Firehose: A Case Study in Big Data Genomics

MIT Lincoln Laboratory
October 4, 2013

Michael S. Noble

Assistant Director for Data Science
Cancer Genome Analysis Group
The Broad Institute of MIT & Harvard

Senior Manager, TCGA Genome Data Analysis Center

Acknowledgements

PI: Lynda Chin, Gaddy Getz

Broad Institute
Douglas Voet
Daniel DiCara
Gordon Saksena
Hailei Zhang
David Heiman
Juok Cho
William Mallard
Harindra Arichchi
Michael Lawrence
Petar Stojanov
Lihua Zou
Chip Stewart
Scott Frazer
Pei Lin
Kristian Cibulskis
Jaegil Kim
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

THE CANCER GENOME ATLAS

THE UNIVERSITY OF TEXAS MD Anderson Cancer Center Making Cancer History®

Outline

1. Why Firehose?
2. What we produce
3. Recent highlights
4. Observations

Outline

1. Why Firehose?
2. What we produce
3. Recent highlights
4. Observations

"Big Data Science Through Software"
More than science or software itself

1. Why?

FIREHOSE
Broad GDAC

Born of the desire to systematize analyses from The Cancer Genome Atlas pilot and scale their execution to the dozens of remaining diseases to be studied, now sits atop ~30 terabytes of TCGA data and reliably executes more than 2300 pipelines per month.

FIREHOSE
Broad GDAC

Born of the desire to systematize analyses from The Cancer Genome Atlas pilot and scale their execution to the dozens of remaining diseases to be studied, now sits atop ~~30~~ 40 terabytes of TCGA data and reliably executes more than ~~2300~~ 6000 pipelines per month.

Because The Bad Old Days ...

Because The Bad Old Days ...

Of solitary, manual experimentation on small sample sets ...

```
% create a folder
% download data.from.some.where
% run_your_computational_analysis
```

Because The Bad Old Days ...

Of solitary, manual experimentation on small sample sets ...

```
% create a folder
% download data.from.some.where
% run_your_computational_analysis
```

Then get distracted, do it again ...
Forget, search ... lose track, search ...
Repeat ... for 20 more disease types
GBM, LUNG, AML, ...

Because The Bad Old Days ...

Of solitary, manual experimentation on small sample sets ...

```
% create a folder
% download data.from.some.where
% run_your_computational_analysis
```

Then get distracted, do it again ...
Forget, search ... lose track, search ...
Repeat ... for 20 more disease types
GBM, LUNG, AML, ...

Then multiply by 5, 10 ... researchers at your site

Don’t Scale to TCGA

November 14, 2012
Firehose Data Snapshot

Tumor	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BLCA	153	108	99	0	138	0	96	0	124	54	28
BRCA	914	866	874	0	889	529	805	0	868	408	507
CESC	122	32	102	0	122	0	0	0	122	0	36
COAD	423	423	413	69	420	155	192	0	407	269	155
COADREAD	592	591	575	104	582	224	264	0	550	399	224
DLBC	28	0	17	0	17	0	0	0	16	0	0
GBM	598	565	563	0	411	542	161	491	0	214	276
HNSC	328	315	294	96	310	0	303	0	309	212	0
KICH	66	0	65	0	65	0	0	0	0	0	0
KIRC	502	502	493	0	500	72	469	0	480	454	403
KIRP	149	103	103	0	103	16	63	0	103	0	0
LAML	202	200	0	0	194	0	179	0	187	0	199
LGG	222	198	180	0	176	27	110	0	180	0	0
LIHC	99	62	97	0	98	0	17	0	96	0	0
LUAD	439	294	356	0	430	32	353	0	365	237	229
LUSC	376	327	343	0	359	154	223	0	332	195	178
OV	592	580	566	0	557	575	297	570	454	412	316
PAAD	57	0	48	0	40	0	0	0	34	0	0
PANCAN8	4086	3882	3907	210	3798	2150	2515	1061	3169	2282	2152
PRAD	180	127	171	0	172	0	140	0	170	0	83
READ	169	168	162	35	162	69	72	0	143	130	69
SARC	29	0	29	0	29	0	0	0	29	0	0
SKCM	273	138	253	101	253	0	247	0	240	164	0
STAD	238	162	144	0	145	0	43	0	134	0	116
THCA	435	218	330	94	353	0	254	0	349	224	323
UCEC	512	451	493	106	500	54	333	0	485	200	248
Totals	7106	5839	6195	501	6443	2225	4357	1061	5627	3173	3166

Don’t Scale to TCGA

November 14, 2012
Firehose Data Snapshot

Tumor	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BLCA	153	108	99	0	138	0	96	0	124	54	28
BRCA	914	866	874	0	889	529	805	0	868	408	507
CESC	122	32	102	0	122	0	0	0	122	0	36
COAD	423	423	413	69	420	155	192	0	407	269	155
COADREAD	592	591	575	104	582	224	264	0	550	399	224
DLBC	28	0	17	0	17	0	0	0	16	0	0
GBM	598	565	563	0	411	542	161	491	0	214	276
HNSC	328	315	294	96	310	0	303	0	309	212	0
KICH	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0	0	0	0	0
KIRC	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	72	469	0	480	454	403
KIRP	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	16	63	0	103	0	0
LAML	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	179	0	187	0	199
LGG	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	27	110	0	180	0	0
LIHC	99	62	97	0	98	0	17	0	96	0	0
LUAD	439	294	356	0	430	32	353	0	365	237	229
LUSC	376	327	343	0	359	154	223	0	332	195	178
OV	592	580	566	0	557	575	297	570	454	412	316
PAAD	57	0	48	0	40	0	0	0	34	0	0
PANCAN8	4086	3882	3907	210	3798	2150	2515	1061	3169	2282	2152
PRAD	180	127	171	0	172	0	140	0	170	0	83
READ	169	168	162	35	162	69	72	0	143	130	69
SARC	29	0	29	0	29	0	0	0	29	0	0
SKCM	273	138	253	101	253	0	247	0	240	164	0
STAD	238	162	144	0	145	0	43	0	134	0	116
THCA	435	218	330	94	353	0	254	0	349	224	323
UCEC	512	451	493	106	500	54	333	0	485	200	248
Totals	7106	5839	6195	501	6443	2225	4357	1061	5627	3173	3166
	+1830	+1665	+2021		+4181						+1142

Diffs since Nov 2011
(~11K samples)

Don't Scale to TCGA

November 14, 2012
Firehose Data Snapshot

Tumor	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BLCA	153	108	99	0	138	0	96	0	124	54	28
BRCA	914	866	874	0	889	529	805	0	868	408	507
CESC	122	32	102	0	122	0	0	0	122	0	36
COAD	423	423	413	69	420	155	192	0	407	269	155
COADREAD	592	591	575	104	582	224	264	0	550	399	224
DLBC	28	0	17	0	17	0	0	0	16	0	0
GBM	598	565	563	0	411	542	161	491	0	214	276
HNSC	328	315	294	96	310	0	303	0	309	212	0
KICH						0					0
KIRC						72					403
KIRP						16					0
LAML						0					199
LGG						27					0
LIHC	99	62	97	0	98	0	17	0	96	0	0
LUAD	439	294	356	0	430	32	353	0	365	237	229
LUSC	376	327	343	0	359	154	223	0	332	195	178
OV	592	580	566	0	557	575	297	570	454	412	316
PAAD	57	0	48	0	40	0	0	0	34	0	0
PANCAN8	4086	3882	3907	210	3798	2150	2515	1061	3169	2282	2152
PRAD	180	127	171	0	172	0	140	0	170	0	83
READ	169	168	162	35	162	69	72	0	143	130	69
SARC	29	0	29	0	29	0	0	0	29	0	0
SKCM	273	138	253	101	253	0	247	0	240	164	0
STAD	238	162	144	0	145	0	43	0	134	0	116
THCA	435	218	330	94	353	0	254	0	349	224	323
UCEC	512	451	493	106	500	54	333	0	485	200	248
Totals	7106	5839	6195	501	6443	2225	4357	1061	5627	3173	3166
	+1830	+1665	+2021	+501	+4181		+4357		+5267	+3173	+1142

Diffs since Nov 2011 (~11K samples)

New data types (12.5K samples)

Don't Scale to TCGA

November 14, 2012
Firehose Data Snapshot

Tumor	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BLCA	153	108	99	0	138	0	96	0	124	54	28
BRCA	914	866	874	0	889	529	805	0	868	408	507
CESC	122	32	102	0	122	0	0	0	122	0	36
COAD	423	423	413	69	420	155	192	0	407	269	155
COADREAD	592	591	575	104	582	224	264	0	550	399	224
DLBC	28	0	17	0	17	0	0	0	16	0	0
GBM	598	565	563	0	411	542	161	491	0	214	276
HNSC	328	315	294	96	310	0	303	0	309	212	0
KICH						0					0
KIRC						72					403
KIRP						16					0
LAML						0					199
LGG						27					0
LIHC	99	62	97	0	98	0	17	0	96	0	0
LUAD	439	294	356	0	430	32	353	0	365	237	229
LUSC	376	327	343	0	359	154	223	0	332	195	178
OV	592	580							454	412	316
PAAD	57	0							34	0	0
PANCAN8	4086	3882							3169	2282	2152
PRAD	180	127							170	0	83
READ	169	168							143	130	69
SARC	29	0	29	0	29	0	0	0	29	0	0
SKCM	273	138	253	101	253	0	247	0	240	164	0
STAD	238	162	144	0	145	0	43	0	134	0	116
THCA	435	218	330	94	353	0	254	0	349	224	323
UCEC	512	451	493	106	500	54	333	0	485	200	248
Totals	7106	5839	6195	501	6443	2225	4357	1061	5627	3173	3166
	+1830	+1665	+2021	+501	+4181		+4357		+5267	+3173	+1142

Diffs since Nov 2011 (~11K samples)

New data types (12.5K samples)

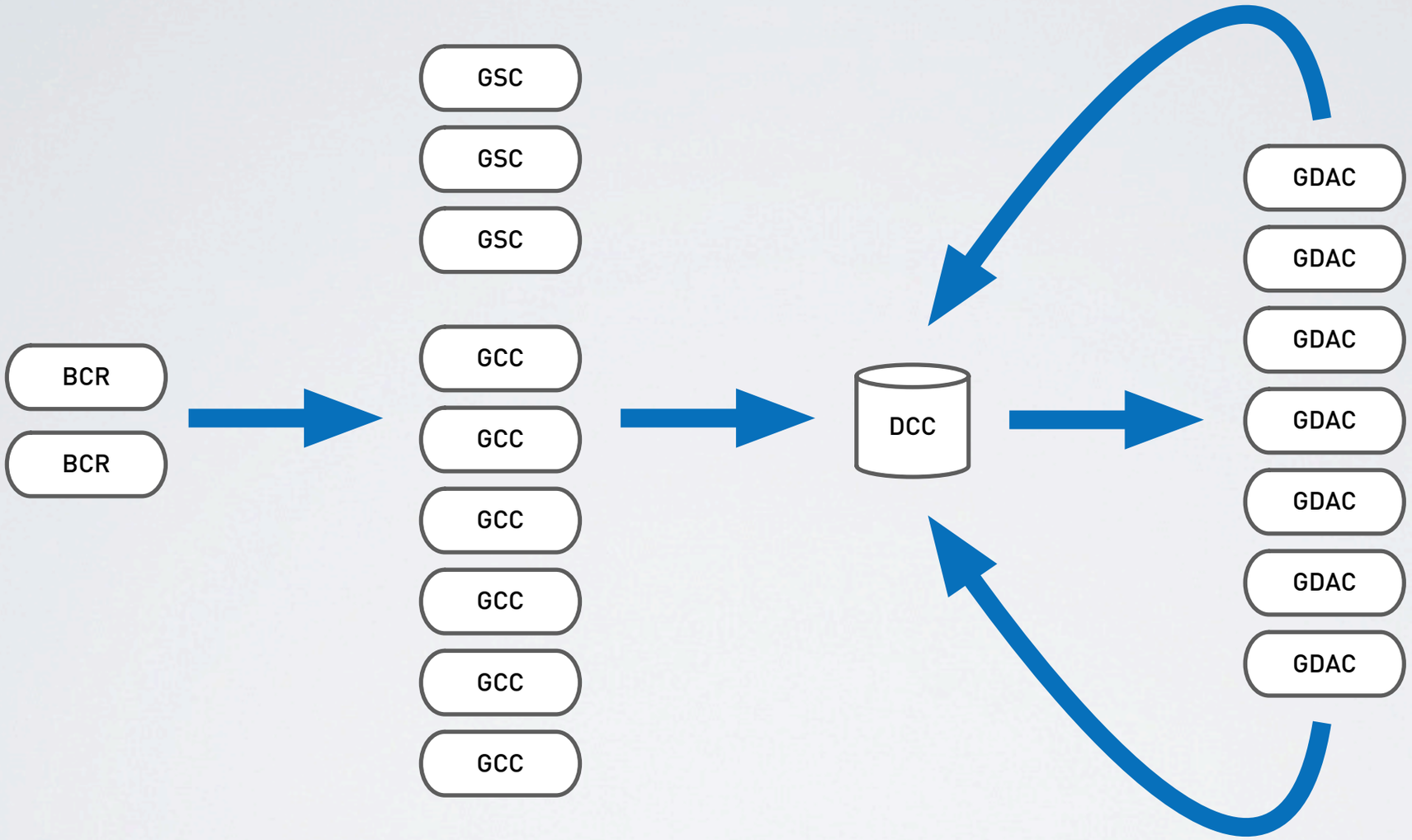
Single year: ~24K new sample aliquots

Context : 2-3 orders magnitude shift

Exome Sequencing Studies of Cancer in 2011			
Cancer Type	**#Samples**	**Key Finding(s)**	**Publication**
Melanoma	14 cases	Frequent mutations in GRIN2A	Wei et al. Nat Genet. 2011.
Metastatic Melanoma	8 cell lines	Mutations in MAP3K5 and MAP3K9	Stark et al. Nat Genet. 2011
Melanoma	7 cell lines	Recurring somatic MAP2K1 and MAP2K2 mutations (8%)	Nikolaev et al. Nat Genet. 2011
Head and neck squamous cell	74 cases	Mutations in TP53, CDKN2A, PIK3CA, HRAS, and squamous differentiation genes.	Stransky et al. Science.
Head and neck squamous cell	32 cases	Mutations in TP53, CDKN2A, PIK3CA, and HRAS, FBXW7 and NOTCH1. Tumor-suppressor role for NOTCH1.	Agrawal et al. Science 2011.
Renal carcinoma	7 cases	Frequent mutation of the SWI/SNF complex gene PBRM1	Varela et al. Nature 2011.
Pancreatic cancer	15 cell lines	Genomic instability caused by MLH1 haploinsufficiency and complete deficiency	Wang et al. Genome Res. 2011
Pancreatic neoplastic cysts	8 cyst resections	Recurrent mutations in components of ubiquitin-dependent pathways	Wu et al. PNAS 2011.
Gastric cancer	22 cases	Frequent mutation of ARID1A	Wang et al. Nat Genet 2011.
Prostate cancer	3 primaries 16 metastases	Recurrent alterations in TP53, DLK2, GPC6, and SDF4	Kumar et al. PNAS 2011

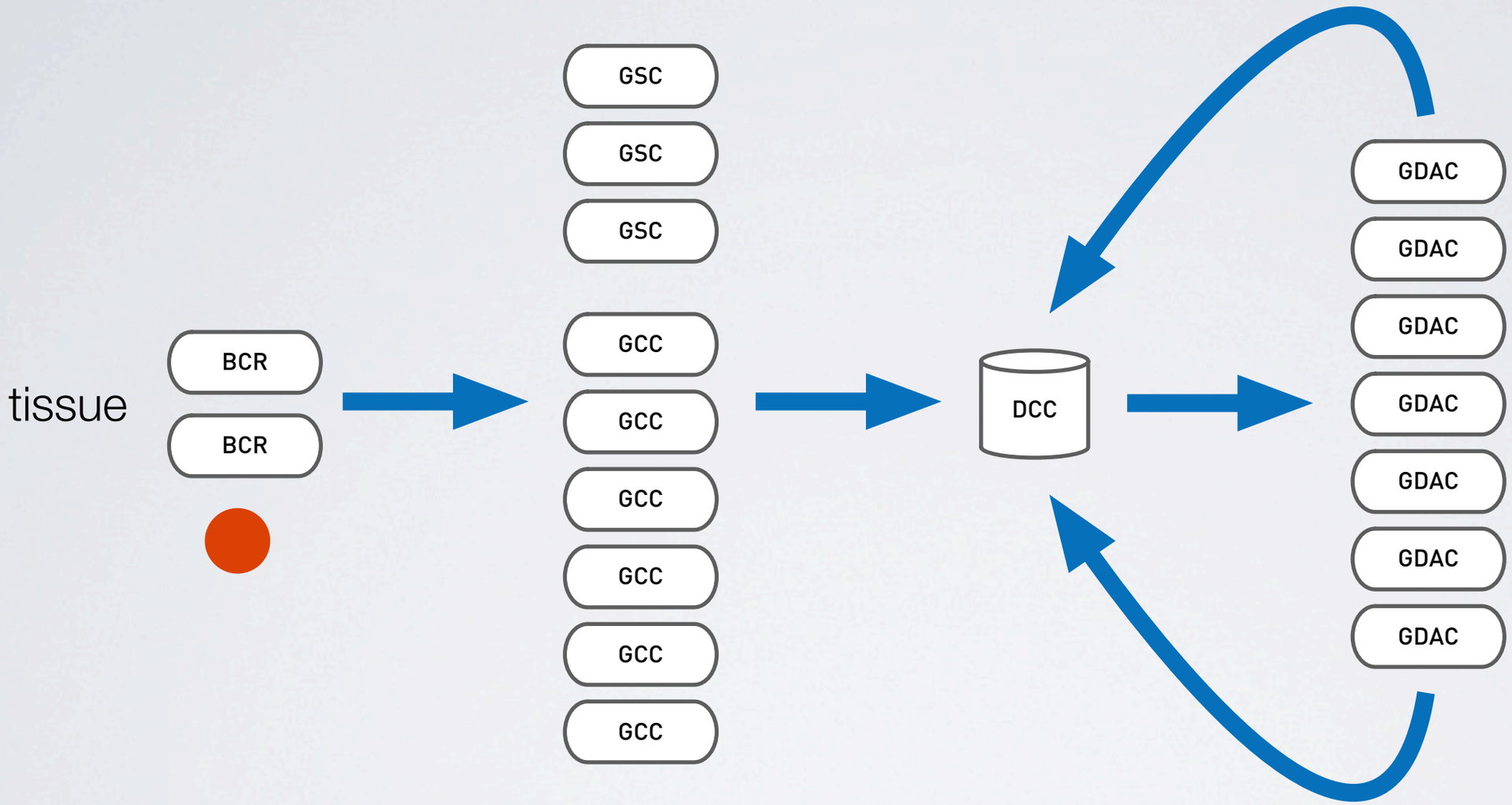
http://massgenomics.org/2012/01/cancer-genome-and-exome-sequencing-in-2011.html

