0
GBM	613	595	577	0	420	540	160	565	0	238	290	290
GBMLGG	1129	1110	1090	52	936	567	676	565	512	668	576	806
HNSC	528	528	522	108	528	0	520	0	523	212	279	510
KICH	113	113	66	0	66	0	66	0	66	63	66	66
KIPAN	973	941	883	0	892	88	889	0	873	756	644	799
KIRC	537	537	528	0	535	72	533	0	516	478	417	451
KIRP	323	291	289	0	291	16	290	0	291	215	161	282
LAML	200	200	197	0	194	0	179	0	188	0	197	0
LGG	516	515	513	52	516	27	516	0	512	430	286	516
LIHC	377	377	370	0	377	0	371	0	372	63	198	373
LUAD	585	522	516	120	578	32	515	0	513	365	230	542
LUSC	504	504	501	0	503	154	501	0	478	328	178	0
MESO	87	87	87	0	87	0	87	0	87	63	0	0
OV	602	591	586	0	594	574	304	570	453	426	316	469
PAAD	185	185	184	0	184	0	178	0	178	123	150	184
PCPG	179	179	175	0	179	0	179	0	179	80	179	0
PRAD	499	499	492	115	498	0	497	0	494	352	332	498
READ	171	171	165	35	165	69	166	0	143	131	69	122
SARC	261	261	257	0	261	0	259	0	259	223	247	0
SKCM	470	470	469	118	470	0	469	0	448	353	343	366
STAD	443	443	442	107	443	0	415	0	436	357	289	395
STES	628	628	626	158	628	0	599	0	620	483	474	395
TGCT	150	134	150	0	150	0	150	0	150	118	149	0
THCA	503	503	499	98	503	0	501	0	502	222	402	496
THYM	124	124	123	0	124	0	120	0	124	90	123	0
UCEC	560	548	540	106	547	54	545	0	538	440	248	0
UCS	57	57	56	0	57	0	57	0	56	48	57	0
UVM	80	80	80	51	80	0	80	0	80	12	80	0
Totals	11368	11196	10987	1211	10972	2217	10267	1135	10156	7429	7099	6322

Last public runs snapshot; only 718 new data aliquots added since then; all RPPA; some clinical data likely refreshed



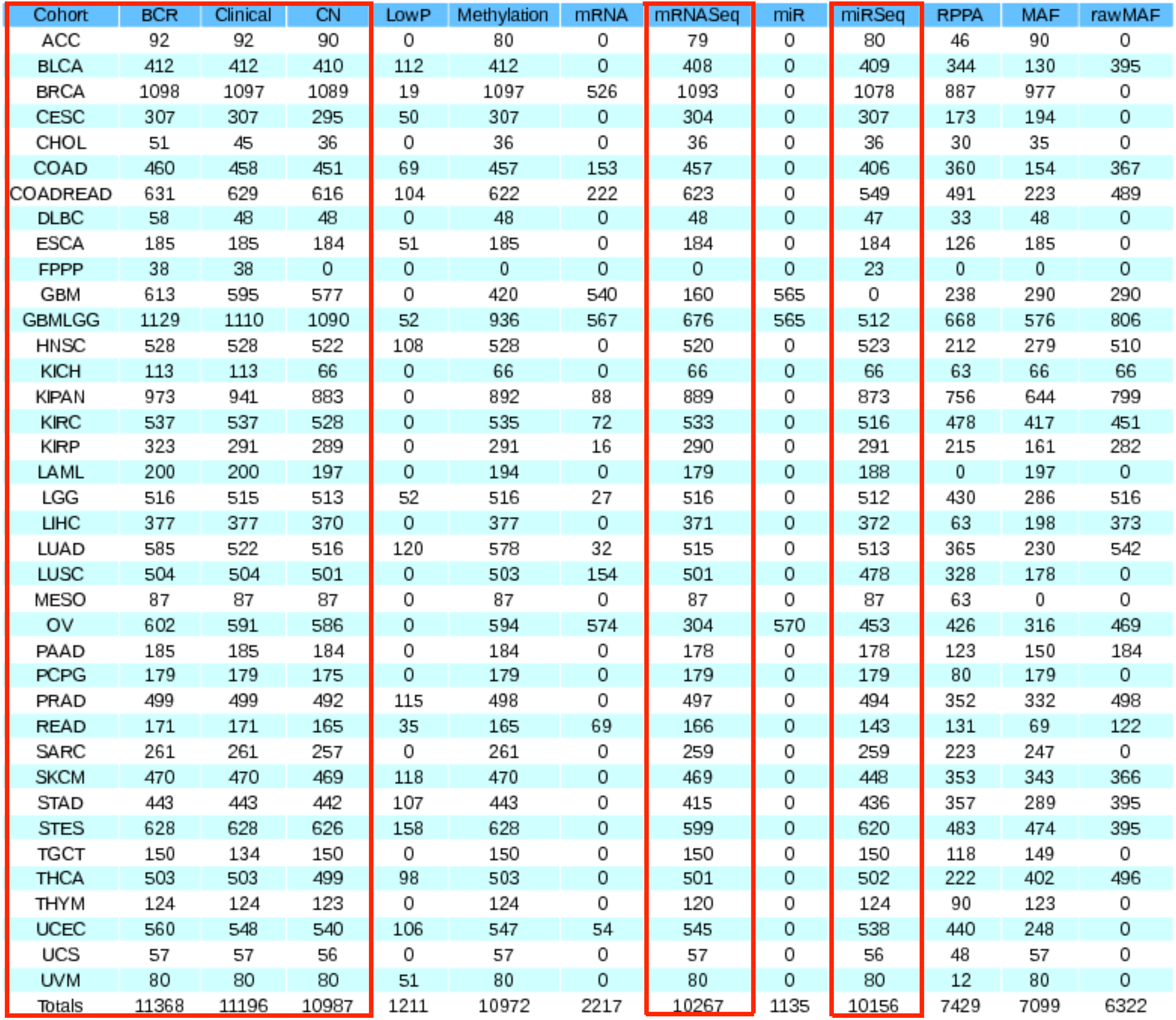
Summary of TCGA Tumor Data Ingested into Broad GDAC Pipeline 2016_01_28 stddata Run

Cohort	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNASeq	miR	miRSeq	RPPA	MAF	rawMAF
ACC	92	92	90	0	80	0	79	0	80	46	90	0
BLCA	412	412	410	112	412	0	408	0	409	344	130	395
BRCA	1098	1097	1089	19	1097	526	1093	0	1078	887	977	0
CESC	307	307	295	50	307	0	304	0	307	173	194	0
CHOL	51	45	36	0	36	0	36	0	36	30	35	0
COAD	460	458	451	69	457	153	457	0	406	360	154	367
COADREAD	631	629	616	104	622	222	623	0	549	491	223	489
DLBC	58	48	48	0	48	0	48	0	47	33	48	0
ESCA	185	185	184	51	185	0	184	0	184	126	185	0
FPPP	38	38	0	0	0	0	0	0	23	0	0	0
GBM	613	595	577	0	420	540	160	565	0	238	290	290
GBMLGG	1129	1110	1090	52	936	567	676	565	512	668	576	806
HNSC	528	528	522	108	528	0	520	0	523	212	279	510
KICH	113	113	66	0	66	0	66	0	66	63	66	66
KIPAN	973	941	883	0	892	88	889	0	873	756	644	799
KIRC	537	537	528	0	535	72	533	0	516	478	417	451
KIRP	323	291	289	0	291	16	290	0	291	215	161	282
LAML	200	200	197	0	194	0	179	0	188	0	197	0
LGG	516	515	513	52	516	27	516	0	512	430	286	516
LIHC	377	377	370	0	377	0	371	0	372	63	198	373
LUAD	585	522	516	120	578	32	515	0	513	365	230	542
LUSC	504	504	501	0	503	154	501	0	478	328	178	0
MESO	87	87	87	0	87	0	87	0	87	63	0	0
OV	602	591	586	0	594	574	304	570	453	426	316	469
PAAD	185	185	184	0	184	0	178	0	178	123	150	184
PCPG	179	179	175	0	179	0	179	0	179	80	179	0
PRAD	499	499	492	115	498	0	497	0	494	352	332	498
READ	171	171	165	35	165	69	166	0	143	131	69	122
SARC	261	261	257	0	261	0	259	0	259	223	247	0
SKCM	470	470	469	118	470	0	469	0	448	353	343	366
STAD	443	443	442	107	443	0	415	0	436	357	289	395
STES	628	628	626	158	628	0	599	0	620	483	474	395
TGCT	150	134	150	0	150	0	150	0	150	118	149	0
THCA	503	503	499	98	503	0	501	0	502	222	402	496
THYM	124	124	123	0	124	0	120	0	124	90	123	0
UCEC	560	548	540	106	547	54	545	0	538	440	248	0
UCS	57	57	56	0	57	0	57	0	56	48	57	0
UVM	80	80	80	51	80	0	80	0	80	12	80	0
Totals	11368	11196	10987	1211	10972	2217	10267	1135	10156	7429	7099	6322

Last public runs snapshot; only 718 new data aliquots added since then; all RPPA; some clinical data likely refreshed

Summary of TCGA Tumor Data Ingested into Broad GDAC Pipeline 2016_01_28 stddata Run

What we have from GDC; sample counts between DCC and GDC mirrors are not exact; more forensics are needed