Life Cycle of A TCGA Sample



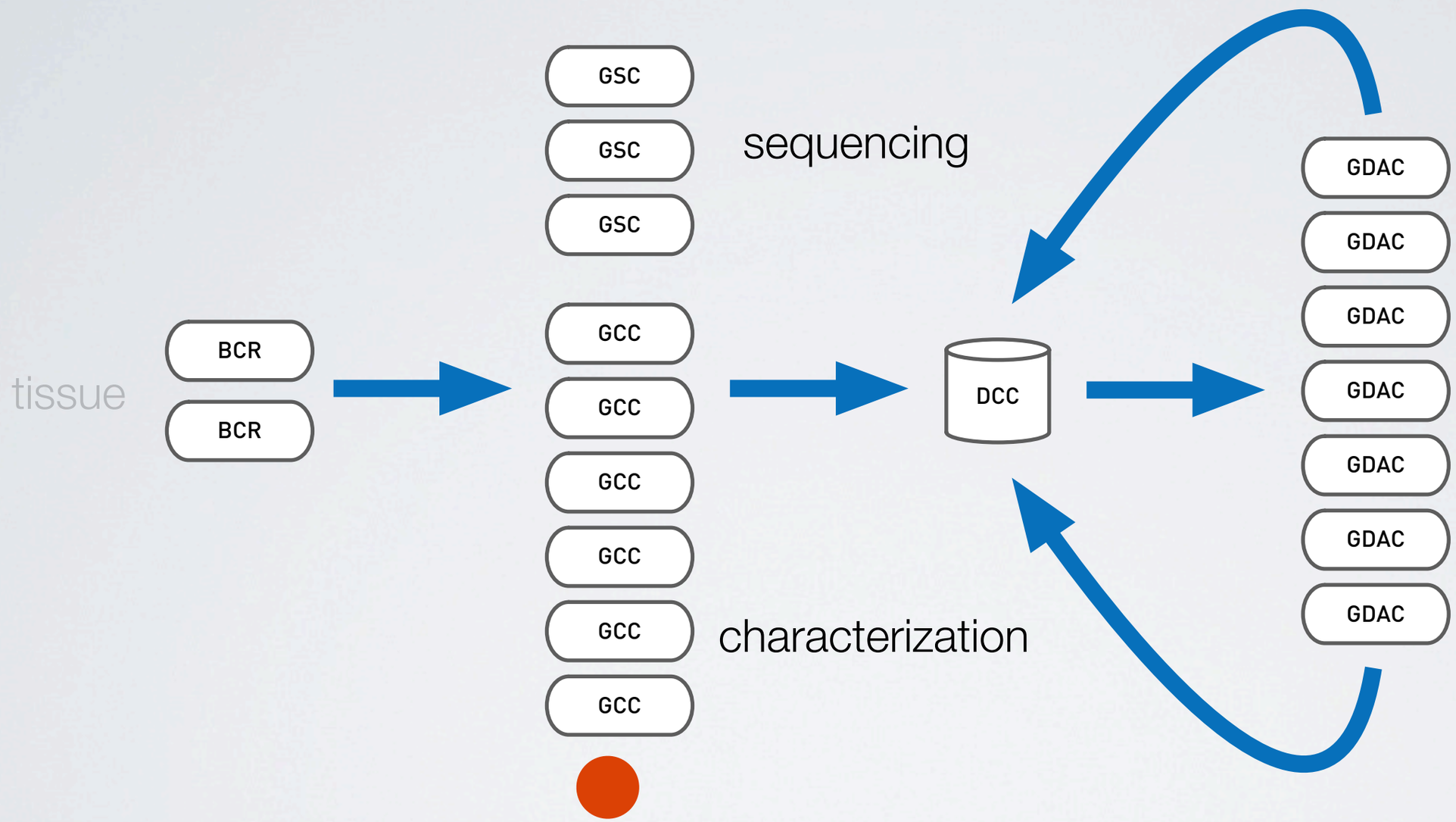
20 centers + Program Office

Life Cycle of A TCGA Sample



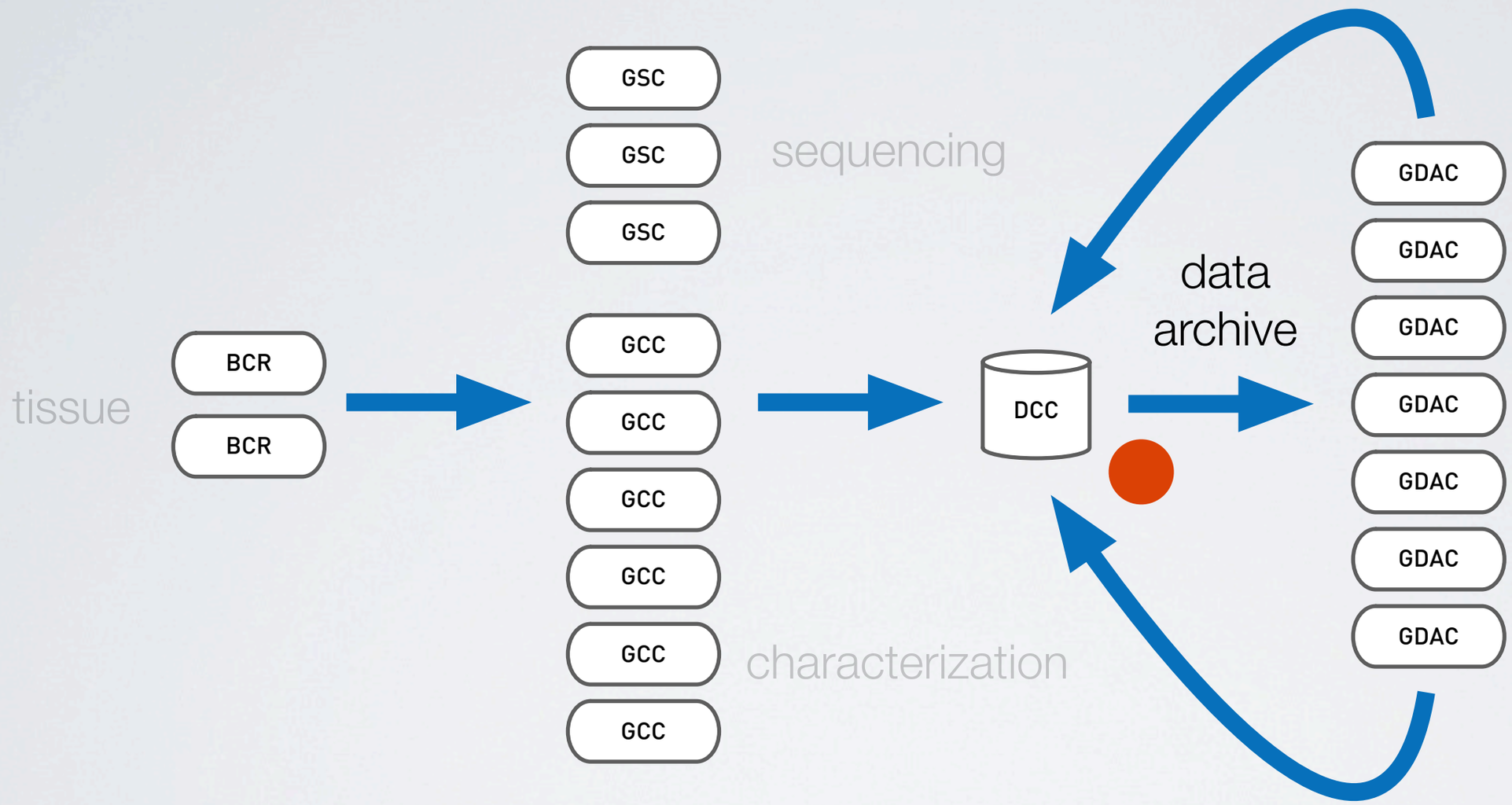
20 centers + Program Office

LIFE CYCLE OF A TCGA SAMPLE



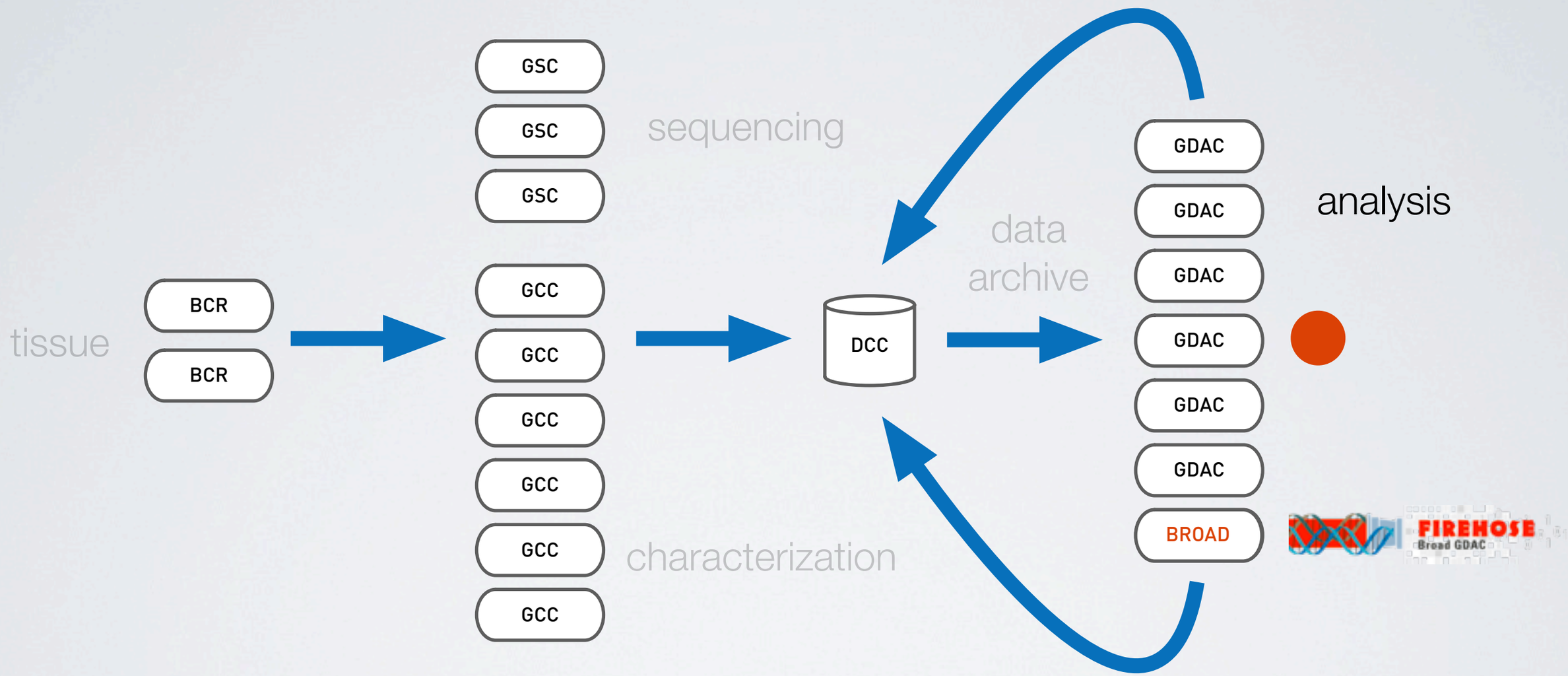
20 CENTERS + PROGRAM OFFICE

LIFE CYCLE OF A TCGA SAMPLE



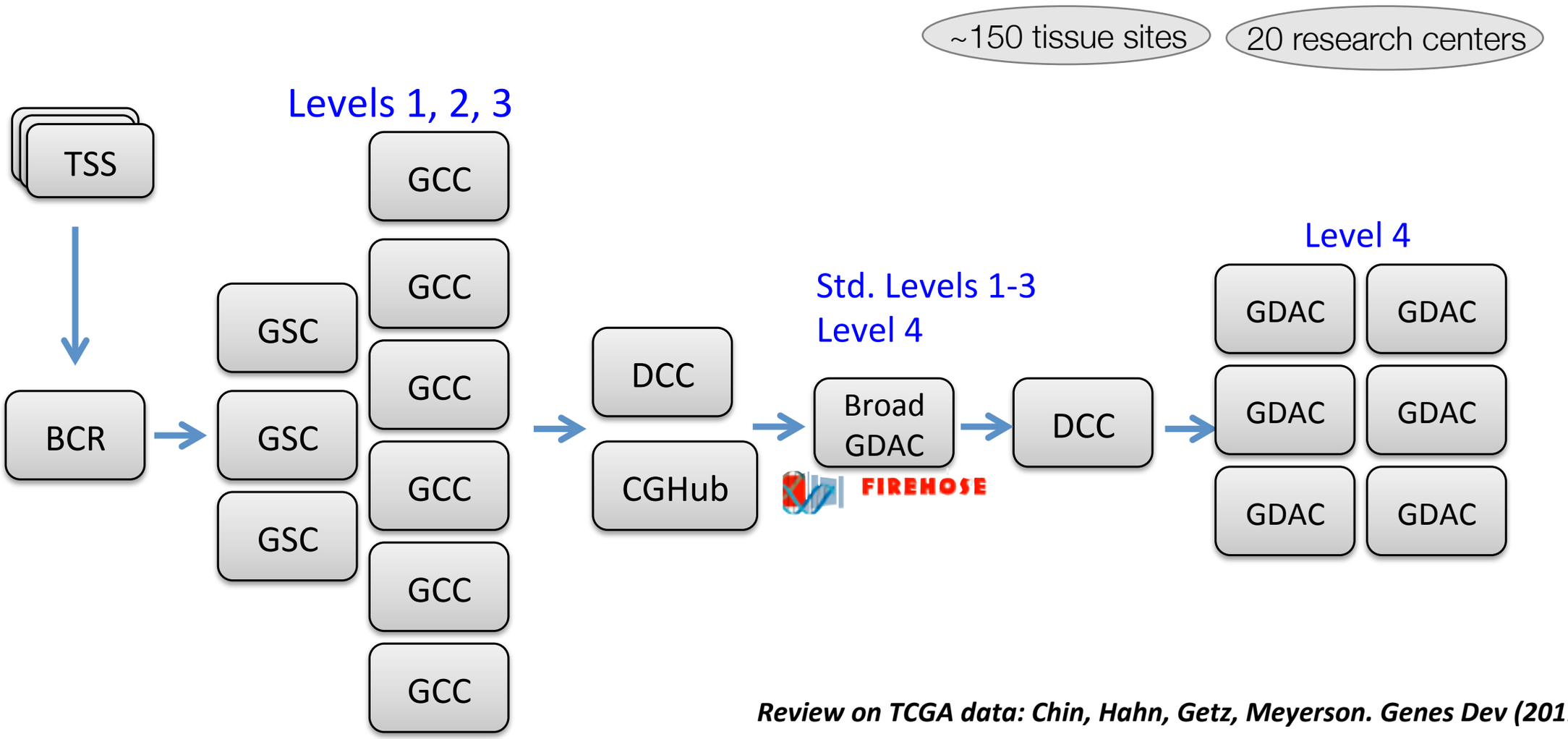
20 CENTERS + PROGRAM OFFICE

LIFE CYCLE OF A TCGA SAMPLE



20 CENTERS + PROGRAM OFFICE

Understanding TCGA : data flow & levels



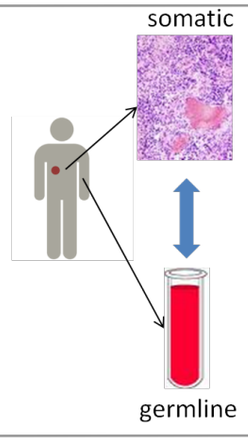
Review on TCGA data: Chin, Hahn, Getz, Meyerson. Genes Dev (2011)

Characterization: (individuals)

Level 1 – Raw data (e.g. raw reads and qualities, Affymetrix CEL files)

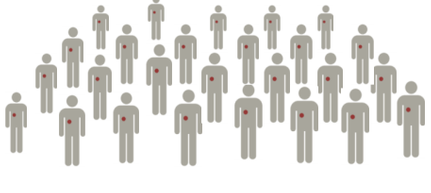
Level 2 – Normalized data (e.g. aligned reads – BAM files, intensity matched files)

Level 3 – Genomic events (e.g. somatic mutations, segments of copy number changes)

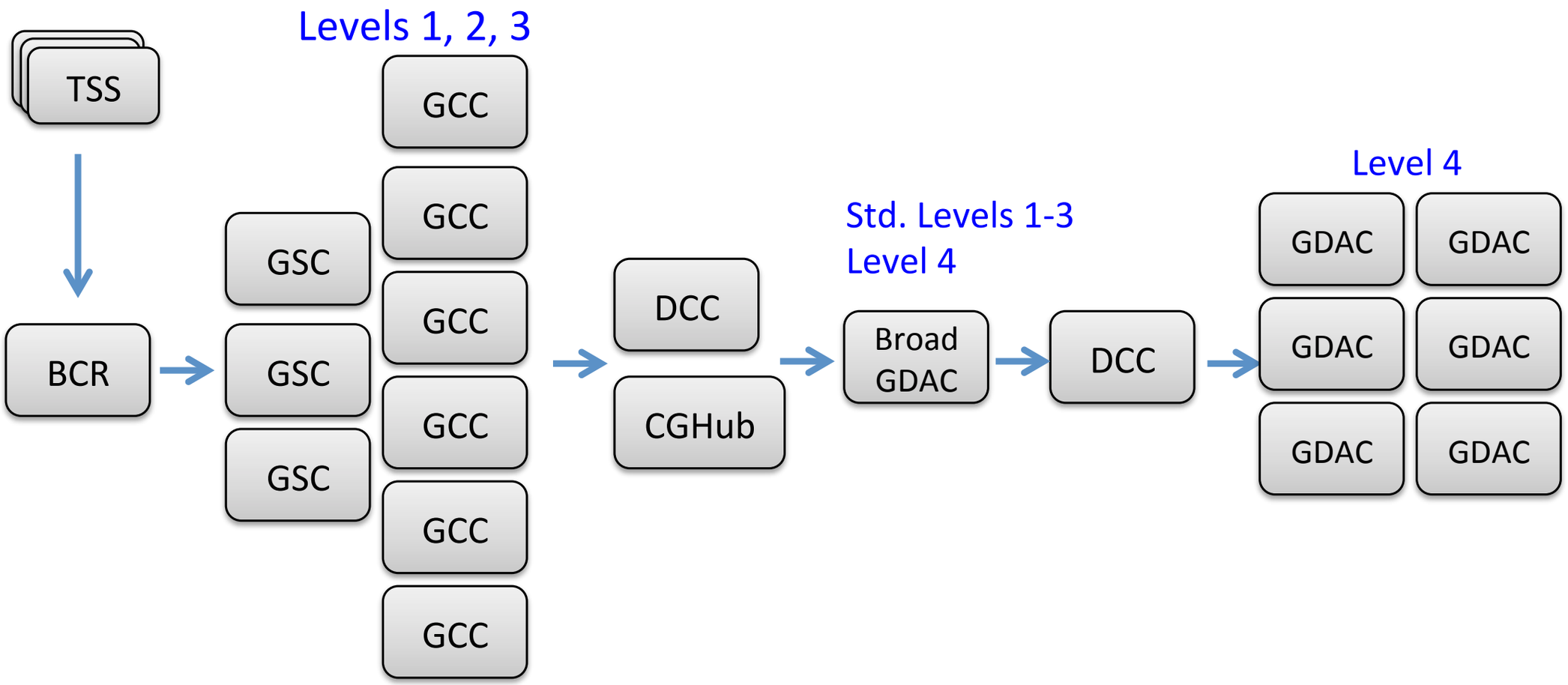


Interpretation: (populations)

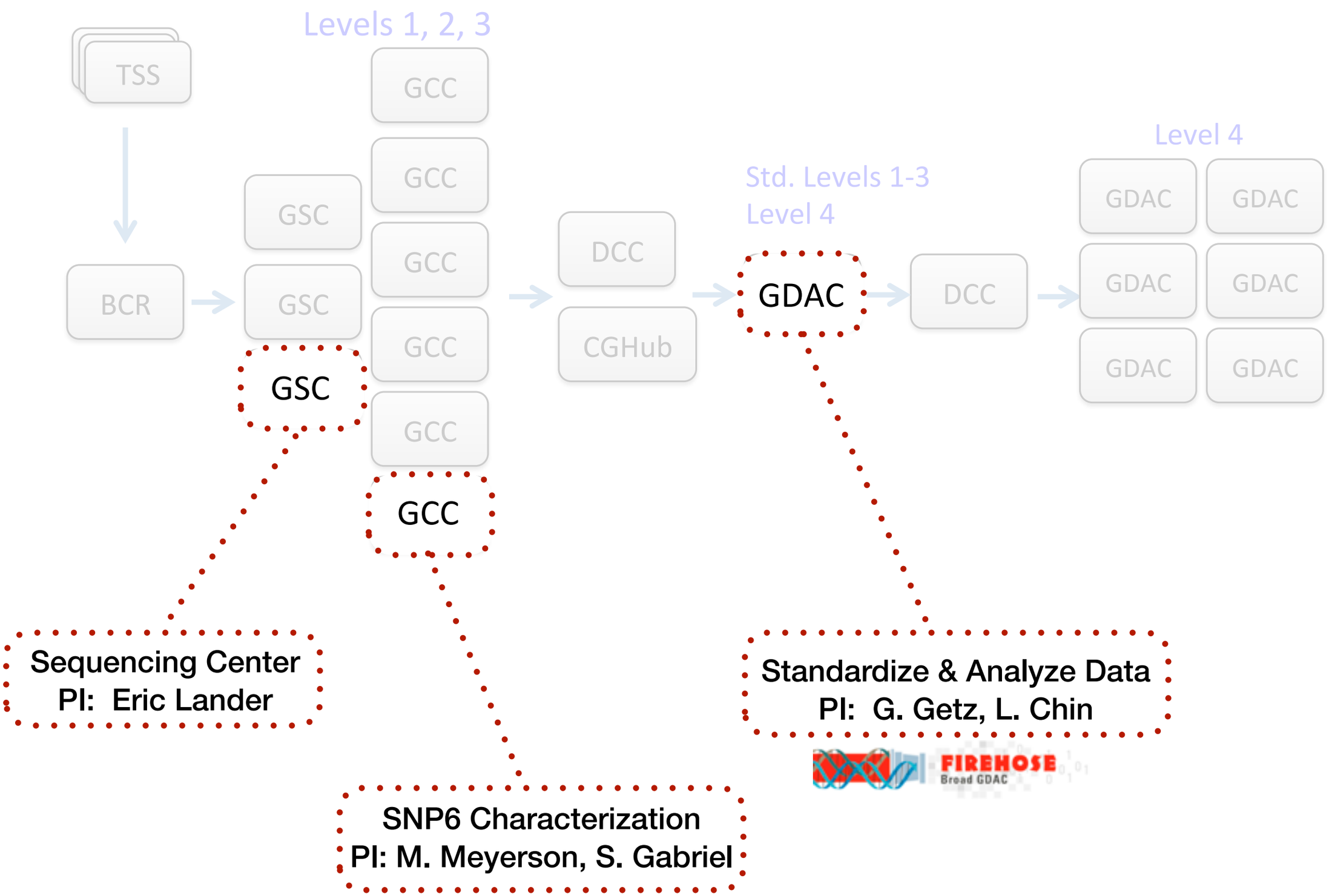
Level 4 – Analysis across a cohort (e.g. sub-types discovery, correlate data types, significantly mutated genes/regions/pathways, correlation to clinical parameters)