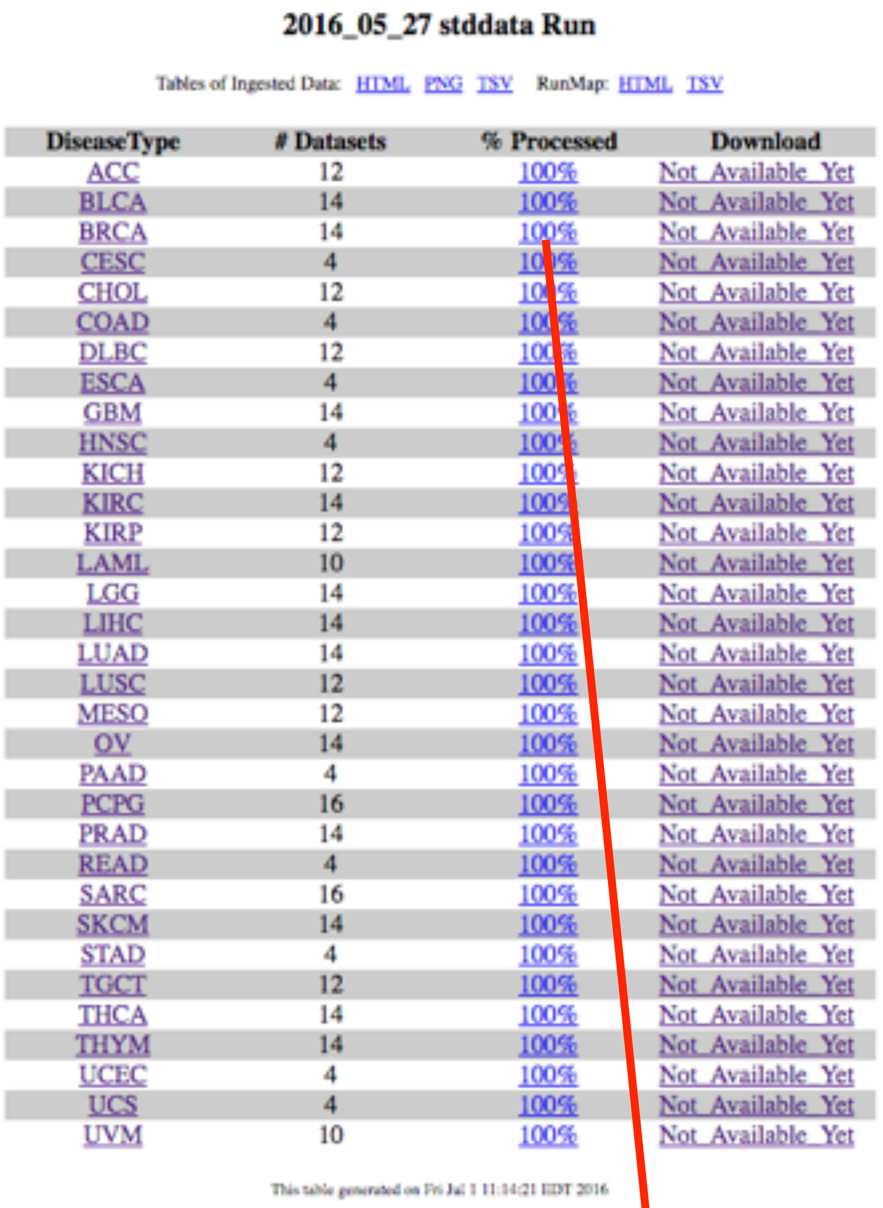
Cohort	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNASeq	miR	miRSeq	RPPA	MAF	rawMAF
ACC	92	92	90	0	80	0	79	0	80	46	90	0
BLCA	412	412	410	112	412	0	408	0	409	344	130	395
BRCA	1098	1097	1089	19	1097	526	1093	0	1078	887	977	0
CESC	307	307	295	50	307	0	304	0	307	173	194	0
CHOL	51	45	36	0	36	0	36	0	36	30	35	0
COAD	460	458	451	69	457	153	457	0	406	360	154	367
COADREAD	631	629	616	104	622	222	623	0	549	491	223	489
DLBC	58	48	48	0	48	0	48	0	47	33	48	0
ESCA	185	185	184	51	185	0	184	0	184	126	185	0
FPPP	38	38	0	0	0	0	0	0	23	0	0	0
GBM	613	595	577	0	420	540	160	565	0	238	290	290
GBMLGG	1129	1110	1090	52	936	567	676	565	512	668	576	806
HNSC	528	528	522	108	528	0	520	0	523	212	279	510
KICH	113	113	66	0	66	0	66	0	66	63	66	66
KIPAN	973	941	883	0	892	88	889	0	873	756	644	799
KIRC	537	537	528	0	535	72	533	0	516	478	417	451
KIRP	323	291	289	0	291	16	290	0	291	215	161	282
LAML	200	200	197	0	194	0	179	0	188	0	197	0
LGG	516	515	513	52	516	27	516	0	512	430	286	516
LIHC	377	377	370	0	377	0	371	0	372	63	198	373
LUAD	585	522	516	120	578	32	515	0	513	365	230	542
LUSC	504	504	501	0	503	154	501	0	478	328	178	0
MESO	87	87	87	0	87	0	87	0	87	63	0	0
OV	602	591	586	0	594	574	304	570	453	426	316	469
PAAD	185	185	184	0	184	0	178	0	178	123	150	184
PCPG	179	179	175	0	179	0	179	0	179	80	179	0
PRAD	499	499	492	115	498	0	497	0	494	352	332	498
READ	171	171	165	35	165	69	166	0	143	131	69	122
SARC	261	261	257	0	261	0	259	0	259	223	247	0
SKCM	470	470	469	118	470	0	469	0	448	353	343	366
STAD	443	443	442	107	443	0	415	0	436	357	289	395
STES	628	628	626	158	628	0	599	0	620	483	474	395
TGCT	150	134	150	0	150	0	150	0	150	118	149	0
THCA	503	503	499	98	503	0	501	0	502	222	402	496
THYM	124	124	123	0	124	0	120	0	124	90	123	0
UCEC	560	548	540	106	547	54	545	0	538	440	248	0
UCS	57	57	56	0	57	0	57	0	56	48	57	0
UVM	80	80	80	51	80	0	80	0	80	12	80	0
Totals	11368	11196	10987	1211	10972	2217	10267	1135	10156	7429	7099	6322

Only data runs so far

Mirror all of TCGA data from GDC
Dice / generate loadfiles

Merge N samples into 1 bolus
(necessary for downstream integrative analysis)

Generate output data archives
Internal reports & dashboards



2016_05_27 stddata Run

Tables of Ingested Data: HTML PNG TSV RunMap: HTML TSV

DiseaseType	# Datasets	% Processed	Download
ACC	12	100%	Not_Available_Yet
BLCA	14	100%	Not_Available_Yet
BRCA	14	100%	Not_Available_Yet
CESC	4	100%	Not_Available_Yet
CHOL	12	100%	Not_Available_Yet
COAD	4	100%	Not_Available_Yet
DLBC	12	100%	Not_Available_Yet
ESCA	4	100%	Not_Available_Yet
GBM	14	100%	Not_Available_Yet
HNSC	4	100%	Not_Available_Yet
KICH	12	100%	Not_Available_Yet
KIRC	14	100%	Not_Available_Yet
KIRP	12	100%	Not_Available_Yet
LAML	10	100%	Not_Available_Yet
LGG	14	100%	Not_Available_Yet
LIHC	14	100%	Not_Available_Yet
LUAD	14	100%	Not_Available_Yet
LUSC	12	100%	Not_Available_Yet
MESO	12	100%	Not_Available_Yet
OV	14	100%	Not_Available_Yet
PAAD	4	100%	Not_Available_Yet
PCPG	16	100%	Not_Available_Yet
PRAD	14	100%	Not_Available_Yet
READ	4	100%	Not_Available_Yet
SARC	16	100%	Not_Available_Yet
SKCM	14	100%	Not_Available_Yet
STAD	4	100%	Not_Available_Yet
TGCT	12	100%	Not_Available_Yet
THCA	14	100%	Not_Available_Yet
THYM	14	100%	Not_Available_Yet
UCEC	4	100%	Not_Available_Yet
UCS	4	100%	Not_Available_Yet
UVM	10	100%	Not_Available_Yet

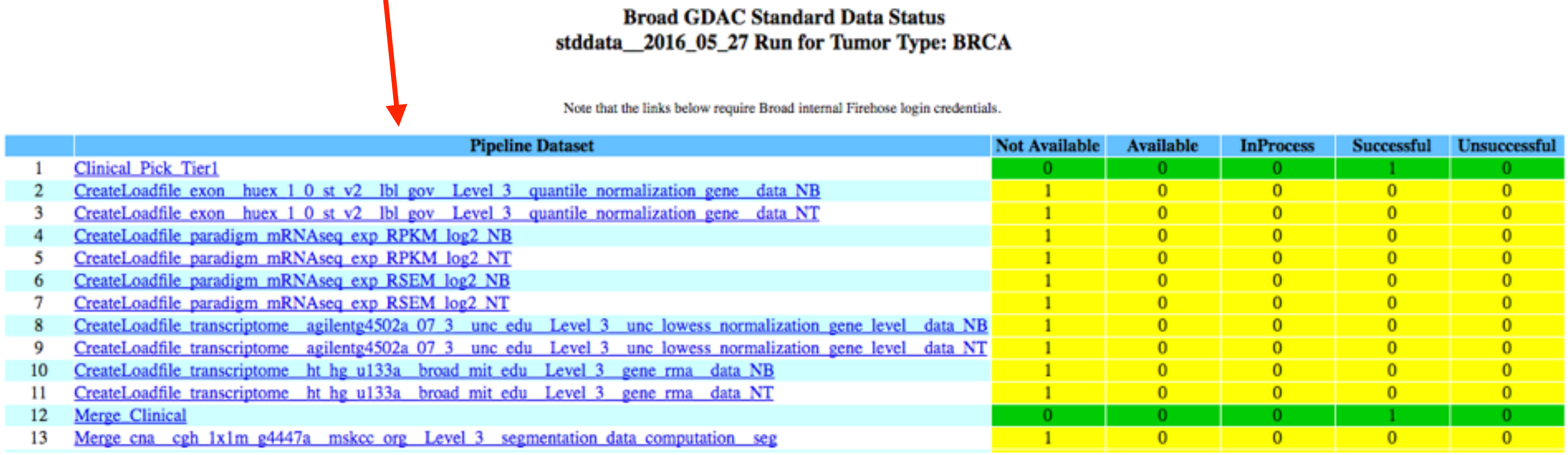
This table generated on Fri Jul 1 11:14:21 EDT 2016