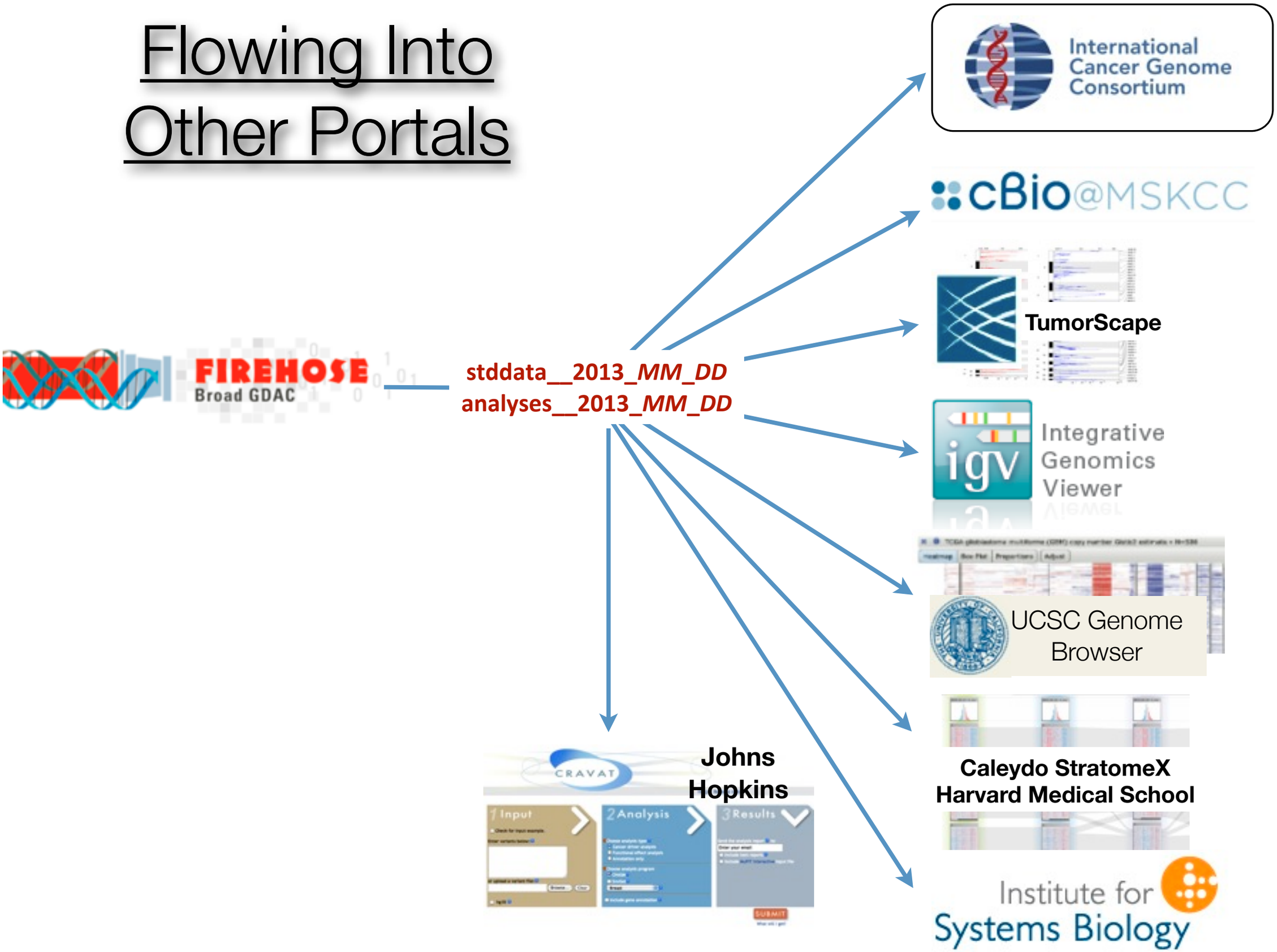
Broad Roles: 3 of 20 TCGA centers



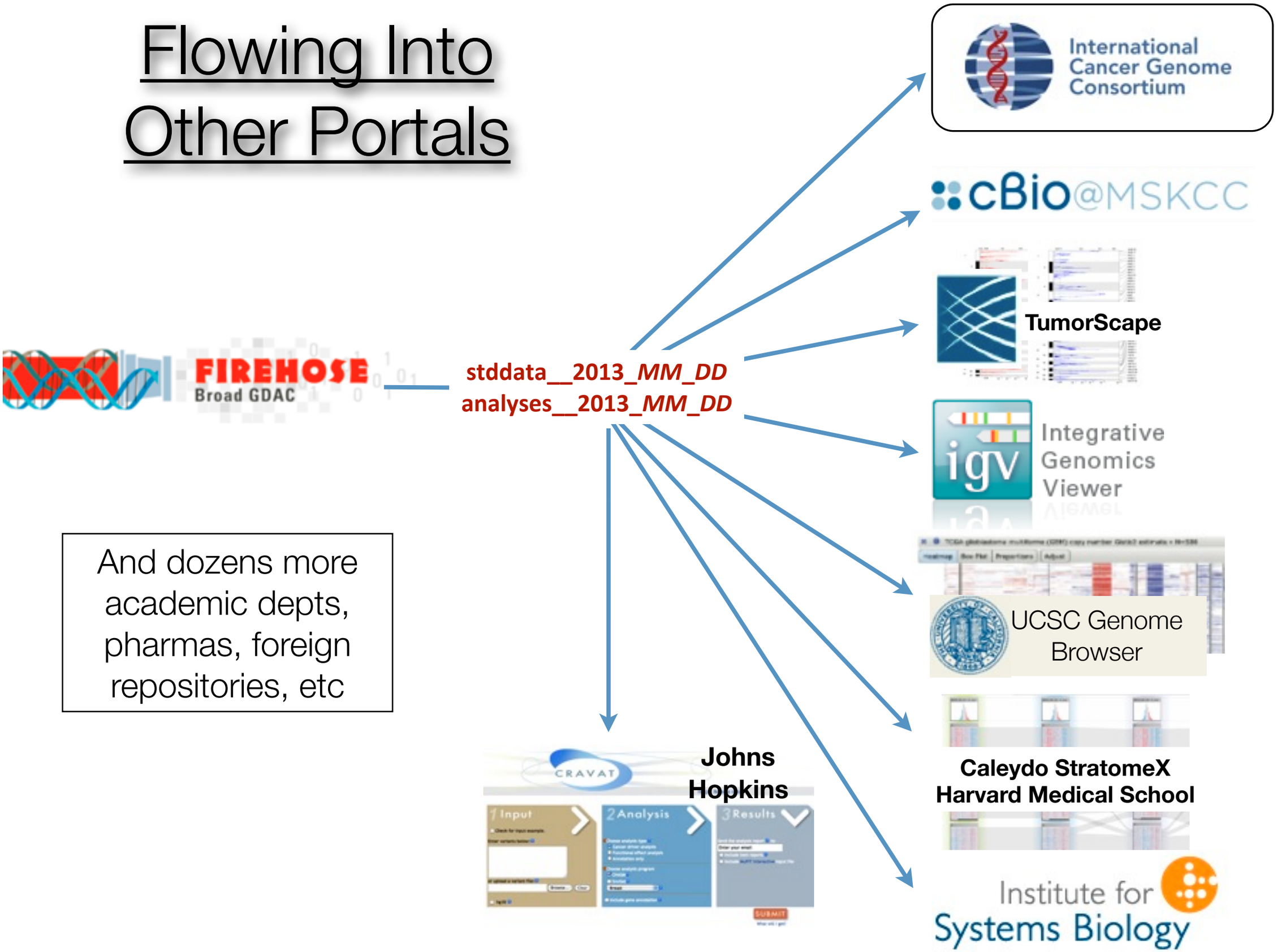
Broad Roles: 3 of 20 TCGA centers



Flowing Into Other Portals

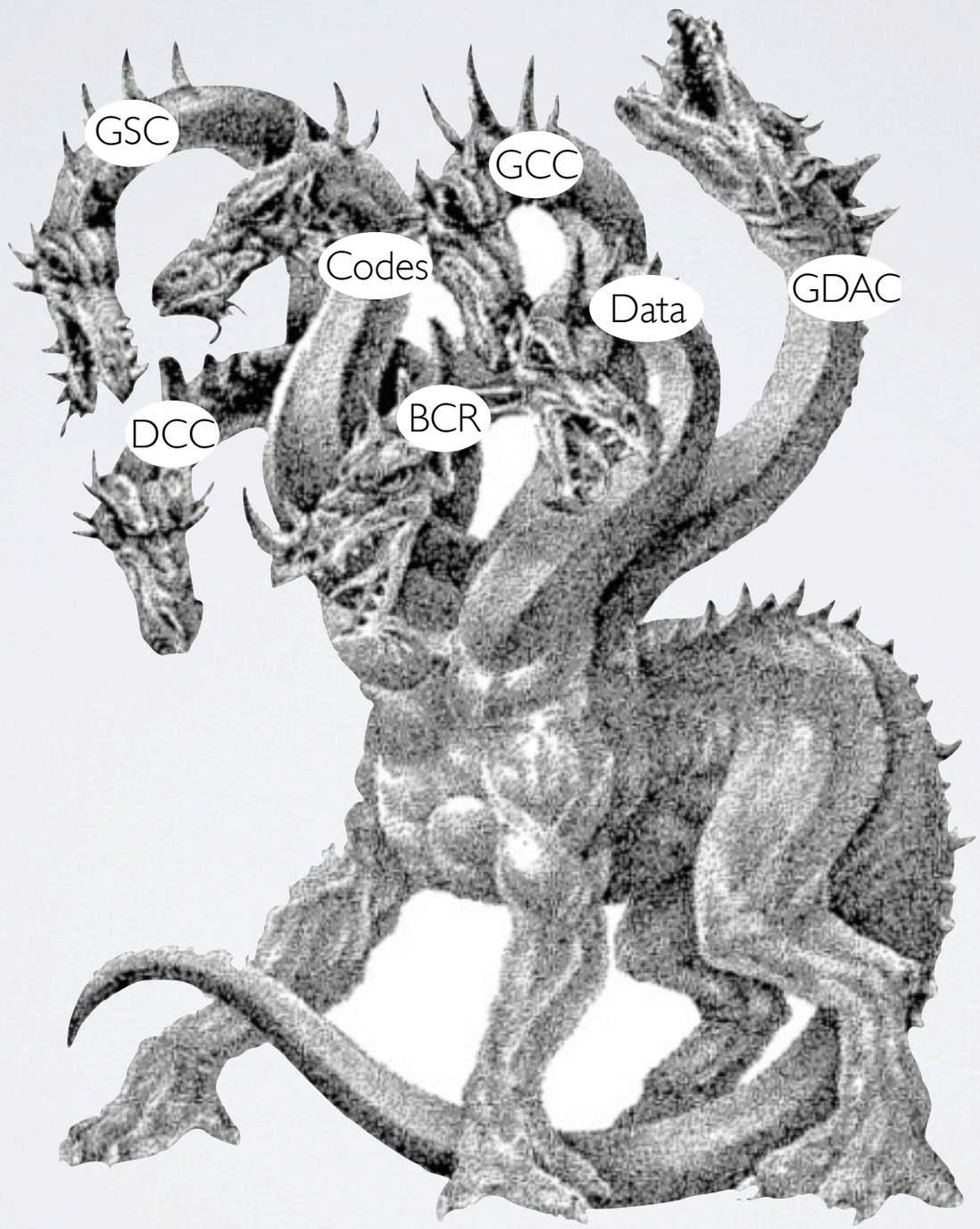


Flowing Into Other Portals

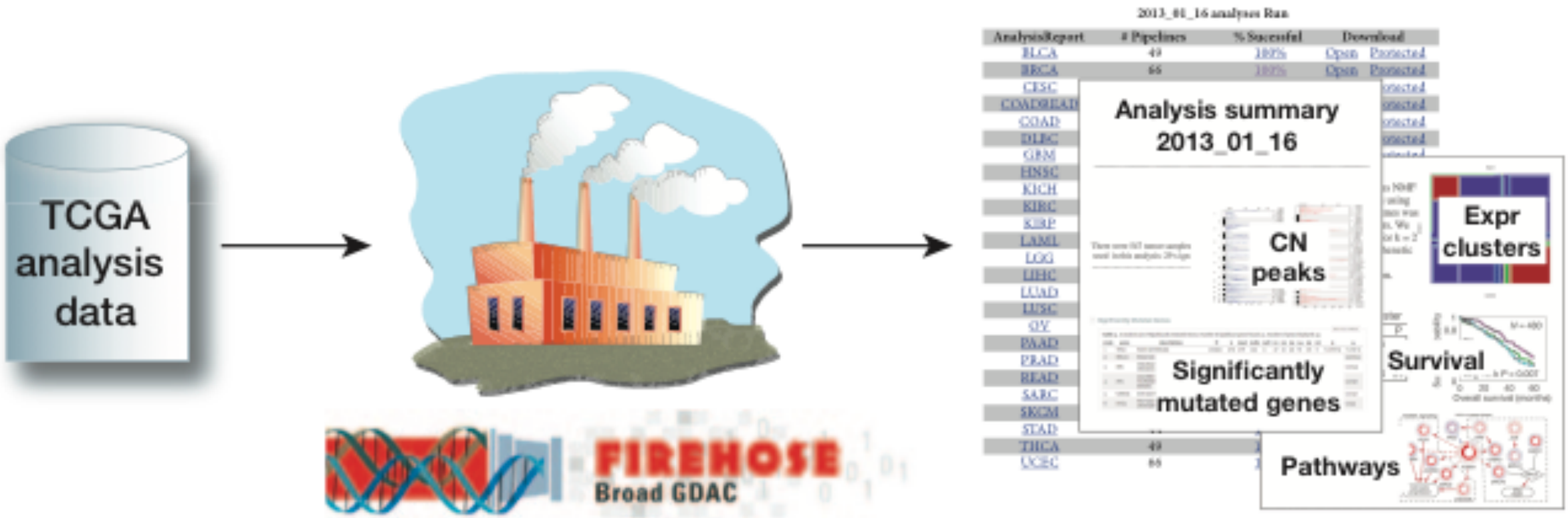


And dozens more academic depts, pharmas, foreign repositories, etc

Tremendous, National-Scale Data Coordination & Standards Challenge



Acute Need for Automation, Systematic Rigor, and Transparency



2,500+ pipelines per month, across all TCGA disease types

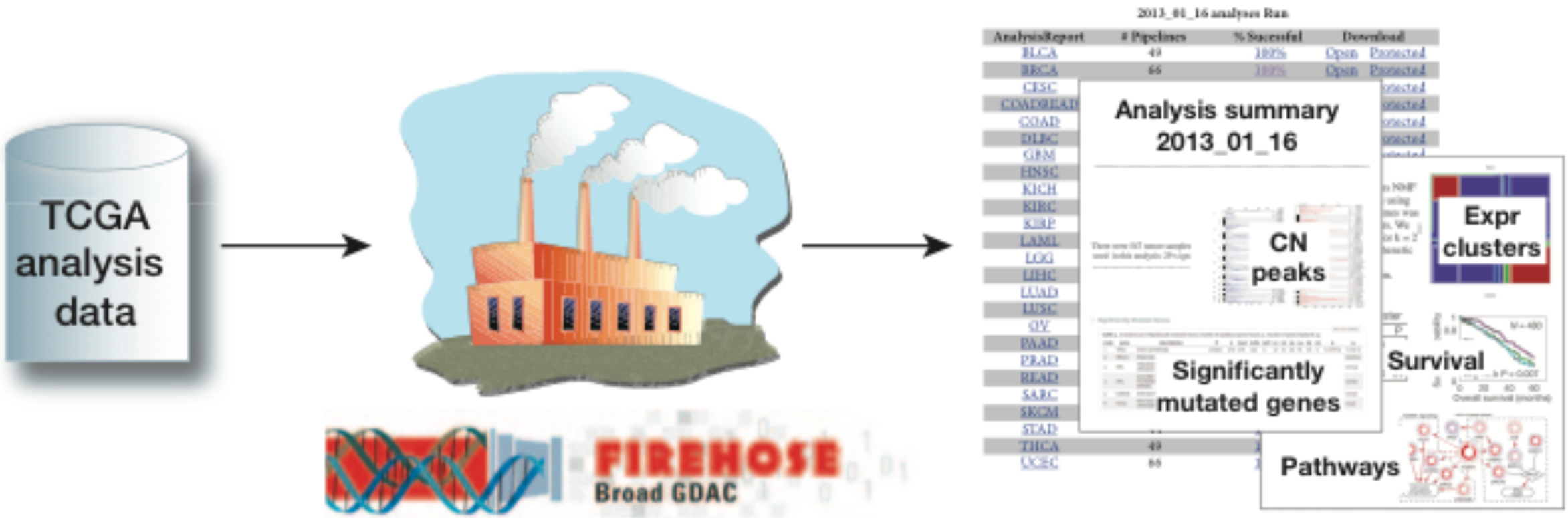
Results dashboards and biologist-friendly reports

Open to public for browsing and automatic download

Democratize TCGA science by lowering entry barriers

Firehose == Data Factory

Acute Need for Automation, Systematic Rigor, and Transparency



2,500+ pipelines per month, across all TCGA disease types

Results dashboards and biologist-friendly reports

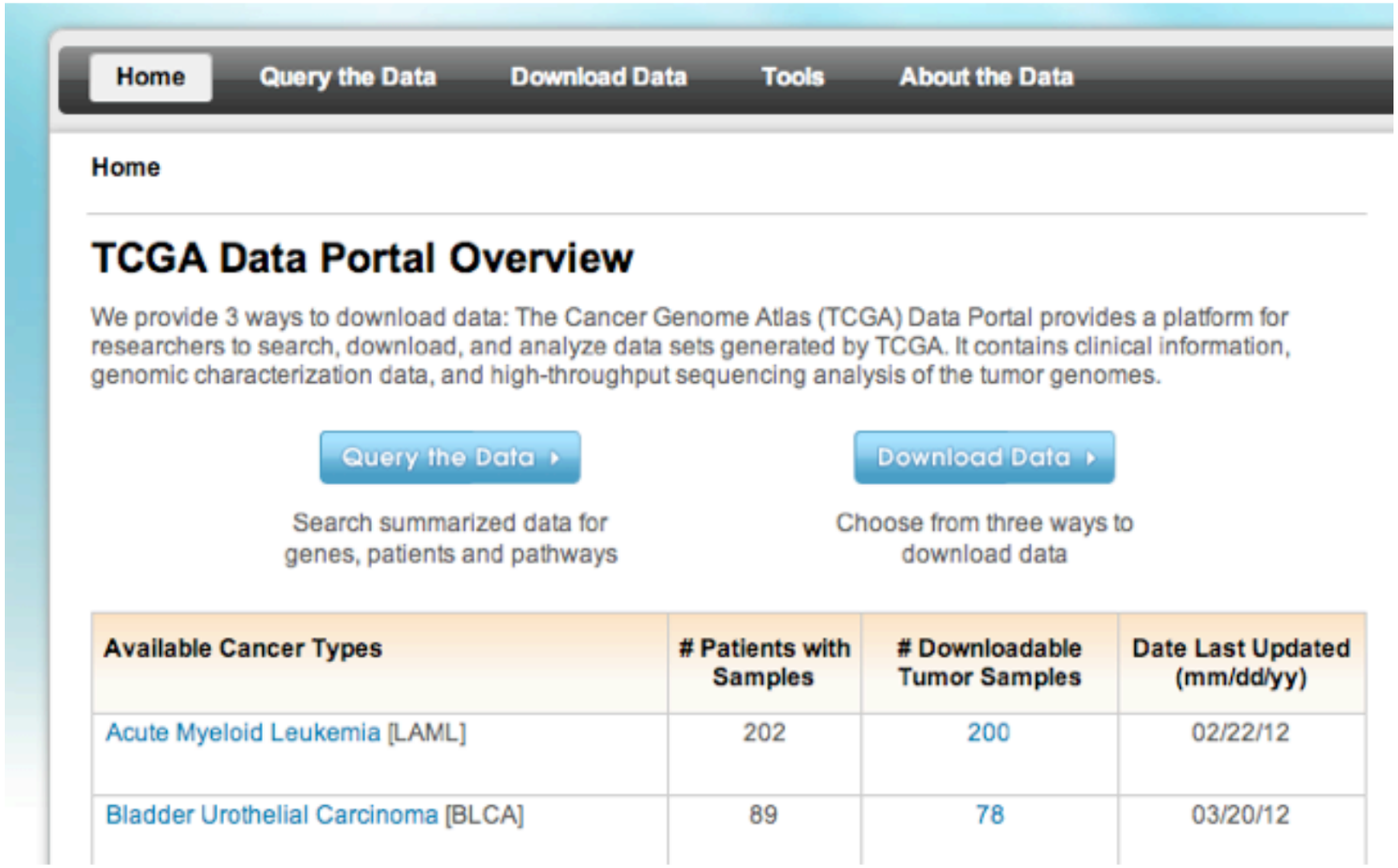
Open to public for browsing and automatic download

Democratize TCGA science by lowering entry barriers

6000

Firehose == Data Factory

But why is this needed when ...



...TCGA already has data archive / portal?

Because TCGA data portal is more “raw” ...

Like giant FTP site: no data aggregate / versioning

Can I easily identify & retrieve ***all*** in one shot?

Because TCGA data portal is more “raw” ...

Like giant FTP site: no data aggregate / versioning

Can I easily identify & retrieve ***all*** in one shot?

How to use portal data directly in my research?

Are they homogeneous?

Or systematically prepared?

To be ready to load in my R or MatLab script?

Because TCGA data portal is more “raw” ...

Like giant FTP site: no data aggregate / versioning

Can I easily identify & retrieve ***all*** in one shot?

How to use portal data directly in my research?

Are they homogeneous?

Or systematically prepared?

To be ready to load in my R or MatLab script?

we had to do this, so would you

Because TCGA data portal is more “raw” ...

Like giant FTP site: no data aggregate / versioning

Can I easily identify & retrieve ***all*** in one shot?

How to use portal data directly in my research?

Are they homogeneous?

Or systematically prepared?

To be ready to load in my R or MatLab script?

we had to do this, so would you

<u>... and does not encompass analyses at all</u>

What if I just want to view copy number peaks in Ovarian (GISTIC)?

Or glance at an expression or methylation cluster?

Must I become an expert first?

One might otherwise need to ...

Spend weeks obtaining protected data credentials

Or becoming a TCGA data guru, obtaining samples spread across many files

And still more time, mastering the analytics

One might otherwise need to ...

Spend weeks obtaining protected data credentials

Or becoming a TCGA data guru, obtaining samples spread across many files

And still more time, mastering the analytics

Complexity & volume preclude this approach for many individuals

2. What?

To Address These Firehose Generates

1. Version-stamped, standardized datasets

   Precursor to automated analyses, durable (DCC)

2. Version-stamped package of standard analyses results

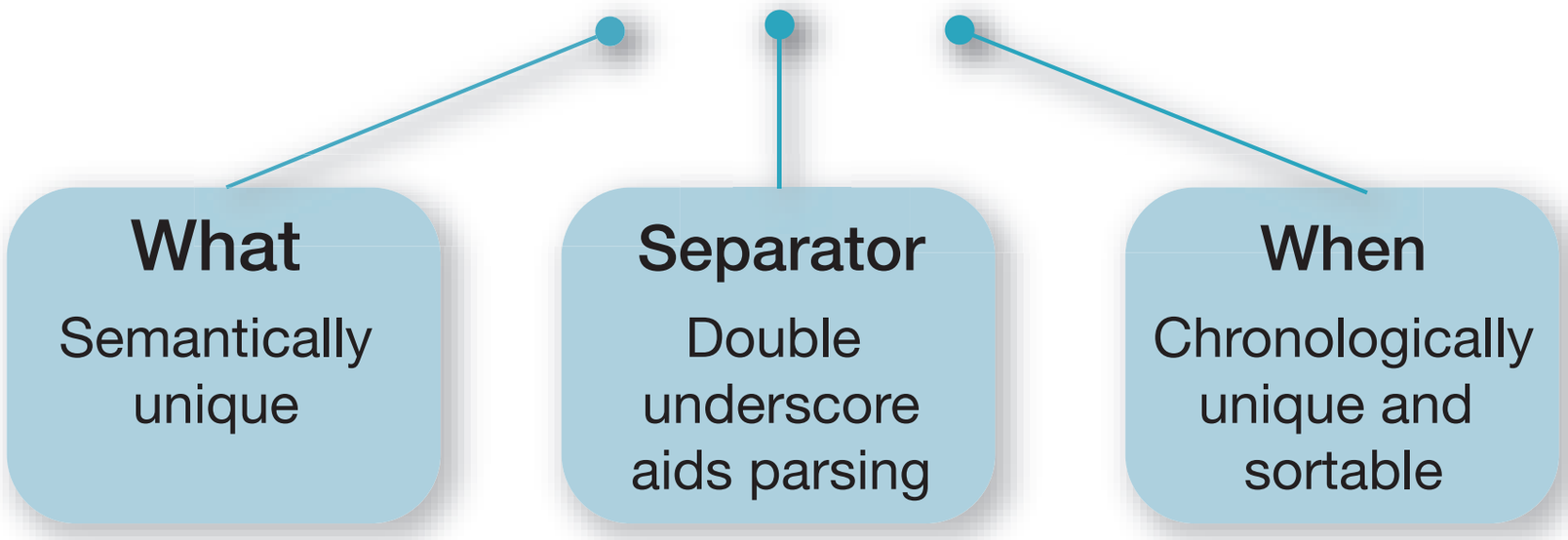
   Dozens of algorithms: GISTIC, MutSig, CNMF, ...

3. With version-stamped, biologist-friendly reports

All of which are citable in the literature (more on that later)

Anatomy of a Firehose Version Stamp

analyses__2013_01_16

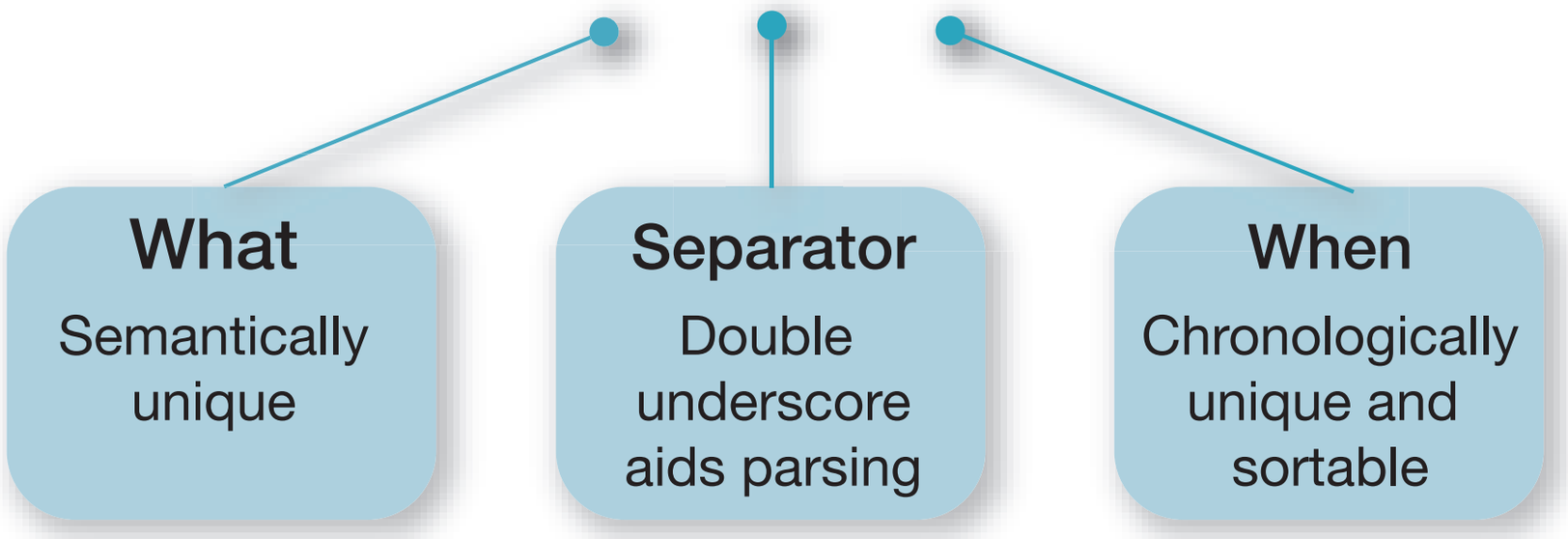


stddata__2013_01_16 — Data snapshot on 16 January 2013, packaged into standardized form

awg_lgg__2013_01_16 — Packages with same date guaranteed to contain same data subset (for example, custom analyses of lower-grade glioma data)

Anatomy of a Firehose Version Stamp

analyses__2013_01_16



stddata__2013_01_16 — Data snapshot on 16 January 2013, packaged into standardized form

awg_lgg__2013_01_16 — Packages with same date guaranteed to contain same data subset (for example, custom analyses of lower-grade glioma data)

<thing>__YYYY_MM_DD

Drilling into big cancer-genome data, ***Nature Methods*** **10**, 293–297 (2013)

Frozen snapshot of all TCGA ***analysis-ready*** data

- ***Cast in a form amenable to immediate algorithmic analysis*** (no additional data preparation required)
- Which provides a ***consistent point of reference*** for analysis and citation by marker papers and users of TCGA data
- Towards a ***formal definition*** of what constitutes a given tumor dataset
- While ***minimizing redundant effort*** across centers and groups to download & prepare data for further analysis
- And ***enhancing provenance and reproducibility***

Frozen snapshot of all TCGA ***analysis-ready*** data

- ***Cast in a form amenable to immediate algorithmic analysis*** (no additional data preparation required)
- Which provides a ***consistent point of reference*** for analysis and citation by marker papers and users of TCGA data
- Towards a ***formal definition*** of what constitutes a given tumor dataset
- While ***minimizing redundant effort*** across centers and groups to download & prepare data for further analysis
- And ***enhancing provenance and reproducibility***