Only data runs so far

Mirror all of TCGA data from GDC
Dice / generate loadfiles

Merge N samples into 1 bolus
(necessary for downstream integrative analysis)

Generate output data archives
Internal reports & dashboards

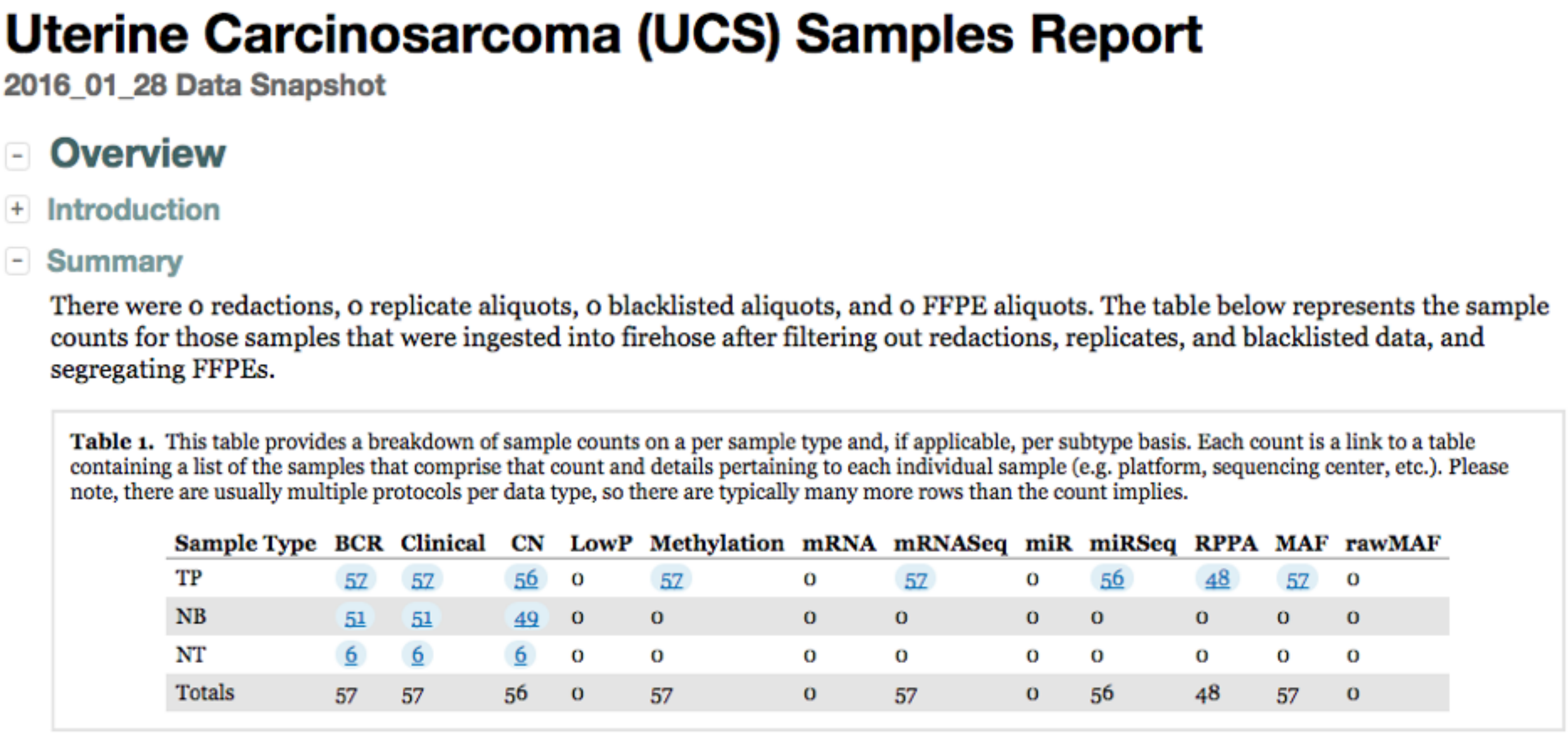


Broad GDAC Standard Data Status
stddata__2016_05_27 Run for Tumor Type: BRCA

Note that the links below require Broad internal Firehose login credentials.