Address BABEL Problem

20 centers in TCGA, little agreement on quantity of samples across analyses

Save time
Decrease waste
Increase quality

How? Many ways ... here are several

1) Because sequencers **create many files**

How? Many ways ... here are several

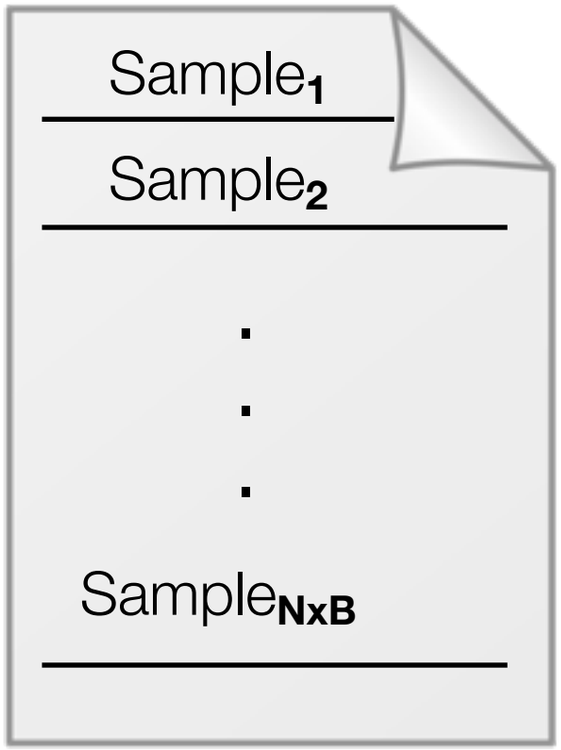
1) Because sequencers **create many files**

One sample = one file

Submitted to DCC in B batches, over months

N samples X B batches = **NxB files**

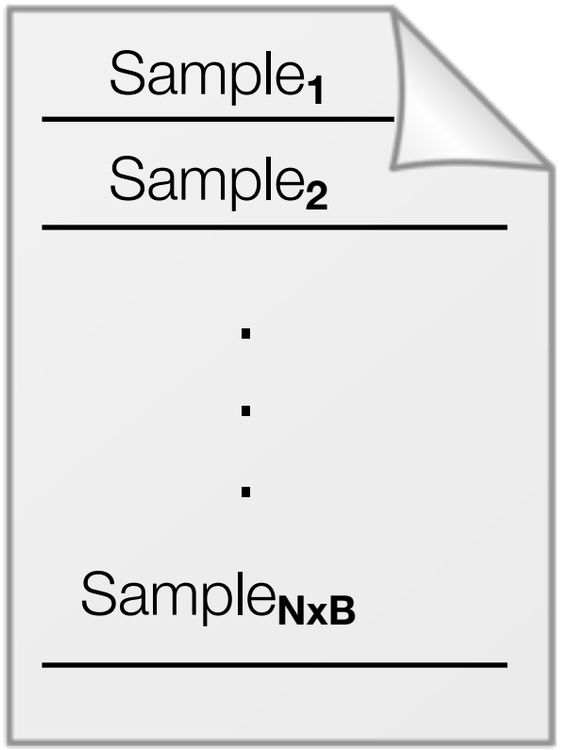
But your brain, R & MatLab code want **one file**



One file = NxB samples

Don't care how/when submitted to DCC

But your brain, R & MatLab code want **one file**

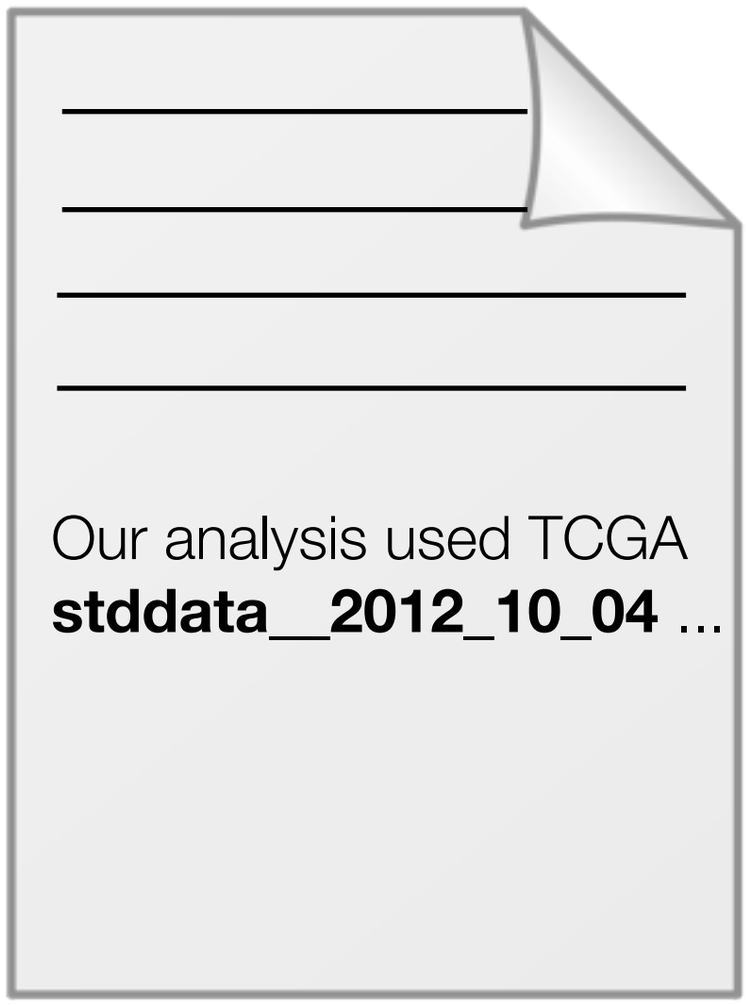


One file = NxB samples
Don't care how/when submitted to DCC

Transparent aggregation over samples, over time (and over operating system: Linux, WinXX, Mac ...)

Wasteful & error-prone to duplicate this at each TCGA center (or at each of your desks)

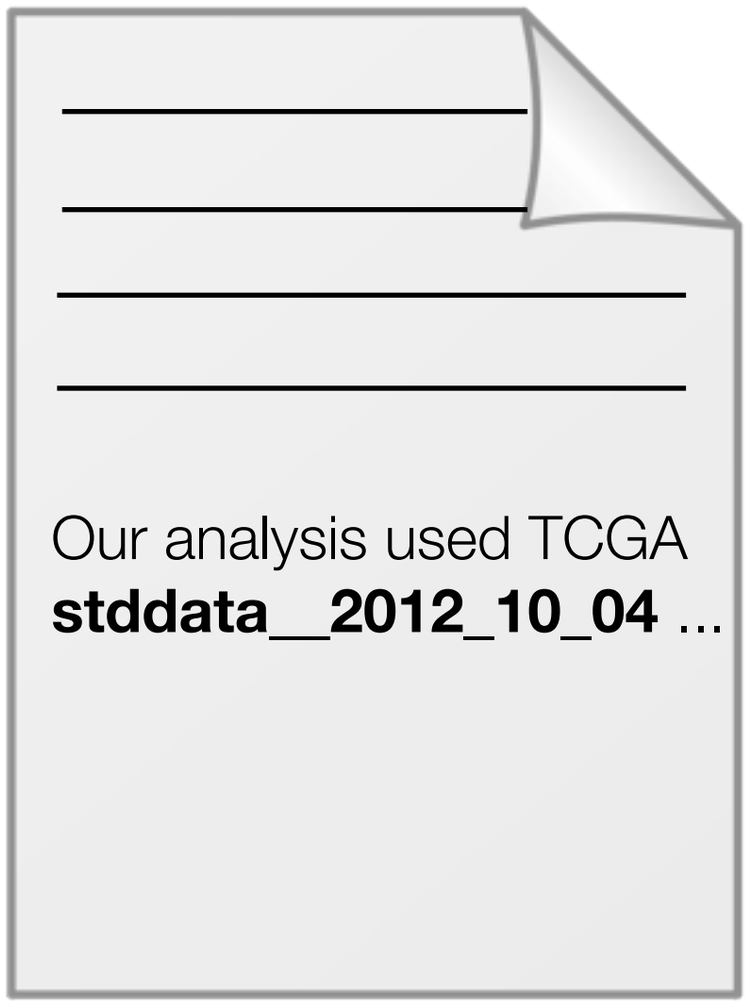
Because you want to **cite one thing**



Consistent point of reference for analysis
and citation by marker papers
and users of TCGA data

Data Science: data must become citable

Because you want to **cite one thing**



Consistent point of reference for analysis
and citation by marker papers
and users of TCGA data

Data Science: data must become citable

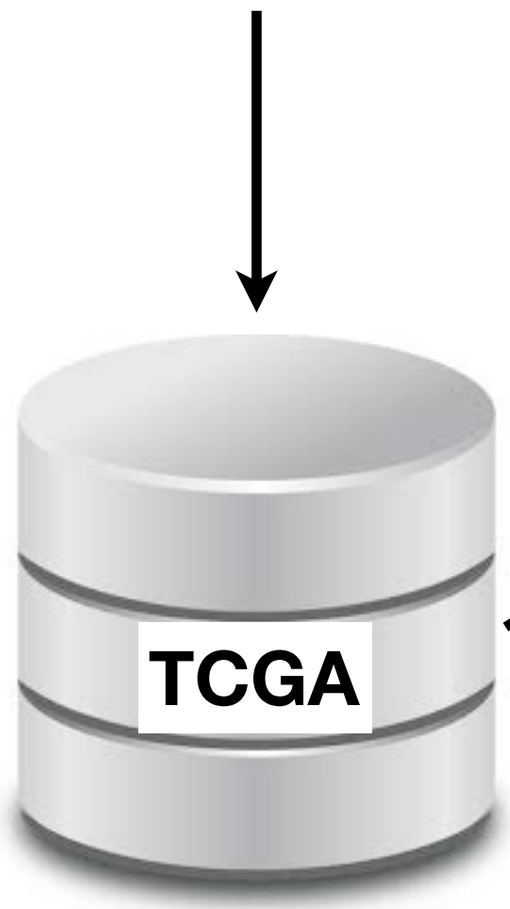
Journals, readers, reproducers want this
Step 1: version-stamping the data aggregates
Step 2: disciplined use of versioned data throughout TCGA

And retrieve it ***clearly & easily***

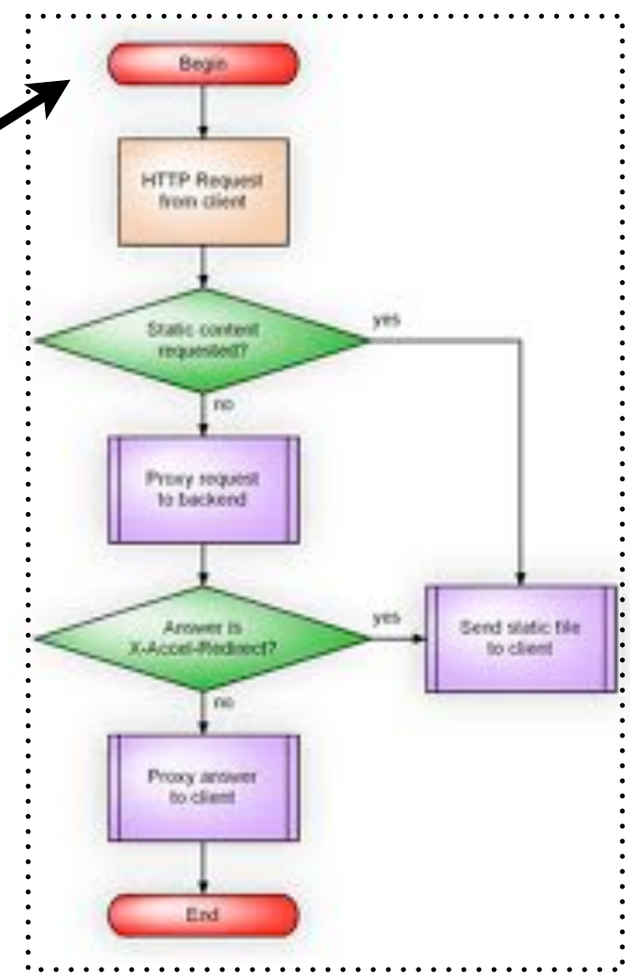
% firehose_get 2012_10_04

And retrieve it ***clearly & easily***

% firehose_get 2012_10_04



Your Algorithm



And **easily** identify what changed

```
%  gdac_diff   2012_09_13   2012_10_04   $PANCAN8

mRNAseq          +161     (2304 total)
CN               +125     (3907 total)
Methylation      +30      (3667 total)
Clinical         +30      (3864 total)
BCR              +16      (4086 total)
```

2 seconds to understand sample accrual differences across 40+ terabytes of data

Unprecedented Scale: KiloPipeline(s) per Month

stddata__2013_04_06	**2192 datasets packaged for DCC**
stddata__2013_04_21	**2265 datasets ...**
analyses__2013_04_21	**942 analyses ...**

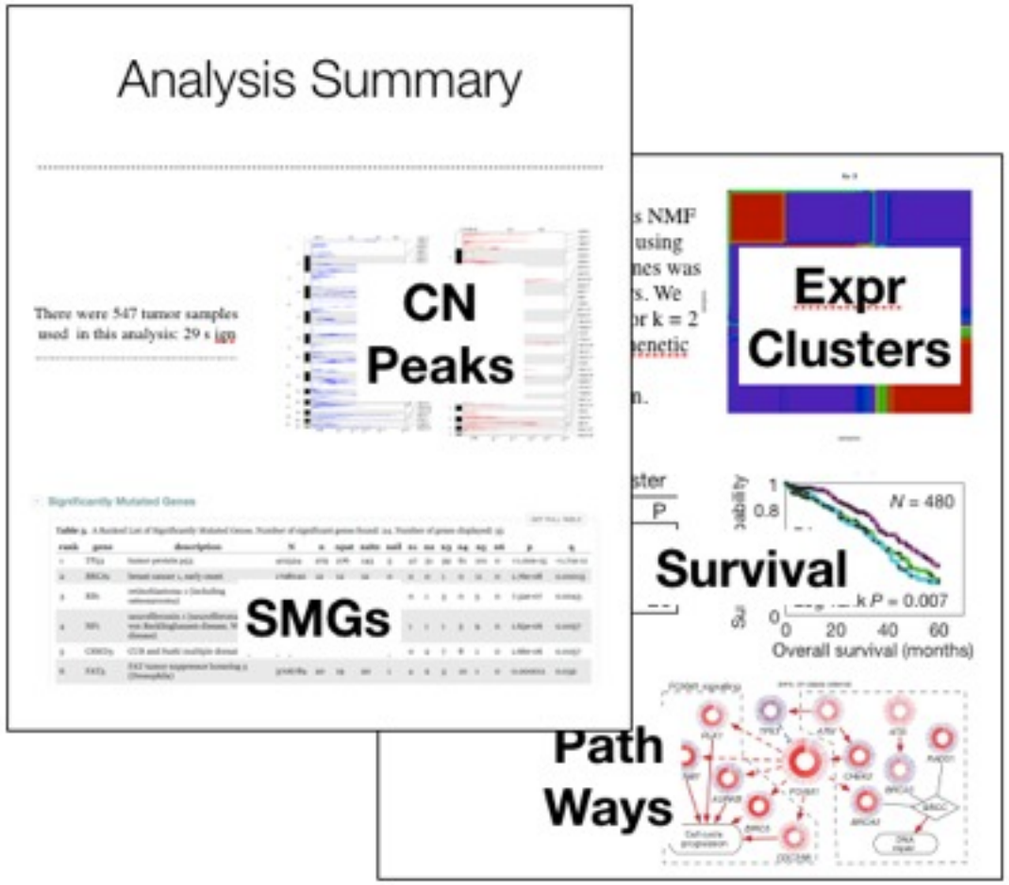
5400 pipelines across 26 disease cohorts

Unprecedented Scale: KiloPipeline(s) per Month

stddata__2013_04_06	**2192 datasets packaged for DCC**
stddata__2013_04_21	**2265 datasets ...**
analyses__2013_04_21	**942 analyses ...**

5400 pipelines across 26 disease cohorts

With up to 40 biologist-friendly analysis reports per disease (~700 total)



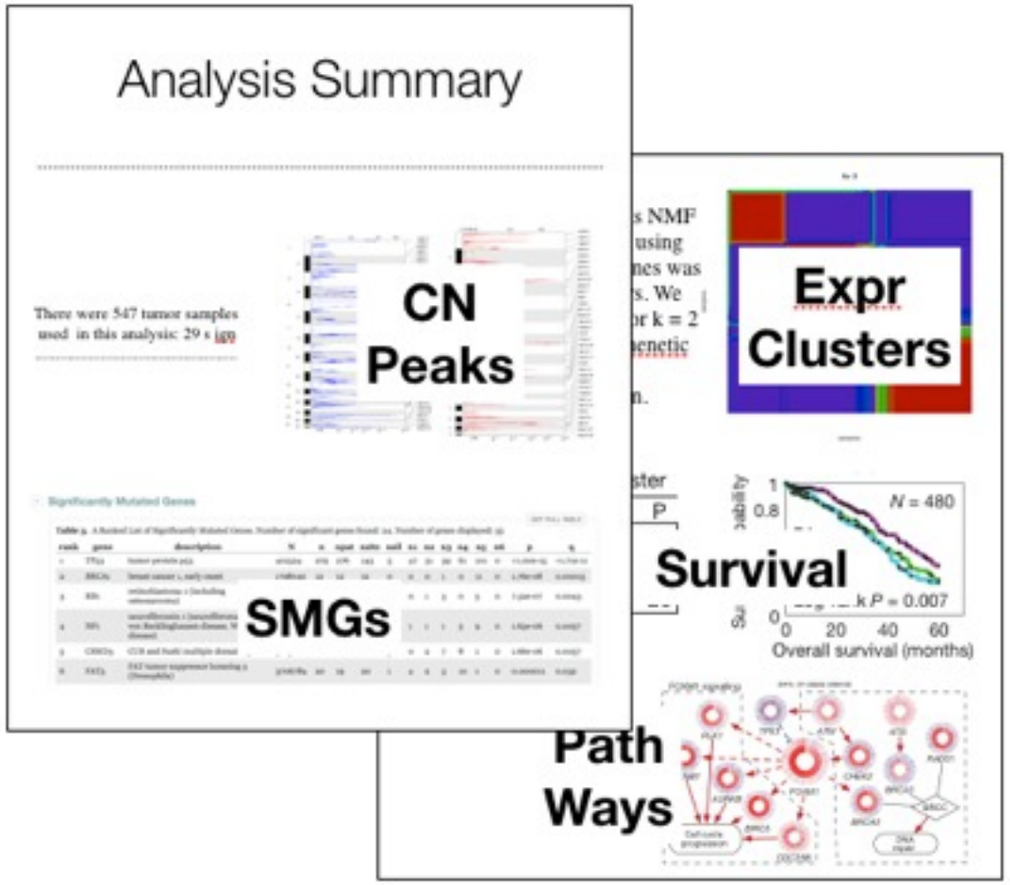
Unprecedented Scale: KiloPipeline(s) per Month

stddata__2013_04_06	**2192 datasets packaged for DCC**
stddata__2013_04_21	**2265 datasets ...**
analyses__2013_04_21	**942 analyses ...**

5400 pipelines across 26 disease cohorts

With up to 40 biologist-friendly analysis reports per disease (~700 total)

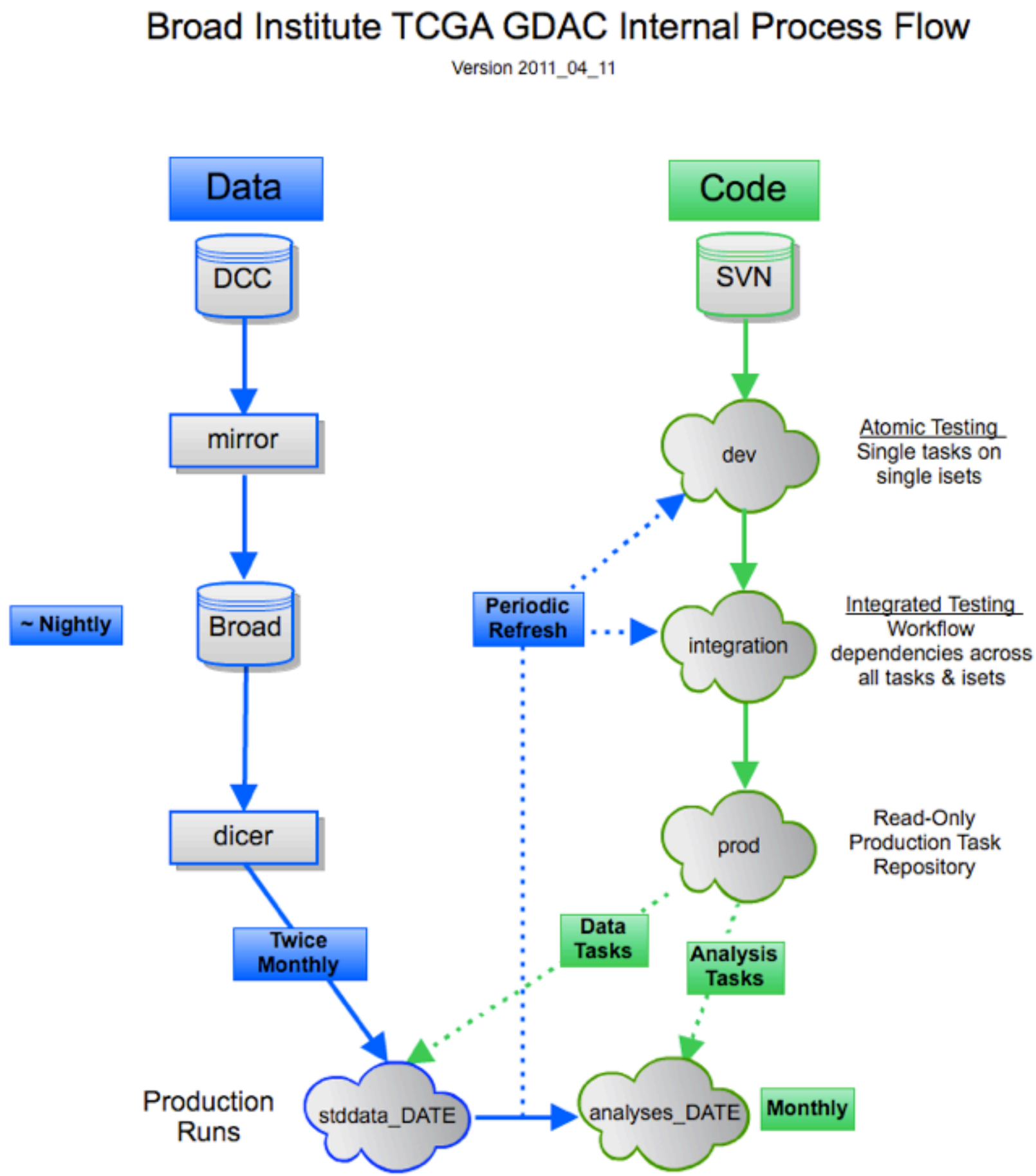
Single Month: April 2013



Subject to Same Engineering Constraints of Timeliness & Reliability as Physical Factories

Not academic one-off

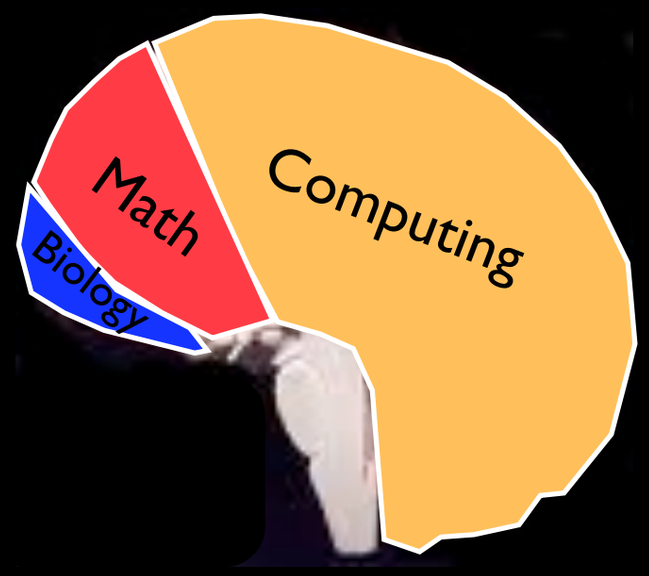
Continuous improvement



But Complex Need Not Be Stupidly Hard

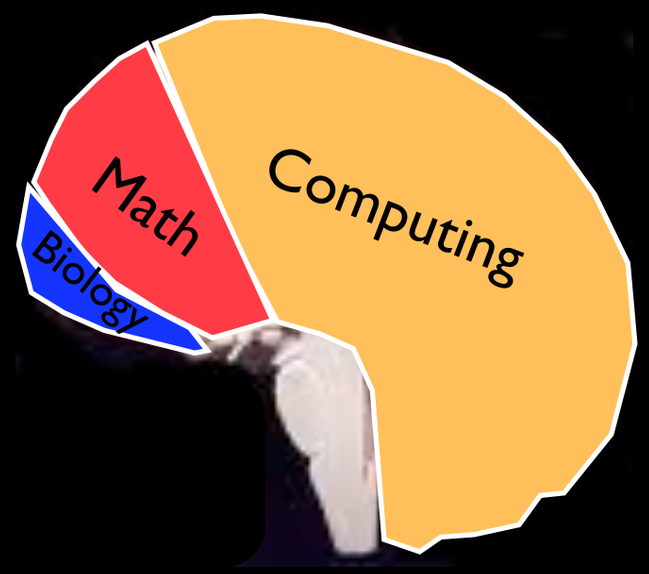
This is Your
Researcher
Brain

But Complex Need Not Be Stupidly Hard

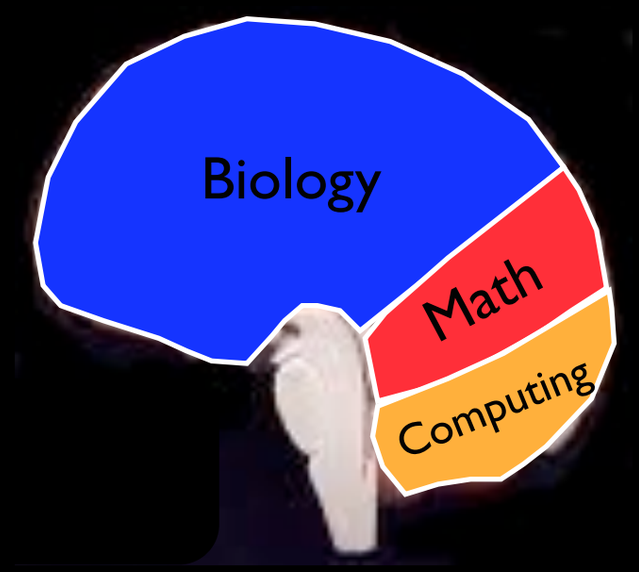


When Coding
Or Data
Exploration
Is Hard

But Complex Need Not Be Stupidly Hard

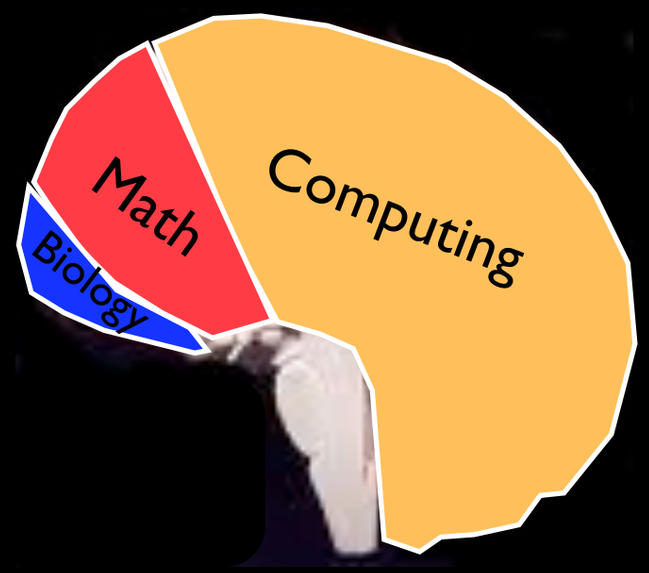


When Coding
Or Data
Exploration
Is Hard

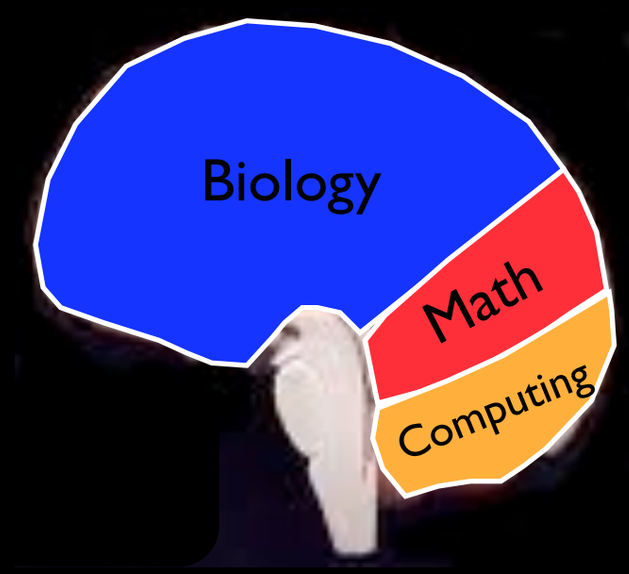


When
Easier

But Complex Need Not Be Stupidly Hard



When Coding
Or Data
Exploration
Is Hard



When
Easier

Civilization advances by extending the number of important operations which we can perform without thought. **A. North Whitehead**

Mission

We strive to lead the world in facilitating the extraction of scientific insight from cancer genomics data. We aim to achieve this through the novel application of quantitative algorithms to cancer genomics data, at unprecedented scale; rigorous & traceable software & process; and lucid, accessible mechanisms of dissemination & exploration.

In this spirit we created ...

Search Results

2013_04_21 stddata Run

DiseaseType	# Datasets	% Processed	Download	
BLCA	24	100%	Open	Protected
BRCA	28	100%	Open	Protected
CESC	18	100%	Open	Protected
COADREAD	34	100%	Open	Protected
COAD	34	100%	Open	Protected
DLBC	8	100%	Open	Protected
ESCA	6	100%	Open	Protected
GBM	35	100%	Open	Protected
HNSC	[illegible]	[illegible]	Open	Protected
KICH	[illegible]	[illegible]	Open	Protected
KIRC	[illegible]	[illegible]	Open	Protected
KIRP	[illegible]	[illegible]	Open	Protected
LAML	[illegible]	[illegible]	Open	Protected
LGG	[illegible]	[illegible]	Open	Protected
LIHC	[illegible]	[illegible]	Open	Protected
LUAD	[illegible]	[illegible]	Open	Protected
LUSC	38	100%	Open	Protected
OV	39	100%	Open	Protected
PRAD	18	100%	Open	Protected
READ	34	100%	Open	Protected
SARC	9	100%	Open	Protected
SKCM	21	100%	Open	Protected
STAD	22	100%	Open	Protected
THCA	24	100%	Open	Protected
UCEC	34	100%	Open	Protected
PANCAN12	58	92%	Open	Protected

Data Dashboard

2013_04_21 analyses Run

AnalysisReport	# Pipelines	% Successful	Download	
BLCA	59	100%	Open	Protected
BRCA	76	100%	Open	Protected
CESC	56	100%	Open	Protected
COADREAD	76	100%	Open	Protected
COAD	76	100%	Open	Protected
DLBC	10	100%	Open	Protected
ESCA	9	100%	Open	Protected
GBM	77	100%	Open	Protected
HNSC	[illegible]	[illegible]	Open	Protected
KICH	[illegible]	[illegible]	Open	Protected
KIRC	[illegible]	[illegible]	Open	Protected
KIRP	[illegible]	[illegible]	Open	Protected
LGG	[illegible]	[illegible]	Open	Protected
LIHC	[illegible]	[illegible]	Open	Protected
LUAD	[illegible]	[illegible]	Open	Protected
LUSC	[illegible]	[illegible]	Open	Protected
OV	81	100%	Open	Protected
PRAD	56	100%	Open	Protected
READ	76	100%	Open	Protected
SARC	13	100%	Open	Protected
SKCM	59	100%	Open	Protected
STAD	56	100%	Open	Protected
THCA	59	100%	Open	Protected
UCEC	76	100%	Open	Protected
LAML	55	98%	Open	Protected
PANCAN12	14	61%	Open	Protected

Analysis Dashboard

Data Notes **FAQ** **Download** **Result Reports** **Analysis Notes**

Welcome to the online home of the Broad Institute's Genome Data Analysis Center (GDAC). On behalf of The Cancer Genome Atlas, we've designed and operate scientific data and analysis pipelines which pump terabyte-scale genomic datasets through scores of quantitative algorithms, in the hope of accelerating the understanding of cancer. See the dashboards above for details of the latest runs, or our presentations page for more background information. Note that downloading data from our site constitutes agreement to this data usage policy.

gdac.broadinstitute.org

With open/public/passwordless dashboards

2013_04_21 analyses Run

Tables of Ingested Data: HTML PNG TSV

AnalysisReport	# Pipelines	% Successful	Download	
BLCA	59	100%	Open	Protected
BRCA	76	100%	Open	Protected
CESC	56	100%	Open	Protected
COADREAD	76	100%	Open	Protected
COAD	76	100%	Open	Protected
DLBC	10	100%	Open	Protected
ESCA	9	100%	Open	Protected
GBM	77	100%	Open	Protected
HNSC	59	100%	Open	Protected
KICH	28	100%	Open	Protected
KIRC	76	100%	Open	Protected
KIRP	73	100%	Open	Protected
LGG	73	100%	Open	Protected
LIHC	34	100%	Open	Protected
LUAD	76	100%	Open	Protected
LUSC	76	100%	Open	Protected
OV	81	100%	Open	Protected
PRAD	56	100%	Open	Protected
READ	76	100%	Open	Protected
SARC	13	100%	Open	Protected
SKCM	59	100%	Open	Protected
STAD	56	100%	Open	Protected
THCA	59	100%	Open	Protected
UCEC	76	100%	Open	Protected
LAML	55	98%	Open	Protected
PANCAN12	14	61%	Open	Protected

Analysis Reports **Release Notes** **Download** **FAQ** **Nomenclature** **Previous Runs**

With open/public/passwordless dashboards

2013_04_21 analyses Run

Tables of Ingested Data: HTML PNG TSV

AnalysisReport	# Pipelines	% Successful	Download	
BLCA	59	100%	Open	Protected
BRCA	76	100%	Open	Protected
CESC	56	100%	Open	Protected
COADREAD	76	100%	Open	Protected
COAD	76	100%	Open	Protected
DLBC	10	100%	Open	Protected
ESCA	9	100%	Open	Protected
GBM	77	100%	Open	Protected
HNSC	59	100%	Open	Protected
KICH	28	100%	Open	Protected
KIRC	76	100%	Open	Protected
KIRP	73	100%	Open	Protected
LGG	73	100%	Open	Protected
LIHC	34	100%	Open	Protected
LUAD	76	100%	Open	Protected
LUSC	76	100%	Open	Protected
OV	81	100%	Open	Protected
PRAD	56	100%	Open	Protected
READ	76	100%	Open	Protected
SARC	13	100%	Open	Protected
SKCM	59	100%	Open	Protected
STAD	56	100%	Open	Protected
THCA	59	100%	Open	Protected
UCEC	76	100%	Open	Protected
LAML	55	98%	Open	Protected
PANCAN12	14	61%	Open	Protected

Analysis Reports Release Notes Download FAQ Nomenclature Previous Runs

Tumor	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BLCA	126	67	58	0	78	0	56	0	88	0	28
BRCA	899	862	833	0	858	529	777	0	809	408	507
CESC	122	31	68	0	0	0	0	0	42	0	36
COADREAD	592	591	575	76	584	224	83	0	255	399	224
DLBC	27	0	0	0	0	0	0	0	0	0	0
GBM	596	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	214	276
HNSC	312	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0
KICH	65	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0
KIRC	502	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	454	403
KIRP	135	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0
LAML	202	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	199
LGG	181	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0
LIHC	99	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0
LNNH	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0
LUAD	439	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	229
LUSC	360	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	178
OV	592	580	564	0	551	568	297	564	454	412	316
PAAD	48	0	14	0	30	0	0	0	0	0	0
PRAD	174	127	100	0	153	0	53	0	81	0	83
SARC	29	0	0	0	0	0	0	0	0	0	0
SKCM	253	129	219	0	240	0	212	0	240	0	0
STAD	226	159	132	0	133	0	57	0	127	0	133
THCA	353	188	228	0	230	0	158	0	138	0	0
UCEC	512	451	430	0	451	54	266	0	367	200	248
PANCANCER	6846	5633	5386	76	5465	2218	3460	1055	3976	2087	2860

Sample Counts

(tabular/programmatic too)

	Pipeline	NotRunnable	Runnable	InProcess	Successful	Unsuccessful
1	Aggregate_Clusters	0	0	0	1	0
2	CopyNumber_GeneBySample	0	0	0	1	0
3	CopyNumber_Gistic2	0	0	0	1	0
4	Correlate_Clinical_vs_CopyNumber_Arm	0	0	0	1	0
5	Correlate_Clinical_vs_CopyNumber_Focal	0	0	0	1	0
6	Correlate_Clinical_vs_miR	0	0	0	1	0
7	Correlate_Clinical_vs_Molecular_Signatures	0	0	0	1	0
8	Correlate_Clinical_vs_mRNA	0	0	0	1	0
9	Correlate_Clinical_vs_Mutation	0	0	0	1	0
10	Correlate_CopyNumber_vs_miR	0	0	0	1	0
11	Correlate_CopyNumber_vs_mRNA	0	0	0	1	0
12	Correlate_CopyNumber_vs_mRNAseq	[illegible]	[illegible]	[illegible]	1	0
13	Correlate_Methylation_v[illegible]	[illegible]	[illegible]	[illegible]	1	0
14	Methylation_Clustering_[illegible]	[illegible]	[illegible]	[illegible]	1	0
15	Methylation_Preprocess	[illegible]	[illegible]	[illegible]	1	0
16	miRseq_Clustering_CN[illegible]	[illegible]	[illegible]	[illegible]	1	0
17	miRseq_Clustering_Con[illegible]	[illegible]	[illegible]	[illegible]	1	0
18	miRseq_Preprocess	[illegible]	[illegible]	[illegible]	1	0
19	miR_Clustering_CNMF	[illegible]	[illegible]	[illegible]	1	0
20	miR_Clustering_Consen[illegible]	[illegible]	[illegible]	[illegible]	1	0
21	miR_FindDirectTargets	[illegible]	[illegible]	[illegible]	1	0
22	miR_Preprocess	[illegible]	[illegible]	[illegible]	1	0
23	mRNAseq_Clustering_CNMF	0	0	0	1	0
24	mRNAseq_Clustering_Consensus	0	0	0	1	0
25	mRNAseq_Preprocess	0	0	0	1	0
26	mRNA_Clustering_CNMF	0	0	0	1	0
27	mRNA_Clustering_Consensus	0	0	0	1	0
28	mRNA_Preprocess_Median	0	0	0	1	0
29	Mutation_Assessor	0	0	0	1	0
30	Mutation_Significance	0	0	0	1	0
31	Pathway_FindEnrichedGenes	0	0	0	1	0
32	Pathway_Paradigm_Expression	0	0	0	1	0
33	Pathway_Paradigm_Expression_CopyNumber	0	0	0	1	0
34	RPPA_Clustering_CNMF	0	0	0	1	0
35	RPPA_Clustering_Consensus	0	0	0	1	0

Analyses Performed

Offering biologist-friendly result reports

2013_04_21 analyses Run

Tables of Ingested Data: HTML PNG TSV Samples Summary: Report

AnalysisReport	# Pipelines	% Successful	Download	
BLCA	59	100%	Open	Protected
BRCA	76	100%	Open	Protected
CESC	56	100%	Open	Protected
COADREAD	76	100%	Open	Protected
COAD	76	100%	Open	Protected
DLBC	10	100%	Open	Protected
ESCA	9	100%	Open	Protected
GBM	77	100%	Open	Protected
HNSC	59	100%	Open	Protected
KICH	28	100%	Open	Protected
KIRC	76	100%	Open	Protected
KIRP	73	100%	Open	Protected
LGG	73	100%	Open	Protected
LIHC	34	100%	Open	Protected
LUAD	76	100%	Open	Protected
LUSC	76	100%	Open	Protected
OV	81	100%	Open	Protected
PRAD	56	100%	Open	Protected
READ	76	100%	Open	Protected
SARC	13	100%	Open	Protected
SKCM	59	100%	Open	Protected
STAD	56	100%	Open	Protected
THCA	59	100%	Open	Protected
UCEC	76	100%	Open	Protected
LAML	55	98%	Open	Protected
PANCAN12	14	61%	Open	Protected

Offering biologist-friendly result reports

2013_04_21 analyses Run

Tables of Ingested Data: HTML PNG TSV Samples Summary: Report

AnalysisReport	# Pipelines	% Successful	Download	
BLCA	59	100%	Open	Protected
BRCA	76	100%	Open	Protected
CESC	56	100%	Open	Protected
COADREAD	76	100%	Open	Protected
COAD	76	100%	Open	Protected
DLBC	10	100%	Open	Protected
ESCA	9	100%	Open	Protected
GBM	77	100%	Open	Protected
HNSC	59	100%	Open	Protected
KICH	28	100%	Open	Protected
KIRC	76	100%	Open	Protected
KIRP	73	100%	Open	Protected
LGG	73	100%	Open	Protected
LIHC	34	100%	Open	Protected
LUAD	76	100%	Open	Protected
LUSC	76	100%	Open	Protected
OV	81	100%	Open	Protected
PRAD	56	100%	Open	Protected
READ	76	100%	Open	Protected
SARC	13	100%	Open	Protected
SKCM	59	100%	Open	Protected
STAD	56	100%	Open	Protected
THCA	59	100%	Open	Protected
UCEC	76	100%	Open	Protected
LAML	55	98%	Open	Protected
PANCAN12	14	61%	Open	Protected

UP < > EXPAND ALL COLLAPSE ALL SET AUTO WIDTH PRINT

Analysis Overview for Ovarian Serous Cystadenocarcinoma

Maintained by TCGA GDAC Team *(Broad Institute/Dana-Farber Cancer Institute/Harvard Medical School)*

Overview

Introduction

Summary

Note: These results are offered to the community as an additional reference point, enabling a wide range of cancer biologists, clinical investigators, and genome and computational scientists to easily incorporate TCGA into the backdrop of ongoing research. While every effort is made to ensure that Firehose input data and algorithms are of the highest possible quality, these analyses have not been reviewed by domain experts.

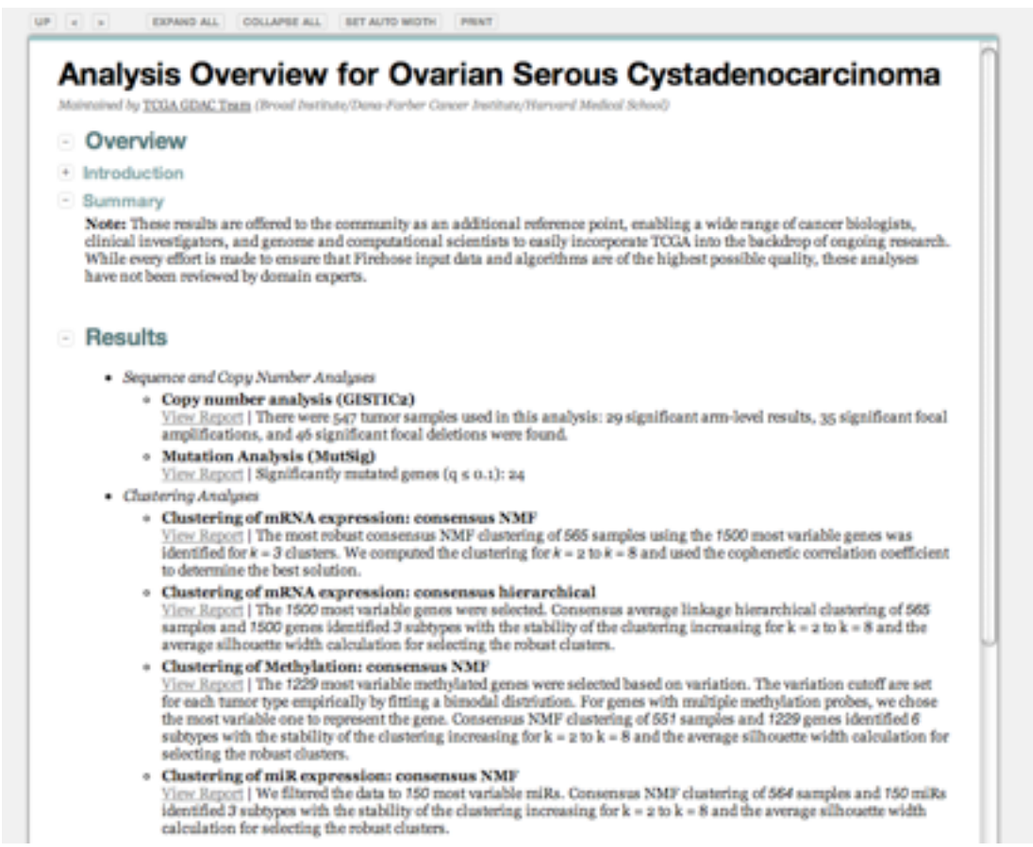
Results

- *Sequence and Copy Number Analyses*
  - **Copy number analysis (GISTIC2)**
    View Report | There were 547 tumor samples used in this analysis: 29 significant arm-level results, 35 significant focal amplifications, and 46 significant focal deletions were found.
  - **Mutation Analysis (MutSig)**
    View Report | Significantly mutated genes ($q \leq 0.1$): 24
- *Clustering Analyses*
  - **Clustering of mRNA expression: consensus NMF**
    View Report | The most robust consensus NMF clustering of *565* samples using the *1500* most variable genes was identified for $k = 3$ clusters. We computed the clustering for $k = 2$ to $k = 8$ and used the cophenetic correlation coefficient to determine the best solution.
  - **Clustering of mRNA expression: consensus hierarchical**
    View Report | The *1500* most variable genes were selected. Consensus average linkage hierarchical clustering of *565* samples and *1500* genes identified *3* subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.
  - **Clustering of Methylation: consensus NMF**
    View Report | The *1229* most variable methylated genes were selected based on variation. The variation cutoff are set for each tumor type empirically by fitting a bimodal distriution. For genes with multiple methylation probes, we chose the most variable one to represent the gene. Consensus NMF clustering of *551* samples and *1229* genes identified *6* subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.
  - **Clustering of miR expression: consensus NMF**
    View Report | We filtered the data to *150* most variable miRs. Consensus NMF clustering of *564* samples and *150* miRs identified *3* subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.

Organized like a paper

- Overview (“Abstract”)
- Results
- Methods & Data

With Browser Convenience



UP < > EXPAND ALL COLLAPSE ALL SET AUTO WIDTH PRINT

Analysis Overview for Ovarian Serous Cystadenocarcinoma

Maintained by TCGA GDAC Team (Broad Institute/Dana-Farber Cancer Institute/Harvard Medical School)

Overview

Introduction

Summary

Note: These results are offered to the community as an additional reference point, enabling a wide range of cancer biologists, clinical investigators, and genome and computational scientists to easily incorporate TCGA into the backdrop of ongoing research. While every effort is made to ensure that Firehose input data and algorithms are of the highest possible quality, these analyses have not been reviewed by domain experts.