	Pipeline Dataset	Not Available	Available	InProcess	Successful	Unsuccessful
1	Clinical_Pick_Tier1	0	0	0	1	0
2	CreateLoadfile_exon__huex_1_0_st_v2__lbl_gov__Level_3__quantile_normalization_gene__data_NB	1	0	0	0	0
3	CreateLoadfile_exon__huex_1_0_st_v2__lbl_gov__Level_3__quantile_normalization_gene__data_NT	1	0	0	0	0
4	CreateLoadfile_paradigm_mRNAseq_exp_RPKM_log2_NB	1	0	0	0	0
5	CreateLoadfile_paradigm_mRNAseq_exp_RPKM_log2_NT	1	0	0	0	0
6	CreateLoadfile_paradigm_mRNAseq_exp_RSEM_log2_NB	1	0	0	0	0
7	CreateLoadfile_paradigm_mRNAseq_exp_RSEM_log2_NT	1	0	0	0	0
8	CreateLoadfile_transcriptome__agilentg4502a_07_3__unc_edu__Level_3__unc_lowess_normalization_gene_level__data_NB	1	0	0	0	0
9	CreateLoadfile_transcriptome__agilentg4502a_07_3__unc_edu__Level_3__unc_lowess_normalization_gene_level__data_NT	1	0	0	0	0
10	CreateLoadfile_transcriptome__ht_hg_u133a__broad_mit_edu__Level_3__gene_rma__data_NB	1	0	0	0	0
11	CreateLoadfile_transcriptome__ht_hg_u133a__broad_mit_edu__Level_3__gene_rma__data_NT	1	0	0	0	0
12	Merge_Clinical	0	0	0	1	0
13	Merge_cna__cgh_1x1m_g4447a__mskcc_org__Level_3__segmentation_data_computation__seg	1	0	0	0	0

Full Sample Reports Available ~ 1 week



Uterine Carcinosarcoma (UCS) Samples Report

2016_01_28 Data Snapshot

- Overview

+ Introduction

- Summary

There were 0 redactions, 0 replicate aliquots, 0 blacklisted aliquots, and 0 FFPE aliquots. The table below represents the sample counts for those samples that were ingested into firehose after filtering out redactions, replicates, and blacklisted data, and segregating FFPEs.

Table 1. This table provides a breakdown of sample counts on a per sample type and, if applicable, per subtype basis. Each count is a link to a table containing a list of the samples that comprise that count and details pertaining to each individual sample (e.g. platform, sequencing center, etc.). Please note, there are usually multiple protocols per data type, so there are typically many more rows than the count implies.

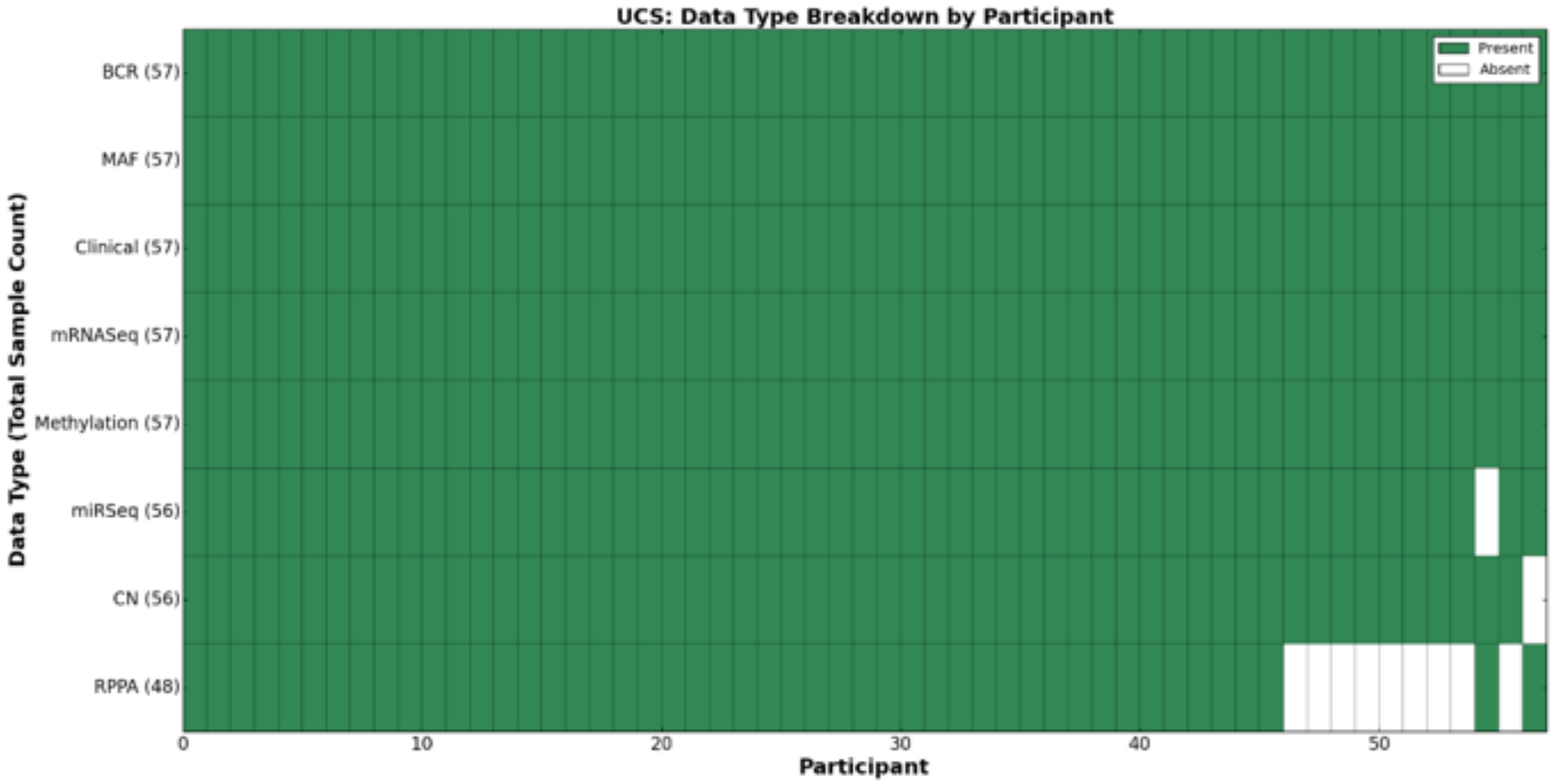
Sample Type	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNASeq	miR	miRSeq	RPPA	MAF	rawMAF
TP	57	57	56	0	57	0	57	0	56	48	57	0
NB	51	51	49	0	0	0	0	0	0	0	0	0
NT	6	6	6	0	0	0	0	0	0	0	0	0
Totals	57	57	56	0	57	0	57	0	56	48	57	0