Results

- *Sequence and Copy Number Analyses*
  - **Copy number analysis (GISTIC2)**
    View Report | There were 547 tumor samples used in this analysis: 29 significant arm-level results, 35 significant focal amplifications, and 46 significant focal deletions were found.
  - **Mutation Analysis (MutSig)**
    View Report | Significantly mutated genes (q ≤ 0.1): 24
- *Clustering Analyses*
  - **Clustering of mRNA expression: consensus NMF**
    View Report | The most robust consensus NMF clustering of 565 samples using the 1500 most variable genes was identified for $k = 3$ clusters. We computed the clustering for $k = 2$ to $k = 8$ and used the cophenetic correlation coefficient to determine the best solution.
  - **Clustering of mRNA expression: consensus hierarchical**
    View Report | The 1500 most variable genes were selected. Consensus average linkage hierarchical clustering of 565 samples and 1500 genes identified 3 subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.
  - **Clustering of Methylation: consensus NMF**
    View Report | The 1229 most variable methylated genes were selected based on variation. The variation cutoff are set for each tumor type empirically by fitting a bimodal distribution. For genes with multiple methylation probes, we chose the most variable one to represent the gene. Consensus NMF clustering of 551 samples and 1229 genes identified 6 subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.
  - **Clustering of miR expression: consensus NMF**
    View Report | We filtered the data to 150 most variable miRs. Consensus NMF clustering of 564 samples and 150 miRs identified 3 subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.

Organized like a paper

- Overview ("Abstract")
- Results
- Methods & Data

With Browser Convenience

UP ‹ › EXPAND ALL COLLAPSE ALL SET AUTO WIDTH PRINT

Analysis Overview for Ovarian Serous Cystadenocarcinoma

Maintained by TCGA GDAC Team (Broad Institute/Dana-Farber Cancer Institute/Harvard Medical School)

Overview

Introduction

Summary

Note: These results are offered to the community as an additional reference point, enabling a wide range of cancer biologists, clinical investigators, and genome and computational scientists to easily incorporate TCGA into the backdrop of ongoing research. While every effort is made to ensure that Firehose input data and algorithms are of the highest possible quality, these analyses have not been reviewed by domain experts.

Results

- *Sequence and Copy Number Analyses*
  - **Copy number analysis (GISTIC2)**
    View Report | There were 547 tumor samples used in this analysis: 29 significant arm-level results, 35 significant focal amplifications, and 46 significant focal deletions were found.
  - **Mutation Analysis (MutSig)**
    View Report | Significantly mutated genes (q ≤ 0.1): 24
- *Clustering Analyses*
  - **Clustering of mRNA expression: consensus NMF**
    View Report | The most robust consensus NMF clustering of 565 samples using the 1500 most variable genes was identified for $k = 3$ clusters. We computed the clustering for $k = 2$ to $k = 8$ and used the cophenetic correlation coefficient to determine the best solution.
  - **Clustering of mRNA expression: consensus hierarchical**
    View Report | The 1500 most variable genes were selected. Consensus average linkage hierarchical clustering of 565 samples and 1500 genes identified 3 subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.
  - **Clustering of Methylation: consensus NMF**
    View Report | The 1229 most variable methylated genes were selected based on variation. The variation cutoff are set for each tumor type empirically by fitting a bimodal distribution. For genes with multiple methylation probes, we chose the most variable one to represent the gene. Consensus NMF clustering of 551 samples and 1229 genes identified 6 subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.
  - **Clustering of miR expression: consensus NMF**
    View Report | We filtered the data to 150 most variable miRs. Consensus NMF clustering of 564 samples and 150 miRs identified 3 subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.

Ovarian Serous Cystadenocarcinoma: Copy number analysis (GISTIC2)

Maintained by Dan DiCara (Broad Institute)

Overview

Introduction

Summary

There were 547 tumor samples used in this analysis: 29 significant arm-level results, 35 significant focal amplifications, and 46 significant focal deletions were found.

Results

Focal results

Figure 1. Genomic positions of amplified regions: the X-axis represents the normalized amplification signals (top) and significance by Q value (bottom). The green line represents the significance cutoff at Q value=0.25.

GET FULL TABLE

Table 1. Amplifications Table - 35 significant amplifications found. Click the link in the last column to view a comprehensive list of candidate genes. If no genes were identified within the peak, the nearest gene appears in brackets.

Cytoband	Q value	Residual Q value	Wide Peak Boundaries	# Genes in Wide Peak
8q24.21	2.645e-77	2.645e-77	chr8:128574848-129810279	5
19q12	1.81476e-87	8.49496e-76	chr19:34947990-35023082	1
3q26.2	1.07220e-60	1.07220e-60	chr3:170905217-170923258	0 [MECOM]

Ovarian Serous Cystadenocarcinoma: Clustering of mRNA expression: consensus NMF

Maintained by Robert Zupko (Broad Institute)

Overview

Introduction

Summary

The most robust consensus NMF clustering of 565 samples using the 1500 most variable genes was identified for $k = 3$ clusters. We computed the clustering for $k = 2$ to $k = 8$ and used the cophenetic correlation coefficient to determine the best solution.

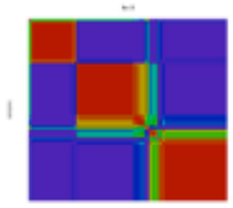
Results

Gene expression patterns of molecular subtypes

Consensus and correlation matrix

GET HIGH-RES IMAGE

Figure 2. The consensus matrix after clustering shows 3 clusters with limited overlap between clusters.



Summary

There were 558 tumor samples used in this analysis: 29 significant arm-level results, 34 significant focal amplifications, and 47 significant focal deletions were found.

Results

Focal results

Arm-level results

GET FULL TABLE

Table 3. Arm-level significance table - 29 significant results found.

Arm	# Genes	Amp Frequency	Amp Z score	Amp Q value	Del Frequency	Del Z score	Del Q value
1p	2121	0.21	0.131	1	0.10	-5.72	1
1q	1955	0.34	6.49	4.26e-10	0.09	-6.29	1
2p	924	0.27	-2.25	1	0.07	-10.7	1
2q	1556	0.22	-2.32	1	0.07	-9.07	1
3p	1062	0.23	-3.6	1	0.20	-4.8	1
3q	1139	0.49	9.71	0			
4p	489	0.14	-7.22	1			
4q	1049	0.07	-7.69	1			

- Standard visual format for ALL pipelines

Summary

There were 558 tumor samples used in this analysis: 29 significant arm-level results, 34 significant focal amplifications, and 47 significant focal deletions were found.

Results

Focal results

Arm-level results

GET FULL TABLE

Table 3. Arm-level significance table - 29 significant results found.

Arm	# Genes	Amp Frequency	Amp Z score	Amp Q value	Del Frequency	Del Z score	Del Q value
1p	2121	0.21	0.131	1	0.10	-5.72	1
1q	1955	0.34	6.49	4.26e-10	0.09	-6.29	1
2p	924	0.27	-2.25	1	0.07	-10.7	1
2q	1556	0.22	-2.32	1	0.07	-9.07	1
3p	1062	0.23	-3.6	1	0.20	-4.8	1
3q	1139	0.49	9.71	0			
4p	489	0.14	-7.22	1			
4q	1049	0.07	-7.69	1			

- Standard visual format for ALL pipelines
- As little as 3-5 simple R calls
- Thoughtfully Scoped:
  - drill from overview to details
  - Significant results “bubble up”

Summary

There were 558 tumor samples used in this analysis: 29 significant arm-level results, 34 significant focal amplifications, and 47 significant focal deletions were found.

Results

Focal results

Arm-level results

GET FULL TABLE

Table 3. Arm-level significance table - 29 significant results found.

Arm	# Genes	Amp Frequency	Amp Z score	Amp Q value	Del Frequency	Del Z score	Del Q value
1p	2121	0.21	0.131	1	0.10	-5.72	1
1q	1955	0.34	6.49	4.26e-10	0.09	-6.29	1
2p	924	0.27	-2.25	1	0.07	-10.7	1
2q	1556	0.22	-2.32	1	0.07	-9.07	1
3p	1062	0.23	-3.6	1	0.20	-4.8	1
3q	1139	0.49	9.71	0			
4p	489	0.14	-7.22	1			
4q	1049	0.07	-7.69	1			

- Standard visual format for ALL pipelines
- As little as 3-5 simple R calls
- Thoughtfully Scoped:
  - drill from overview to details
  - Significant results “bubble up”
  - **don’t miss needle in haystack**

Summary

There were 558 tumor samples used in this analysis: 29 significant arm-level results, 34 significant focal amplifications, and 47 significant focal deletions were found.

Results

Focal results

Arm-level results

GET FULL TABLE

Table 3. Arm-level significance table - 29 significant results found.

Arm	# Genes	Amp Frequency	Amp Z score	Amp Q value	Del Frequency	Del Z score	Del Q value
1p	2121	0.21	0.131	1	0.10	-5.72	1
1q	1955	0.34	6.49	4.26e-10	0.09	-6.29	1
2p	924	0.27	-2.25	1	0.07	-10.7	1
2q	1556	0.22	-2.32	1	0.07	-9.07	1
3p	1062	0.23	-3.6	1	0.20	-4.8	1
3q	1139	0.49	9.71	0			
4p	489	0.14	-7.22	1			
4q	1049	0.07	-7.69	1			

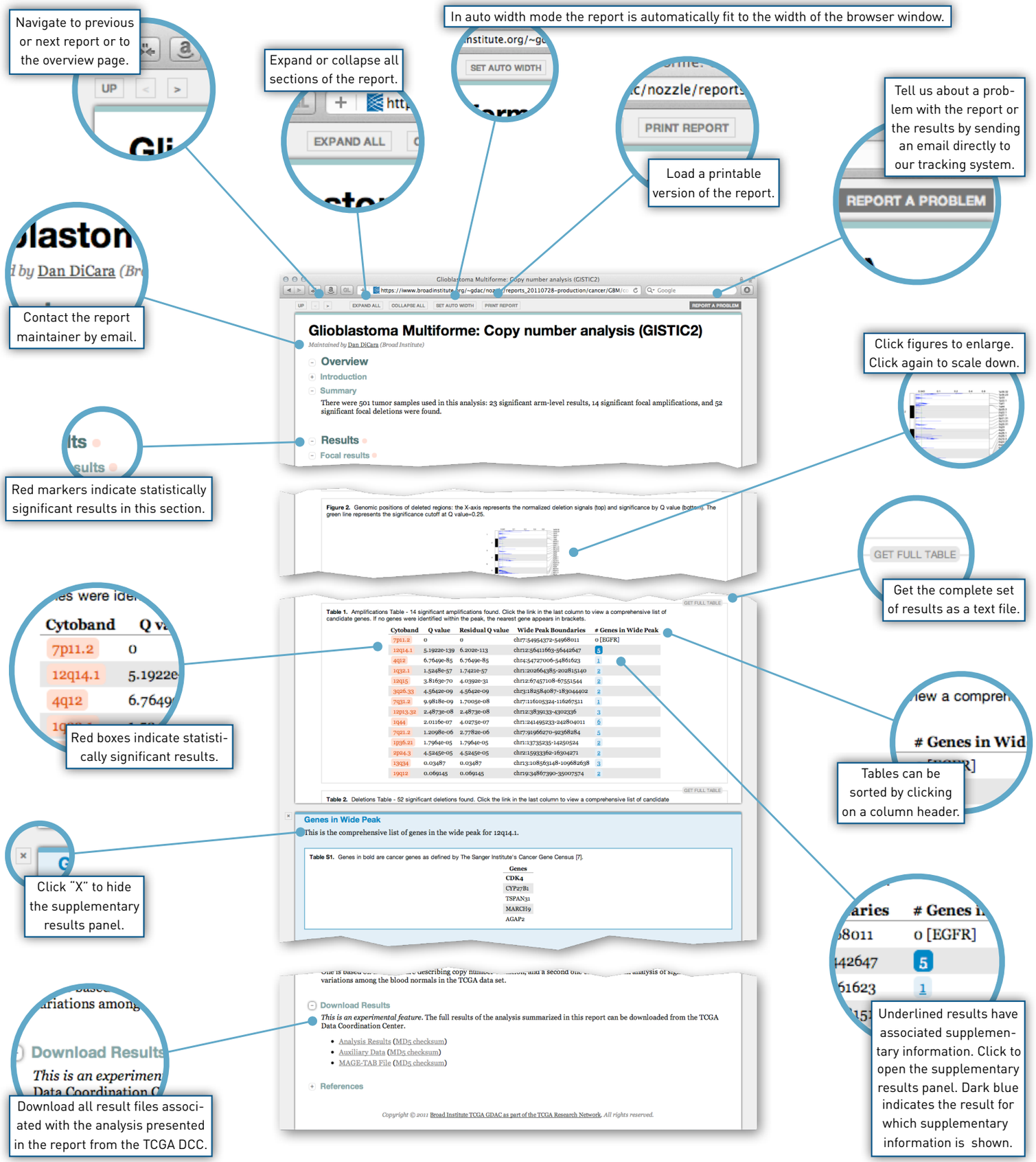
RIGOR: nothing thrown away

- Standard visual format for ALL pipelines
- As little as 3-5 simple R calls
- Thoughtfully Scoped:
  - drill from overview to details
  - Significant results “bubble up”
  - **don’t miss needle in haystack**

Firehose Reports | At-a-Glance

BROAD INSTITUTE

➜ Reports are compatible with Firefox 4+, Chrome 12+, Safari 5+, Opera 11+ and Internet Explorer 9+.



Again, aimed at solid design & engineering

Nozzle package downloadable as open source

Used in multiple external projects

Dead Simple Bulk Retrieval

```
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 posted runs
- **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

20K script

```
% firehose_get -runs
```

Run	At_DCC	Available_From_Broad_GDAC
...		
analyses__2012_04_25	yes	yes
analyses__2012_05_25	yes	yes
analyses__2012_06_23	yes	yes
analyses__2012_07_25	no	yes

```
% firehose_get -tasks analyses 2012_07_25
```

```
...
CopyNumber_Gistic2
Correlate_Clinical_vs_CopyNumber_Arm
Correlate_Clinical_vs_Molecular_Signatures
Correlate_Clinical_vs_Mutation
...
Correlate_CopyNumber_vs_miR
Correlate_CopyNumber_vs_mRNAseq
Correlate_Methylation_vs_mRNA
...
Methylation_Clustering_CNMF
miRseq_Clustering_CNMF
miRseq_Clustering_Consensus
miR_Clustering_CNMF
...
mRNAseq_Clustering_CNMF
...
mRNAseq_Clustering_Consensus
mRNAseq_Preprocess
Mutation_Significance
...
Pathway_FindEnrichedGenes
Pathway_Paradigm_Expression
...
RPPA_Clustering_CNMF
...
```

These analyses are what is described by the reports on our GDAC dashboards

Democratize TCGA science: lower entry barriers

- Enable readers (PIs, bench bios, clinical trialists, DotComs)
- To quickly take pulse of TCGA for given disease type(s)
- With just a few glances at common representational figures
- Not deep head-scratching or big time investment

Democratize TCGA science: lower entry barriers

- Enable readers (PIs, bench bios, clinical trialists, DotComs)
- To quickly take pulse of TCGA for given disease type(s)
- With just a few glances at common representational figures
- Not deep head-scratching or big time investment

"Oh, that's interesting, maybe my code has found something here ... I wonder if this is seen in the Firehose version 2013_04_21 results, too?"

When easy things kept easy, harder things become possible

Established Traction as Nexus Resource

	Pages	Hits	Bandwidth
Interactive Use	**643,858** (221.18 Pages/Visit)	**757,376** (260.17 Hits/Visit)	**277.61 GB** (99997.5 KB/Visit)
firehose_get downloads		108,397+1198	1567.50 GB

May 2013	640K pages	860K hits	1.8 TB traffic
July 2013			> 2 TB traffic

- Across dozens of centers & portals
- Research / Academic / Commercial
- National & International

With Open (-Source) / Transparent Look & Feel

Q: **Why does your table of ingested data show that *disease type XYZ* has *N* mutation samples?**
A: Our precedence rules for ingesting mutation samples are:

1. Prefer manually-curated MAF from the respective analysis working group (AWG), on the premise t
2. When no AWG MAF is available, fall back to using what is available in the DCC by automatic subn
3. Otherwise Firehose will contain zero mutation samples for that disease type.

We're in the process of defining a fourth rule, however, to account for the evolving nature of TCGA mutati
accrue at the DCC (again, automatically submitted by the respective GSCs), and it is natural for analysts

For more information, please consult our provenance table for mutation data, the TCGA MAF workflow ar
will likely support VCFs once they become sufficiently prevalent in the TCGA dataflow.

Q: **Why does your table of ingested data show that *disease type XYZ* has *N* methylation samples**
A: We ingest and support both of the major methylation platforms (meth450 and meth27), therefore the
statistical algorithms used by TCGA AWGs to merge both of these methylation platforms into a single bol
higher resolution data.

Q: **What TCGA sample types are Firehose pipelines executed upon?**
A: Since inception Firehose analyses have been executed upon tumor samples and then correlated with
exception is melanoma (SKCM), which we analyze using metastatic tumor samples (code 06) as it is usu
we will include a larger range of sample types, including normals.

Q: **What do you do when multiple aliquot barcodes exist for a given sample/portion/analyte comb**
A: To date GDAC analyses have proceeded upon one single tumor sample per patient, so when multiple
metrics, we use the following rules to make such selections:

FAQ

Dashboard › Broad TCGA GDAC › Browse Space › Mail Archive › Thread Browse ▾ Log In Search

› Re: [GDAC-users] firehose - download normal expression values

FIREHOSE Broad GDAC

Re: [GDAC-users] firehose - download normal expression values

Subject: **Re: [GDAC-users] firehose - download normal expression values** (find more)
From: David Tamborero <hidden> (find more)
Date: Aug 26, 2012 14:22

```
Thank you very much, your work and help is priceless.