Users find these very helpful, by showing:

Counts broken down by tissue and data type

Redactions and other filtered samples (replicate aliquots, FFPEs)

Provenance of every aliquot in run (all the way back to submitting center)



And sample heatmaps

What about TARGET?

We have no data or analysis pipelines for TARGET

But GDCtools aims to be flexible & easily configured to mirror any PROGRAM or PROJECT from GDC

We verified this morning by initiating TARGET mirror

Remaining Work

After establishing full confidence in sample counts
(we're about 95% confident right now)

1. Perform production data run, with reports (2-3 weeks)
2. Then kick off an GDC-only analysis run (Aug)
3. Finish phase 1 of Python bindings in *GDCtools*
4. Document *GDCtools* repo & initial release (Sept)

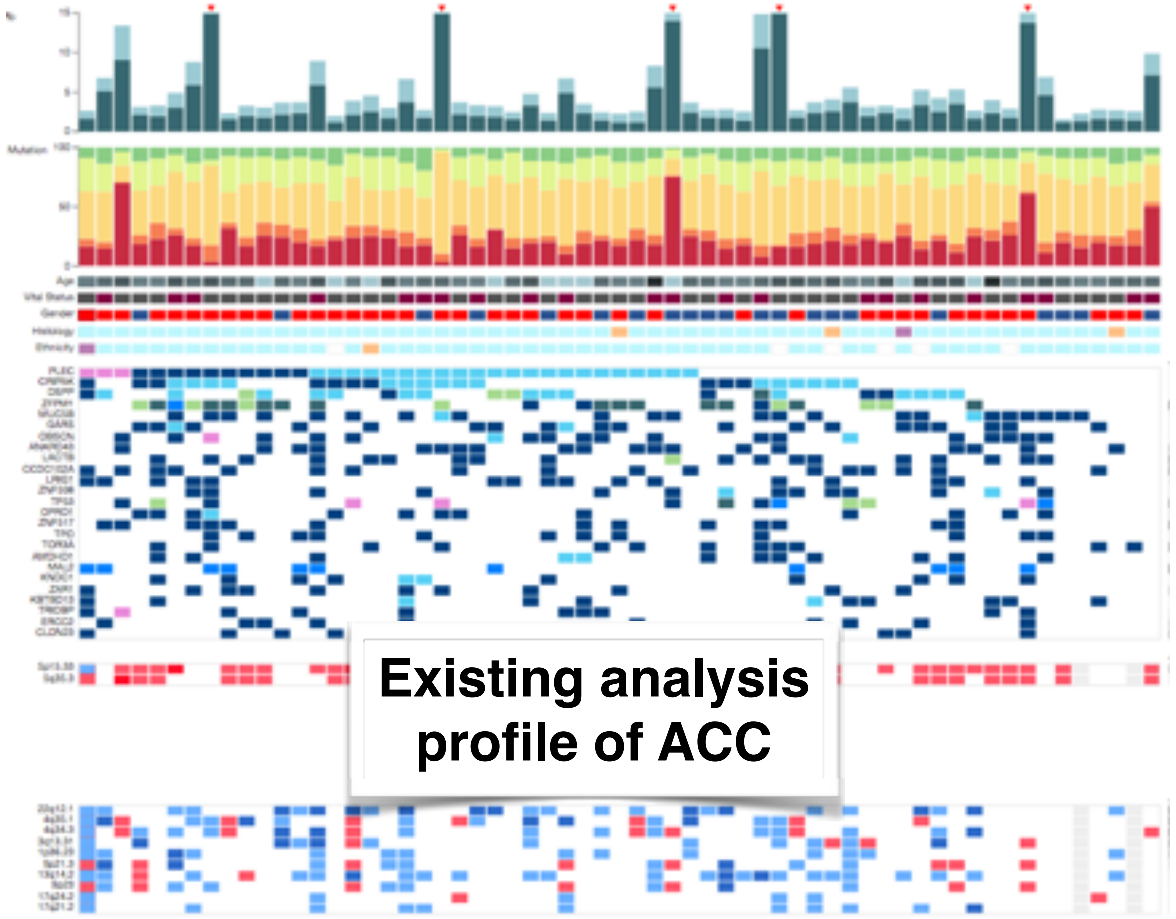
Public Release Criteria

Data Runs

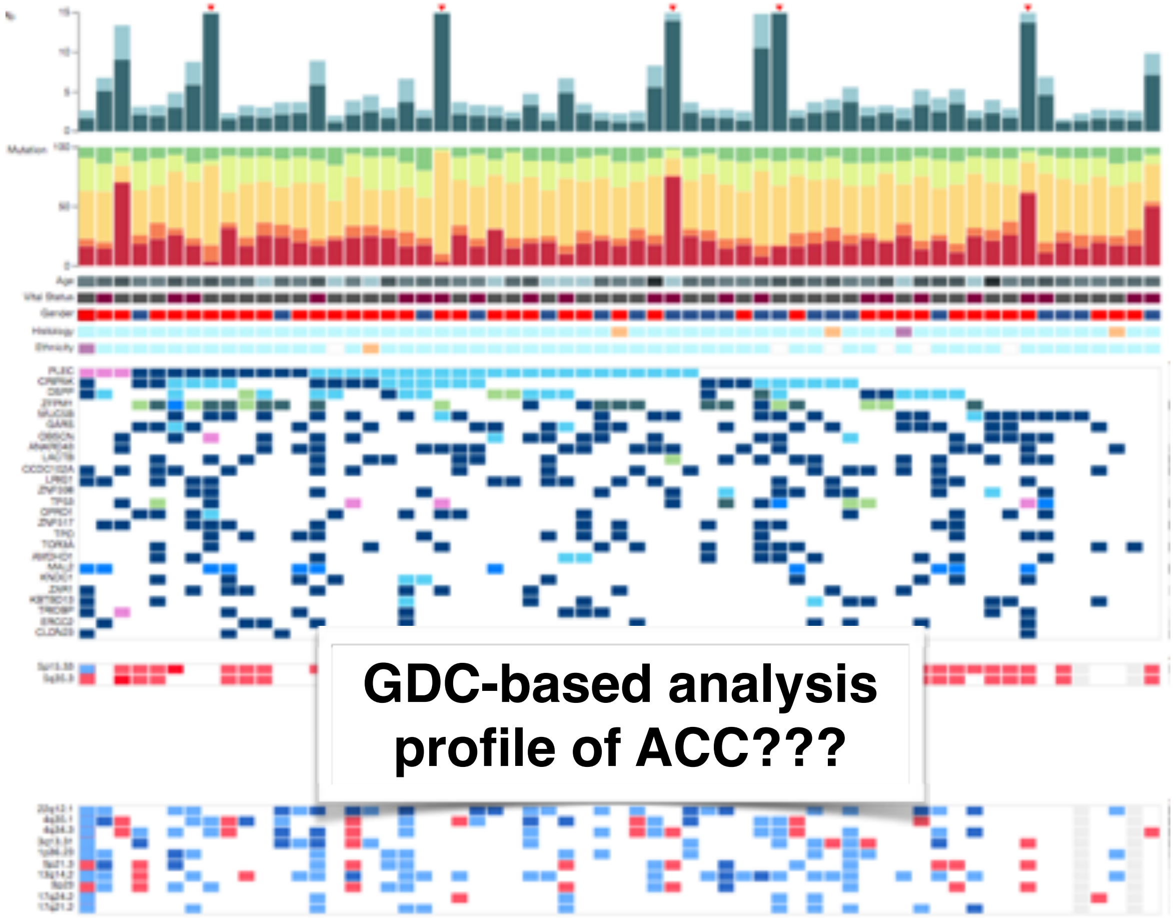
When our sample counts of DCC vs GDC are plausible (effectively identical)

Analysis Runs

Establish concordance of analysis results from GDC and DCC data by visual diff with iCoMut



Existing analysis profile of ACC



GDC-based analysis profile of ACC???

http://firebrowse.org/iCoMut/?cohort=ACC

GDAN readiness, GDC and the 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 vendor-neutral 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