2012/8/24 Michael S. Noble <hidden>

>
> Dear David,
>
> Apologies for the delay in responding.  Yes, you are right: our outputs do
> not
> contain normals.  This is partly a legacy held over from the TCGA pilot
> studies, which is where many of the analyses in our GDAC originally stem
> from.  Our FAQ online at gdac.broadinstitute.org discusses this in the
> section
>
>     Q: What TCGA sample types are Firehose pipelines executed upon?
>
> and points out that we aim to support normals in the Fall of 2012.
>
> Regards,
> Mike Noble
>
```

Searchable Mail Archive

Dashboard › Broad TCGA GDAC › Browse Space › Mail Archive › Thread Browse ▾ Log In Search

› Re: [GDAC-users] firehose - download normal expression values

FIREHOSE Broad GDAC

Re: [GDAC-users] firehose - download normal expression values

Subject: **Re: [GDAC-users] firehose - download normal expression values** (find more)
From: David Tamborero <hidden> (find more)
Date: Aug 26, 2012 14:22

Thank you very much, your work and help is priceless.

2012/8/24 Michael S. Noble <hidden>

>
> Dear David,
>
> Apologies for the delay in responding. Yes, you are right: our outputs do
> not
> contain normals. This is partly a legacy held over from the TCGA pilot
> studies, which is where many of the analyses in our GDAC originally stem
> from. Our FAQ online at gdac.broadinstitute.org discusses this in the
> section
>
> Q: What TCGA sample types are Firehose pipelines
>
> and points out that we aim to support normals in the
>

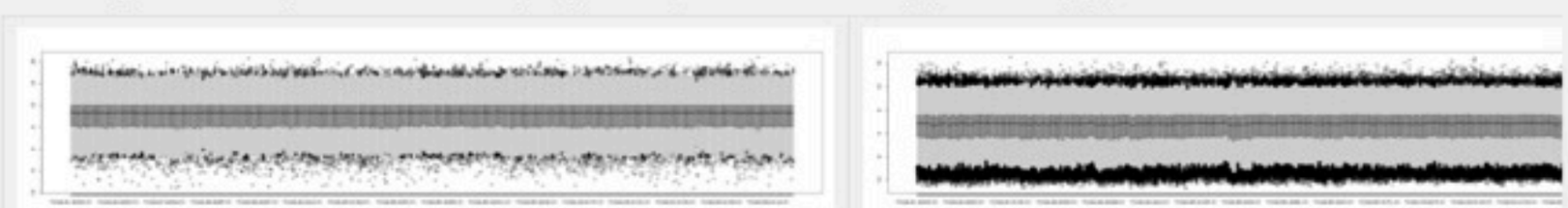
> Regards,
> Mike Noble
>

Searchable Mail Archive

Detailed Release Notes

June 2012 (2012_06_23)

1. Increased number of archives generated from 777 to 993
2. Increased number of reports from 227 to 252
3. 2,244 new samples reflected since May analyses run, due to more data and better counting:
   - 76 LowP (new sample type - Low Pass DNAseq)
   - 230 BCR
   - 307 Clinical
   - 618 mRNAseq
   - 937 miRseq
   - 76 MAF
4. GISTIC2 report now includes a description of both the input and output files in the *Methods & Data* section
5. Methylation data:
   - Rewired pipelines to include meth450 platform, and also give it preference over meth27 when both are present. (Methods to combine 450 & 27 analytically are not in Firehose: would be nice for AWGs to provide if possible)
   - This greatly increases count of methylation samples flowing through analyses (e.g. UCEC 117–>363)
   - Most clusterings show similar results, but some are discordant with previous runs: we could use AWG help to evaluate, and will post comparative analysis online towards that end
6. New clustering pipelines heuristic: a sample will be dropped from analyses when 80% or more genes are absent.
7. mRNAseq: we now utilize rnaseqv2 archives, but fall back to v1 rnaseq when v2 is not available for a given tumor type
   - RSEM estimation used for downstream clustering & correlation analysis, when available, otherwise RPKM estimation will used
   - RSEM is used to estimate gene and transcript abundances (http://deweylab.biostat.wisc.edu/rsem/rsem-calculate-expression.html); values are normalized to a fixed upper quartile value of 1000 for gene and 300 for transcript level estimates, and the normalized values are placed in a separate file (From the DCC document).
   - The following showed the boxplot of BRCA mRNAseq samples with log2 transformed RESM (left) and RPKM (right).



3. Recent Highlights

Custom Runs for Analysis Working Groups

- limited to a single disease cohort
- *and in particular subtypes thereof*
- executed by request of the AWG
- on latest snapshot of data from DCC
- avoid time & sample lag of monthly runs

Custom Runs for Analysis Working Groups

- limited to a single disease cohort
- *and in particular subtypes thereof*
- executed by request of the AWG
- on latest snapshot of data from DCC
- avoid time & sample lag of monthly runs

Provides *real time* scientific value to TCGA AWGs

Using same internal Firehose machinery, public-facing dashboards, ***Nozzle*** reports, **`firehose_get`** etc, known to community

DiseaseType	AWG Run Dashboard
GBM	2013_02_17
LGG	2013_04_06
	2013_02_03
	2013_01_16
HNSC	2013_03_30
LUAD	2013_02_07
	2012_11_15
PANCAN8	2012_08_25
SKCM	2013_01_16
	2012_12_21
STAD	2013_04_17
THCA	2013_03_18
	2012_10_24

TCGA AWG analyses for Lower Grade Glioma: 2013_04_06

Maintained by TCGA GDAC Team *(Broad Institute/MD Anderson Cancer Center/Harvard Medical School)*

Unique Tumor Sample Counts

Tumor	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
LGG	247	208	220	0	217	27	220	0	221	0	217

Download run results with firehose_get

Example download command: ***firehose_get awg_lgg 2013_04_06***

For More Help: ***firehose_get --help***

Overview

Introduction

Summary

Note: These results are offered to the community as an additional reference point, enabling a wide range of cancer biologists, clinical investigators, and genome and computational scientists to easily incorporate TCGA into the backdrop of ongoing research. While every effort is made to ensure that Firehose input data and algorithms are of the highest possible quality, these analyses have not been reviewed by domain experts.

A total of 121 reports are available for analysis run "06 April 2013".

Results

Cancer Types

Table 1. Click "Browse" to view reports for a cancer type of interest. If you prefer to view reports on your own computer, you may download a ZIP archive containing all reports for a cancer type by clicking "Download".

Cancer Type	Cohort	Reports	HTML	ZIP
Brain Lower Grade Glioma	LGG-IDH1_IDH2_mutant_1p_19q_co-deleted	30	Browse	Download
Brain Lower Grade Glioma	LGG-IDH1_IDH2_mutant_1p_19q_intact	31	Browse	Download
Brain Lower Grade Glioma	LGG-IDH_WT	27	Browse	Download
Brain Lower Grade Glioma	LGG-TP	33	Browse	Download

Example

121 analyses reports

Across 4 cohort subsets

2-3 days turnaround

gdac.broadinstitute.org/runs/awg_lgg__2013_04_06

Custom Google-Powered Search Engine

gistic thyroid

Search Results

Thyroid Adenocarcinoma: Correlation between copy number ...
Mar 13, 2013 ... Testing the association between copy number variation of 19 ...
gdac.broadinstitute.org/runs/analyses__latest/.../nozzle.html

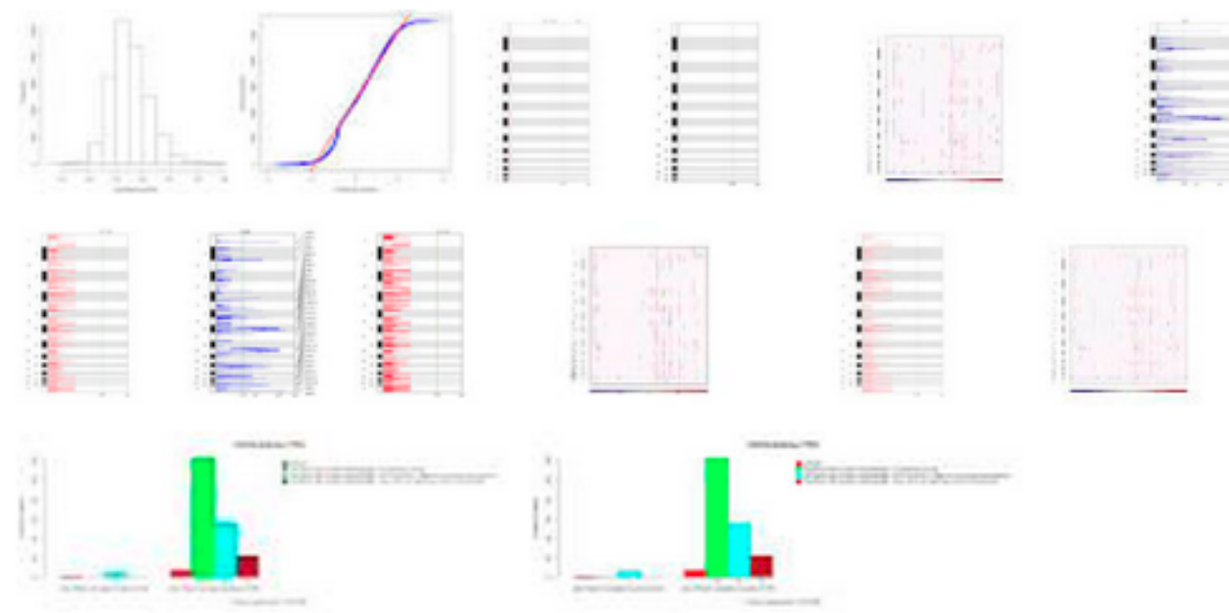
Thyroid Adenocarcinoma: Clustering of copy number data ...
Mar 13, 2013 ... *Thyroid* Adenocarcinoma: Clustering of copy number data: consen
NMF .... The all lesions file is from *GISTIC* pipeline and summarizes the results fro
...
gdac.broadinstitute.org/runs/analyses__latest/.../nozzle.html

Thyroid Adenocarcinoma: Copy number analysis (GISTIC2)
Mar 12, 2013 ... *Thyroid* Adenocarcinoma: Copy number analysis (GIST
(primary solid tumor cohort). Maintained by Dan DiCara (Broad Institu
Overview. Introduction ...
gdac.broadinstitute.org/runs/analyses...TP/.../**gistic**2/nozzle.html

Thyroid Adenocarcinoma: Correlations between copy number and ...
Feb 21, 2013 ... *Thyroid* Adenocarcinoma: Correlations between copy number and
gene derived by *GISTIC* pipelinePearson correlation coefficients were calculated fo
each ...
gdac.broadinstitute.org/runs/analyses__2013.../nozzle.html

WEB IMAGE

About 13 results (0.35 seconds)



Streamline extraction of meaning from TCGA data & Firehose analyses

Digital Object Identifiers (DOIs)

Analysis Overview

Ovarian Serous Cystadenocarcinoma (Primary solid tumor)

21 April 2013 | analyses__2013_04_21 Maintainer Information Citation Information doi:10.7908/C1BV7DK1

- Overview
- Introduction
- Summary

Note: These results are offered to the community as an additional reference point, enabling a wide range of cancer biologists, clinical investigators, and genome and computational scientists to easily incorporate TCGA into the backdrop of ongoing research. While every effort is made to ensure that Firehose input data and algorithms are of the highest possible quality, these analyses have not been reviewed by domain experts.

- Results
  - *Sequence and Copy Number Analyses*
    - **Copy number analysis (GISTIC2)**
      View Report | There were 569 tumor samples used in this analysis: 32 significant arm-level results, 32 significant focal amplifications, and 37 significant focal deletions were found.
    - **Mutation Analysis (MutSig v1.5)**
      View Report |
    - **Mutation Analysis (MutSig v2.0)**
      View Report |
    - **Mutation Analysis (MutSigCV v0.9)**
      View Report |

Analysis Overview

Ovarian Serous Cystadenocarcinoma (Primary solid tumor)

21 April 2013 | analyses__2013_04_21 Maintainer Information Citation Information doi:10.7908/C1BV7DK1

Maintained by TCGA GDAC Team (Broad Institute/MD Anderson Cancer Center/Harvard Medical School)

Copy number analysis (GISTIC2)

Ovarian Serous Cystadenocarcinoma (Primary solid tumor)

21 April 2013 | analyses__2013_04_21 Maintainer Information Citation Information doi:10.7908/C1CZ3544

Cite as Broad Institute TCGA Genome Data Analysis Center (2013): Ovarian Serous Cystadenocarcinoma (Primary solid tumor cohort) - 21 April 2013: Copy number analysis (GISTIC2). Broad Institute of MIT and Harvard doi:10.7908/C1CZ3544

Hundreds of reports generated per month, citable directly in literature
First of its kind at Broad Institute: nothing at this scale, anywhere?

High Resolution Sample Provenance

2013_04_21 analyses Run

Tables of Ingested Data: HTML PNG TSV Samples Summary: Report

AnalysisReport	# Pipelines	% Successful	Download	
BLCA	59	100%	Open	Protected
BRCA	76	100%	Open	Protected
CESC	56	100%	Open	Protected
COADREAD	76	100%	Open	Protected
COAD	76	100%	Open	Protected
DLBC	10	100%	Open	Protected
ESCA	9	100%	Open	Protected
GBM	77	100%	Open	Protected
HNSC	59	100%	Open	Protected
KICH	28	100%	Open	Protected
KIRC	76	100%	Open	Protected
KIRP	73	100		
LGG	73	100		
LIHC	34	100		
LUAD	76	100		
LUSC	76	100		
OV	81	100		
PRAD	56	100		
READ	76	100		
SARC	13	100		
SKCM	59	100		
STAD	56	100%	Open	Protected
THCA	59	100%	Open	Protected
UCEC	76	100%	Open	Protected
LAML	55	98%	Open	Protected
PANCAN12	14	61%	Open	Protected

Overview

Introduction

Summary

There were 70 redacted samples, 2787 replicate aliquots, and 155 blacklisted aliquots. The table below represents the sam
counts for those samples that were ingested into firehose after filtering out redactions, replicates, and blacklisted data.

GET FULL TAB

Table 1. Summary of TCGA Tumor Data. Click on a tumor type to display a tumor type specific Samples Summary Report.

Tumor	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNASeq	miR	miRSeq	RPPA	MAF
BLCA	171	135	153	105	153	0	122	0	150	54	28
BRCA	967	878	927	0	928	526	842	0	892	408	772
CESC	144	40	129	0	134	0	116	0	122	0	39
COAD	423	423	422	69	420	153	364	0	407	269	155
COADREAD	592	591	586	104	582	222	488	0	550	399	224
							0	0	16	0	0
							0	0	0	0	0
							160	491	0	214	291
							304	0	356	212	306
							66	0	66	0	65
							480	0	481	454	293
							76	0	117	0	111
							179	0	187	0	197
							220	0	221	0	217
LIHC	126	73	99	0	98	0	69	0	96	0	0
LUAD	563	376	474	0	533	32	353	0	401	237	229
LUSC	494	327	398	0	385	154	261	0	349	195	178
OV	592	580	579	0	584	574	296	570	453	412	316
PAAD	73	1	57	0	49	0	40	0	34	0	34
PANCAN12	5591	4936	5248	423	5074	2176	3857	1061	4306	2785	3082
PRAD	200	156	188	0	188	0	176	0	177	0	83
READ	169	168	164	35	162	69	124	0	143	130	69
SARC	73	18	51	0	52	0	0	0	29	0	0
SKCM	336	191	288	119	316	0	265	0	272	164	253
STAD	308	178	308	0	308	0	43	0	237	0	116
THCA	500	318	473	94	500	0	461	0	426	224	323
UCEC	525	455	498	106	500	54	372	0	487	200	248
Totals	7916	6244	7333	636	7195	2219	5389	1061	6119	3173	4323

Linked to every dashboard

Raises bar for clarity, comprehensiveness & ease of access for data resolution in TCGA

Lung Adenocarcinoma (LUAD) Samples Summary Report

Overview

Introduction

Summary

There were 1 redacted samples, 126 replicate aliquots, and 6 blacklisted aliquots. The table below represents the sample counts for those samples that were ingested into firehose after filtering out redactions, replicates, and blacklisted data.

Table 1. Summary of TCGA Tumor Data.

Tumor	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNASeq	miR	miRSeq	RPPA
LUAD	563	376	474	0	533	32	353	0	401	237

Results

Ingested Samples

This section includes a more granular look at the samples ingested into Firehose. A sample counts table is provided type (e.g. Primary Solid Tumor, Recurrent Solid Tumor, Normal Blood, etc.). Furthermore, each count is a link to a breakdown of the samples and their specific details (e.g. platform, sequencing center, etc.) The following platforms included in the counts depicted in the table below.

- Agilent SurePrint G3 Human CGH Microarray Kit 1x1M
- Agilent Human Genome CGH Microarray 244A
- Agilent Human Genome CGH Custom Microarray 2x415K
- Affymetrix Human Exon 1.0 ST Array
- Illumina DNA Methylation OMA002 Cancer Panel I
- Illumina DNA Methylation OMA003 Cancer Panel I
- Illumina Human1M-Duo BeadChip
- Illumina 550K Infinium HumanHap550 SNP Chip

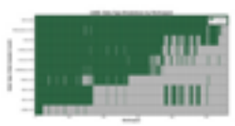
The sample type short letter codes in the table below are defined in the following list.

- TP: Primary solid Tumor
- TR: Recurrent Solid Tumor
- TB: Primary Blood Derived Cancer - Peripheral Blood
- TM: Metastatic
- TAM: Additional Metastatic
- NB: Blood Derived Normal
- NT: Solid Tissue Normal

Table 2. Click on any sample count to display a table detailing all the samples that comprise that count. Please note, there are usually mu are typically many more rows than the count implies.

Sample Type	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNASeq	miR	miRSeq	RI
TP	563	376	474	0	533	32	353	0	401	2
TR	2	2	2	0	2	0	2	0	2	0
NB	383	284	369	0	0	0	0	0	0	0
NT	173	159	165	0	56	0	57	0	46	0
Totals	563	376	474	0	533	32	353	0	401	23

Figure 1. This figure depicts the distribution of available data on a per participant basis.



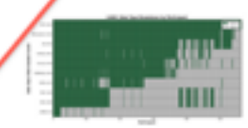
Redactions

Table 2. Click on any sample count to display a table detailing all the samples that comprise that count. 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
TP	563	376	474	0	533	32	353	0	401	237	229
TR	2	2	2	0	2	0	2	0	2	0	0
NB	383	284	369	0	0	0	0	0	0	0	0
NT	173	159	165	0	56	0	57	0	46	0	0
Totals	563	376	474	0	533	32	353	0	401	237	229

GET HIGH-RES IMAGE

Figure 1. This figure depicts the distribution of available data on a per participant basis.



Redactions

Replicate Filtered Samples

Blacklisted Samples

Methods & Data

MADE WITH NOZZLE

LUAD Primary solid Tumor mRNA Data

Table S5.

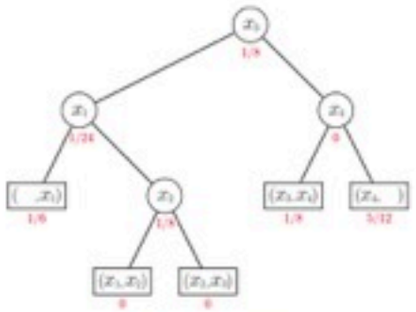
TCGA Barcode	Platform	Center	Data Level	Protocol
TCGA-05-4244-01A-01R-1107-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	2	unc_lowess_normalization_probe_level
TCGA-05-4244-01A-01R-1107-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	3	unc_lowess_normalization_gene_level
TCGA-05-4249-01A-01R-1107-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	2	unc_lowess_normalization_probe_level
TCGA-05-4249-01A-01R-1107-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	3	unc_lowess_normalization_gene_level
TCGA-05-4250-01A-01R-1107-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	2	unc_lowess_normalization_probe_level
TCGA-05-4250-01A-01R-1107-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	3	unc_lowess_normalization_gene_level
TCGA-35-3615-01A-01R-0946-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	2	unc_lowess_normalization_probe_level
TCGA-35-3615-01A-01R-0946-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	3	unc_lowess_normalization_gene_level
TCGA-35-4122-01A-01R-1107-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	2	unc_lowess_normalization_probe_level
TCGA-35-4122-01A-01R-1107-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	3	unc_lowess_normalization_gene_level
TCGA-35-4123-01A-01R-1107-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	2	unc_lowess_normalization_probe_level
TCGA-35-4123-01A-01R-1107-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	3	unc_lowess_normalization_gene_level
TCGA-44-2655-01A-01R-0946-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	2	unc_lowess_normalization_probe_level
TCGA-44-2655-01A-01R-0946-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	3	unc_lowess_normalization_gene_level
TCGA-44-2656-01A-02R-0946-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	2	unc_lowess_normalization_probe_level
TCGA-44-2656-01A-02R-0946-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	3	unc_lowess_normalization_gene_level
TCGA-44-2657-01A-01R-1107-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	2	unc_lowess_normalization_probe_level
TCGA-44-2657-01A-01R-1107-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	3	unc_lowess_normalization_gene_level
TCGA-44-2659-01A-01R-0946-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	2	unc_lowess_normalization_probe_level
TCGA-44-2659-01A-01R-0946-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	3	unc_lowess_normalization_gene_level
TCGA-44-2661-01A-01R-1107-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	2	unc_lowess_normalization_probe_level
TCGA-44-2661-01A-01R-1107-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	3	unc_lowess_normalization_gene_level
TCGA-44-2662-01A-01R-0946-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	2	unc_lowess_normalization_probe_level
TCGA-44-2662-01A-01R-0946-07	Agilent 244K Custom Gene Expression G4502A-07-3	University of North Carolina	3	unc_lowess_normalization_gene_level

ReproD™

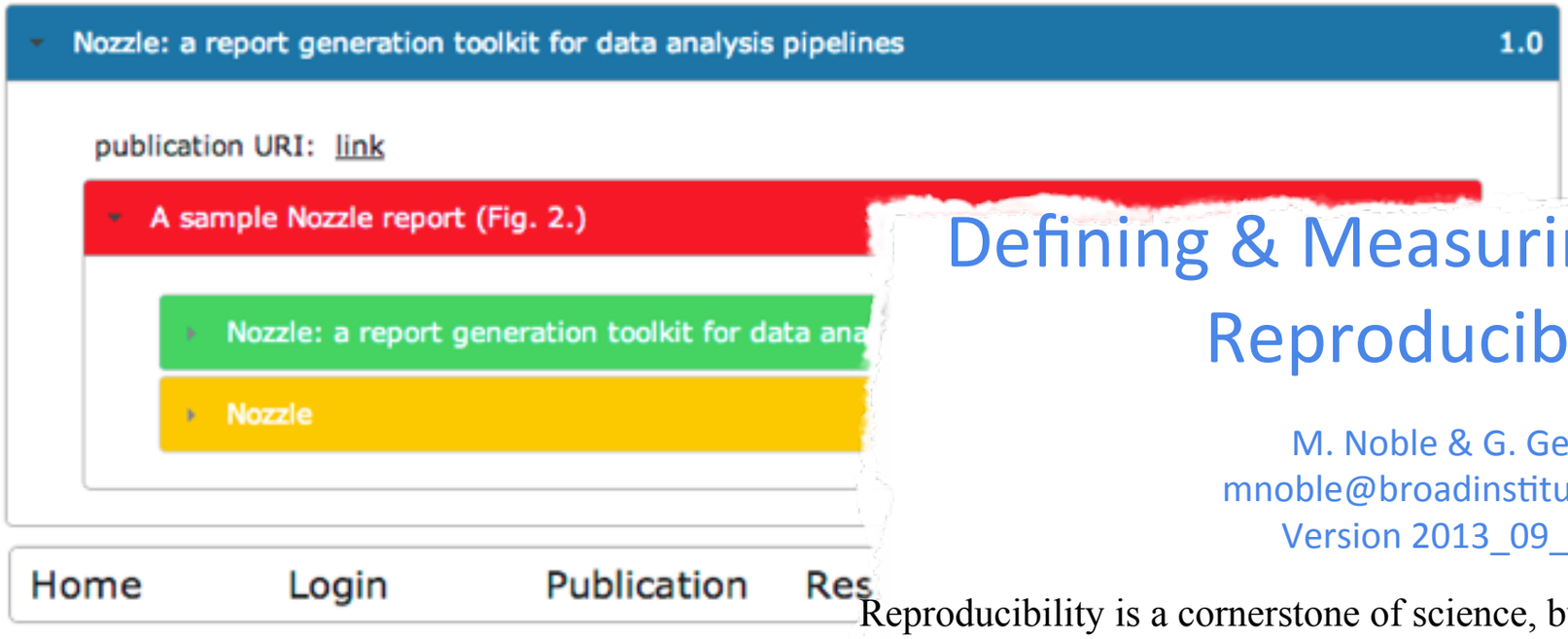
The Reproducibility Dashboard v0.05 beta

Publications

Data



Software Methods



Defining & Measuring Research Reproducibility

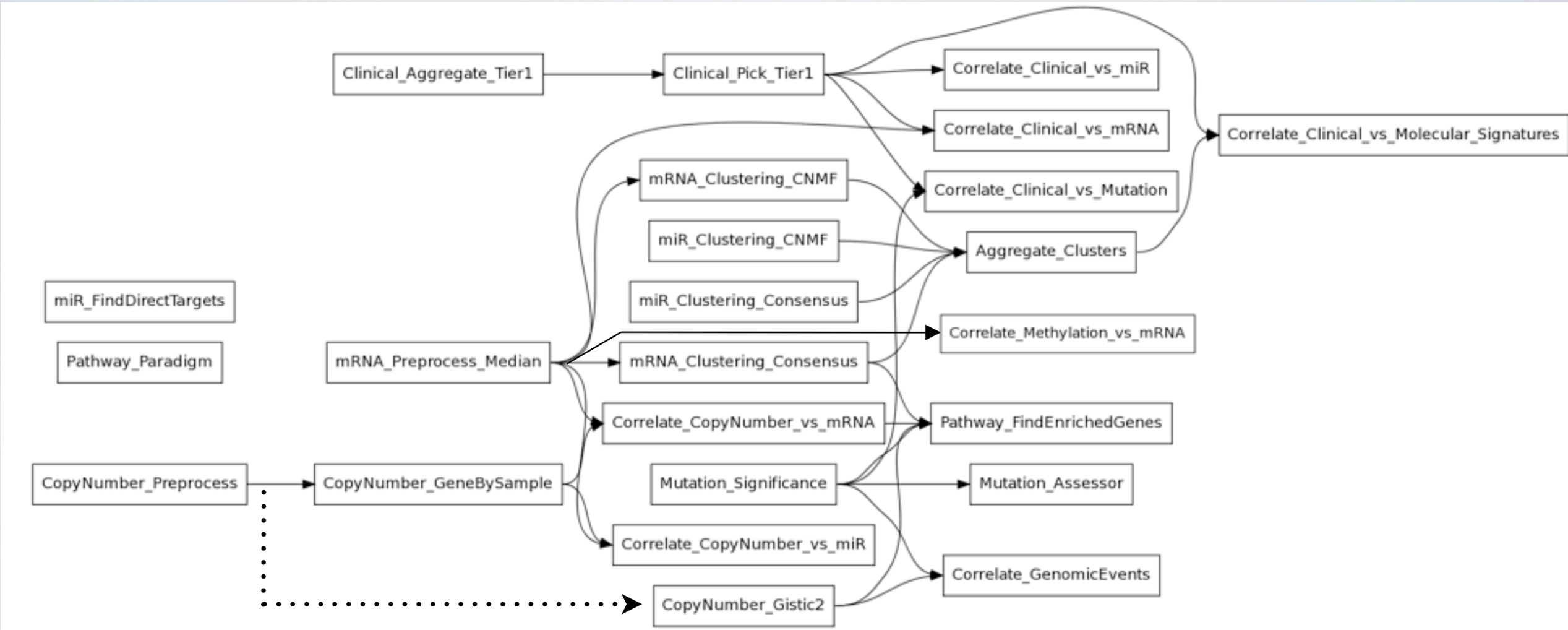
M. Noble & G. Getz
mnoble@broadinstitute.org
Version 2013_09_23

Reproducibility is a cornerstone of science, but its widespread practice remains stubbornly inconsistent. Time consuming and complex, for both producers and verifiers of results, reproducibility demands attention that might otherwise be applied towards novelty & innovation.

4. Observations

Observation 1

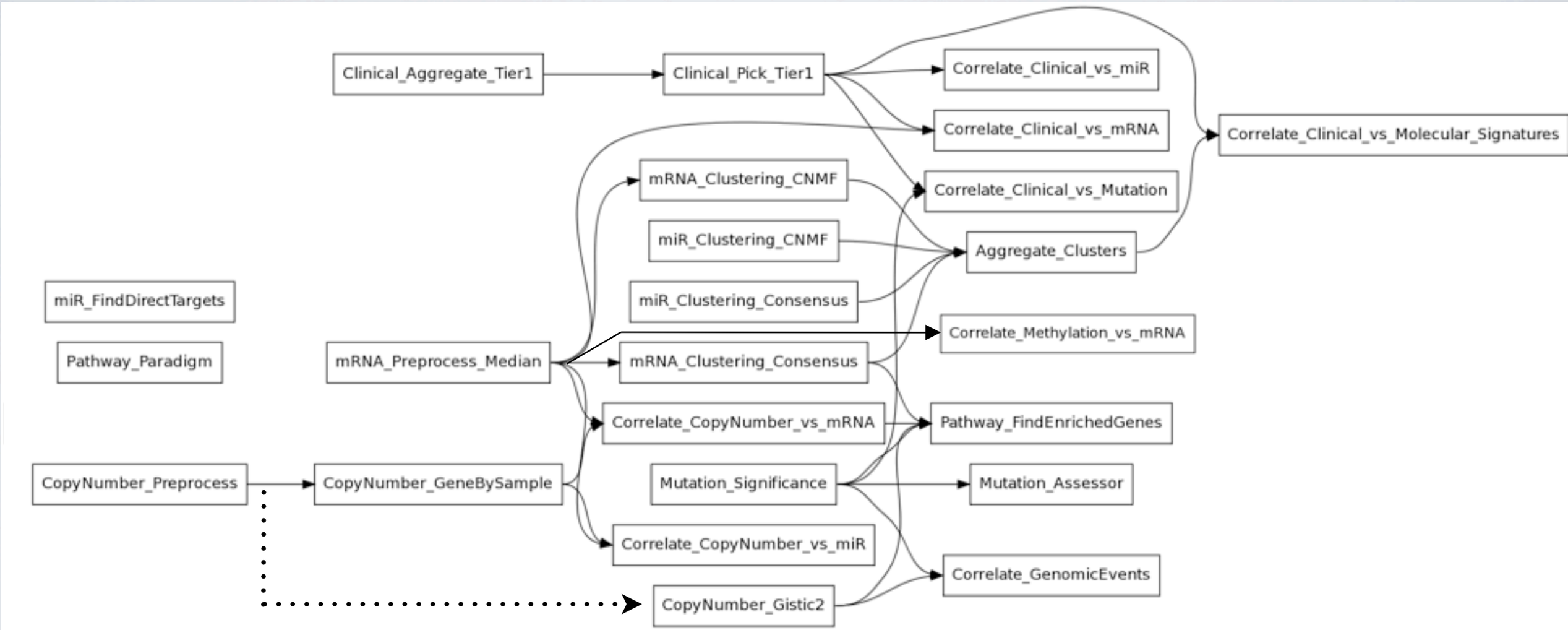
This workflow ...



Hundreds of tasks & modules, ***per disease***

Observation 1

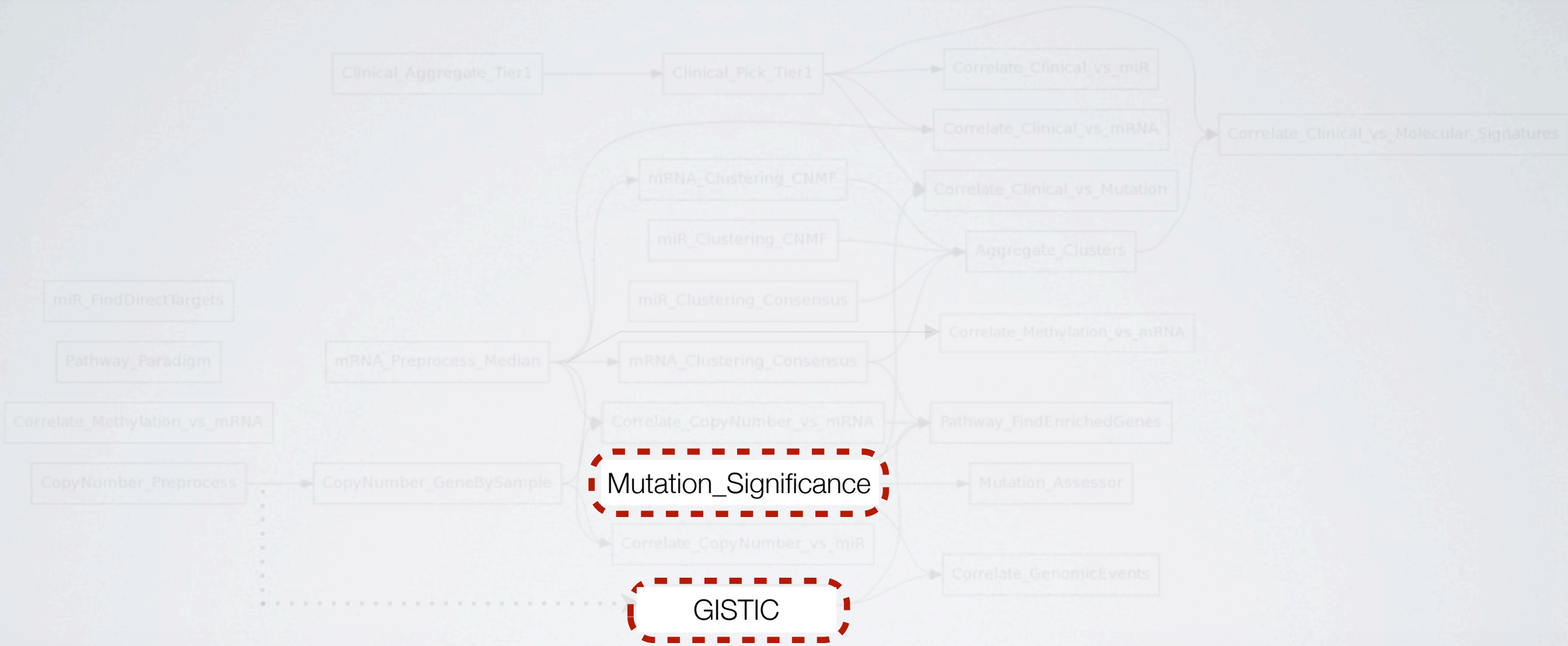
This workflow ... is really a <u>META-pipeline</u> of pipelines



Hundreds of tasks & modules, ***per disease***

Observation 1

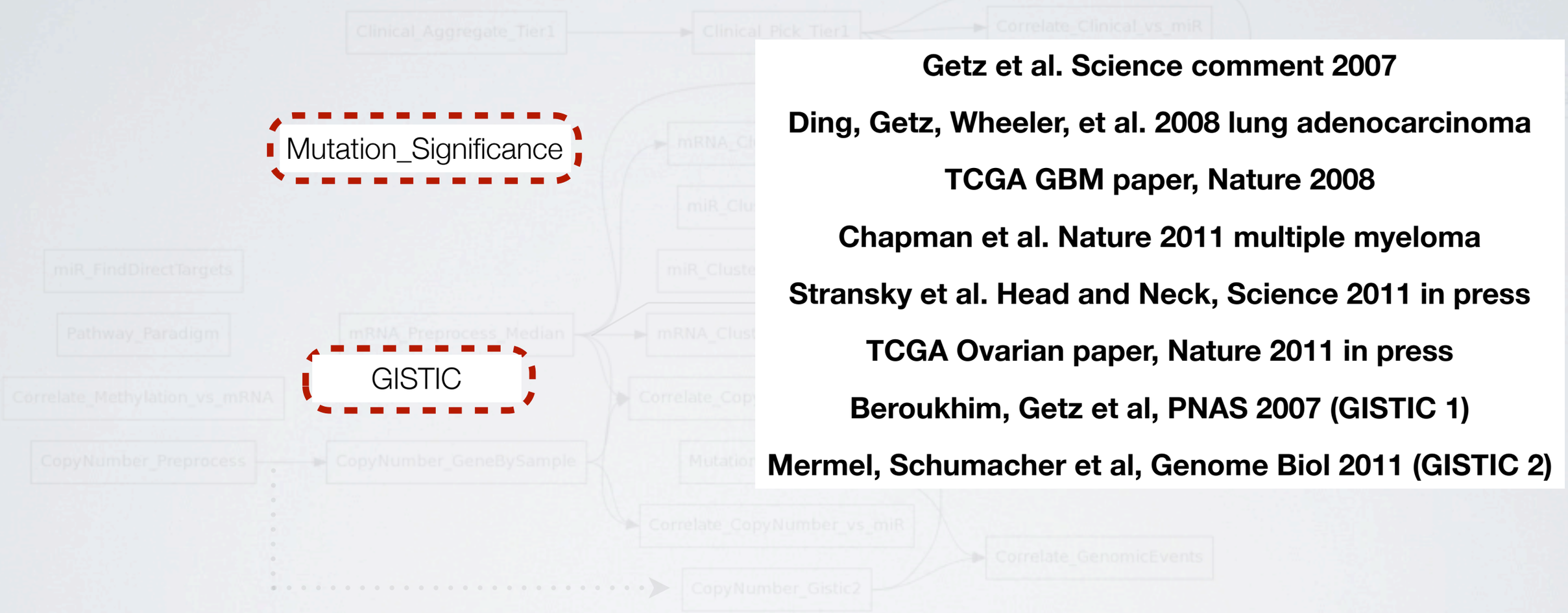
This workflow ... is really a <u>META-pipeline</u> of pipelines



Some of which are themselves complex pipelined codes.

Observation 1

This workflow ... is really a <u>META-pipeline</u> of pipelines



Getz et al. Science comment 2007

Ding, Getz, Wheeler, et al. 2008 lung adenocarcinoma

TCGA GBM paper, Nature 2008

Chapman et al. Nature 2011 multiple myeloma

Stransky et al. Head and Neck, Science 2011 in press

TCGA Ovarian paper, Nature 2011 in press

Beroukhim, Getz et al, PNAS 2007 (GISTIC 1)

Mermel, Schumacher et al, Genome Biol 2011 (GISTIC 2)

Some of which are themselves complex pipelined codes.

<u>Continuously evolving</u> through years of publication use.

Like ENIAC, no simple task
to keep it all running

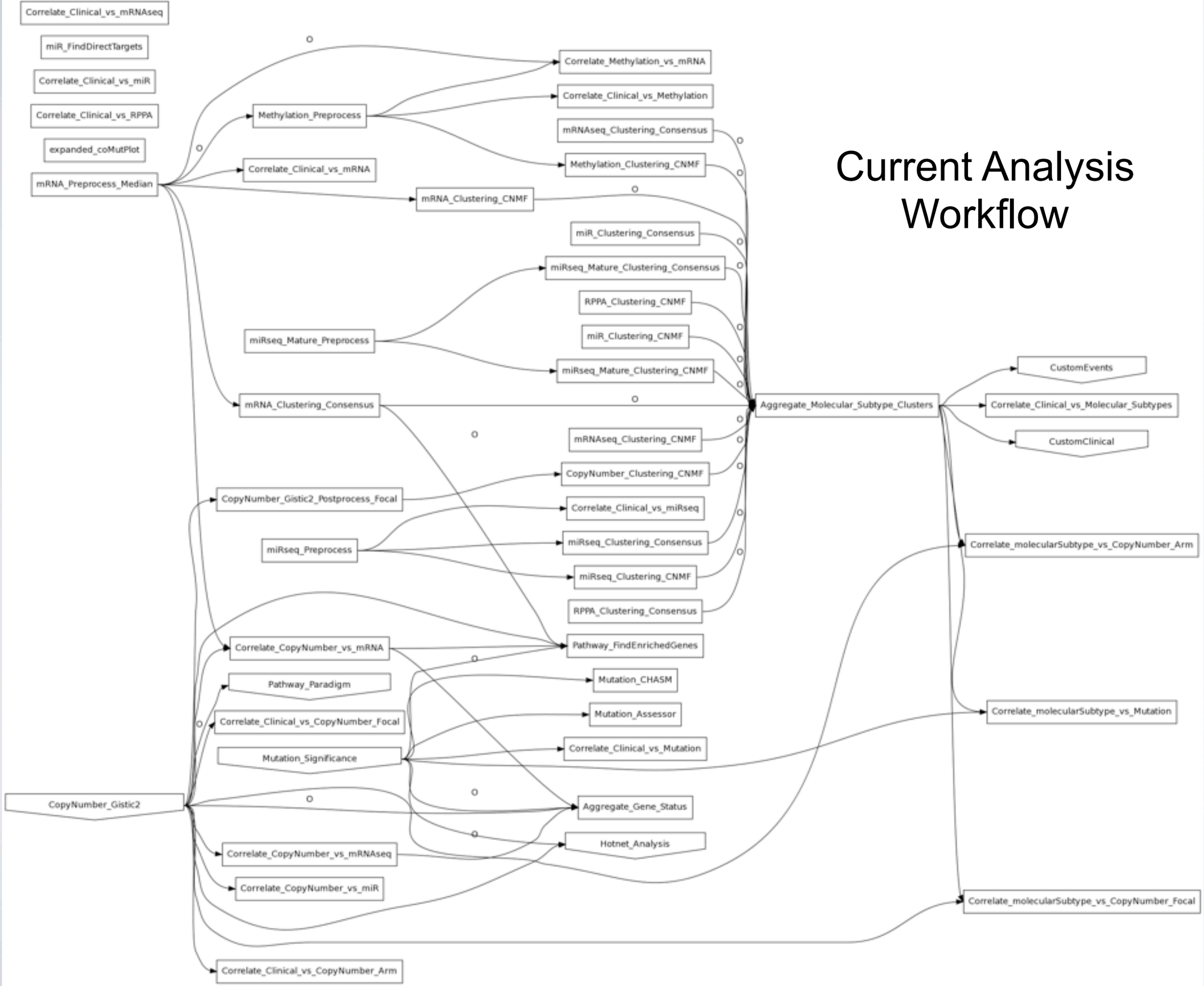
... in part because ...

A Tale of Two Coders

Software Engineer	Comp Bio / Researcher
Careful, deliberate design Towards production deployment Must be fastidious	Exploratory, open-ended analysis Towards publication Can be messy

Overlapping, But Not Identical, Aims

Current Analysis Workflow



Observation 2: Unit Testing Not Enough

Individual researcher invoking THEIR code against THEIR data for THEIR paper, to establish that, in isolation, it runs to completion.

Observation 2: Unit Testing Not Enough

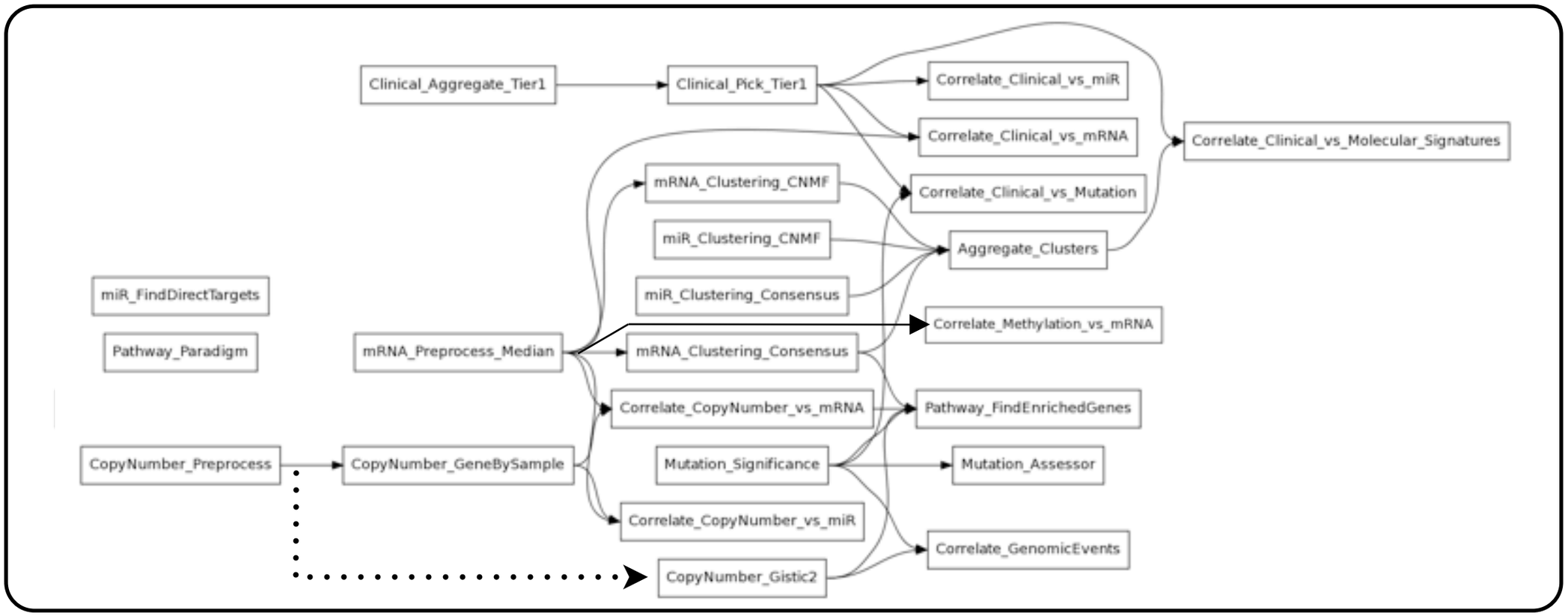
Individual researcher invoking THEIR code against THEIR data for THEIR paper, to establish that, in isolation, it runs to completion.

<u>INTEGRATION TESTING</u> must establish that (changes to) codes plays nice with rest of system.

VITAL to maintain production operation of Firehose “data factory”

Observation 2: Unit Testing Not Enough

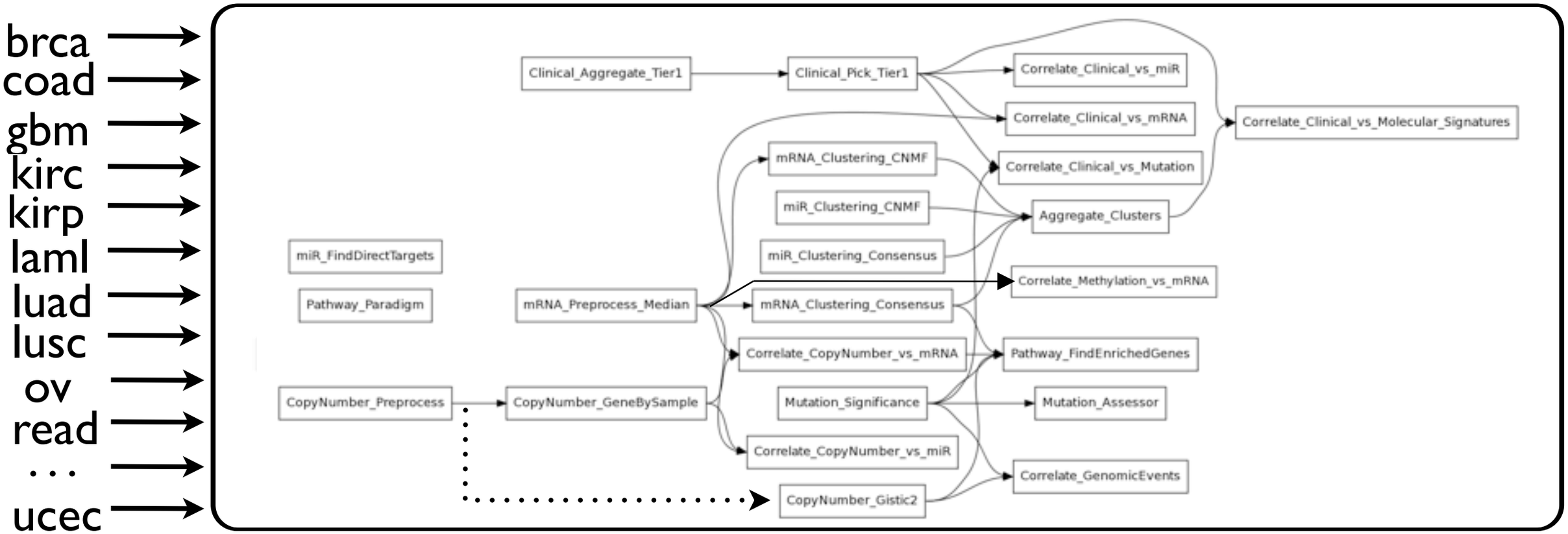
Individual researcher invoking THEIR code against THEIR data for THEIR paper, to establish that, in isolation, it runs to completion.



<u>INTEGRATION TESTING</u> must establish that (changes to) codes plays nice with rest of system.

Observation 2: Unit Testing Not Enough

Individual researcher invoking THEIR code against THEIR data for THEIR paper, to establish that, in isolation, it runs to completion.

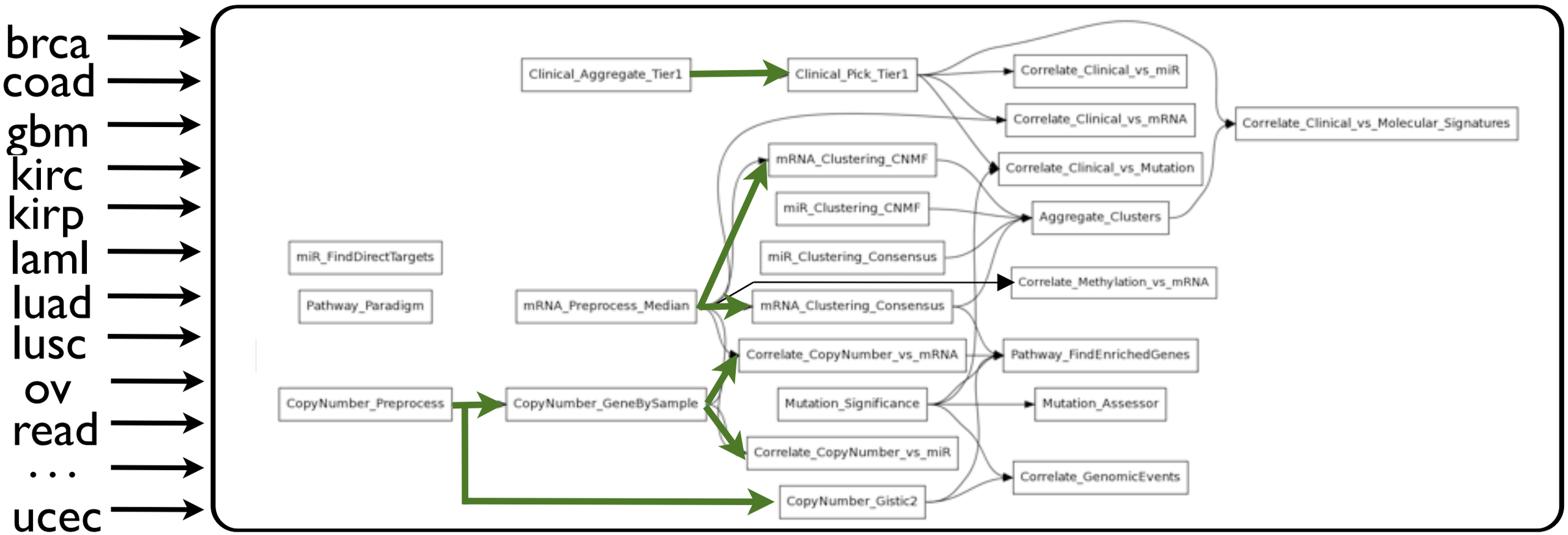


<u>INTEGRATION TESTING</u> must establish that (changes to) codes plays nice with rest of system.

Across datasets

Observation 2: Unit Testing Not Enough

Individual researcher invoking THEIR code against THEIR data for THEIR paper, to establish that, in isolation, it runs to completion.



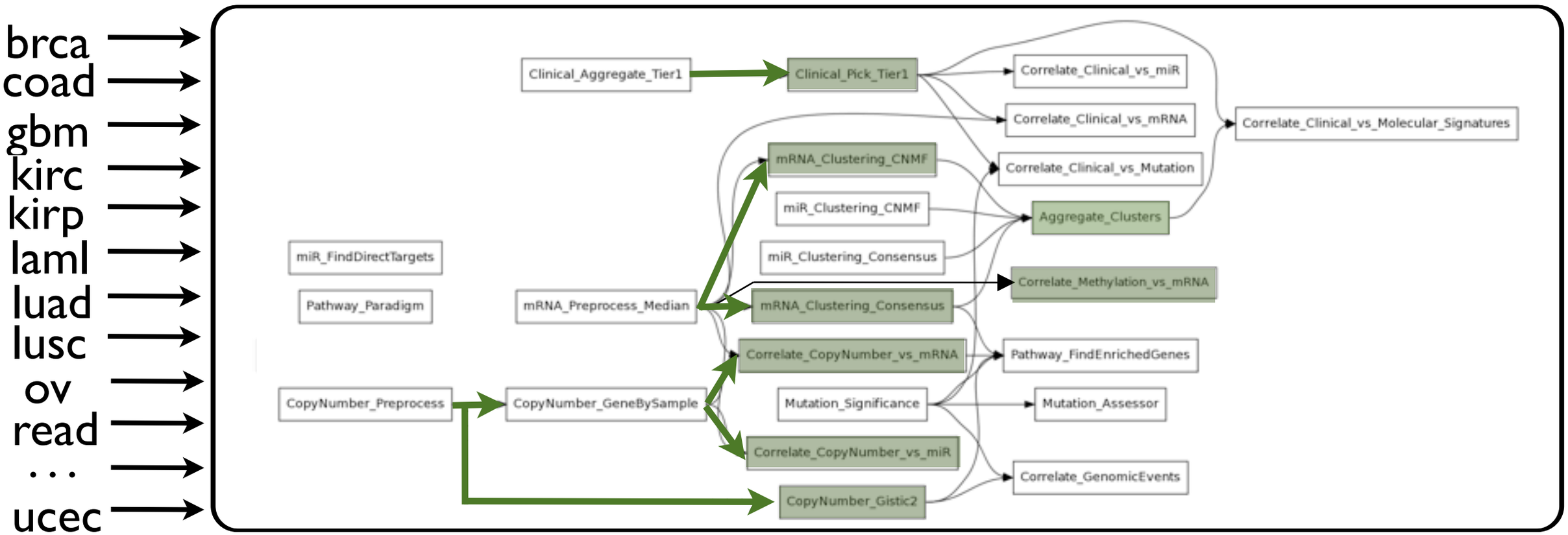
<u>INTEGRATION TESTING</u> must establish that (changes to) codes plays nice with rest of system.

Across datasets

With O's correctly wired to I's

Observation 2: Unit Testing Not Enough

Individual researcher invoking THEIR code against THEIR data for THEIR paper, to establish that, in isolation, it runs to completion.



<u>INTEGRATION TESTING</u> must establish that (changes to) codes plays nice with rest of system.

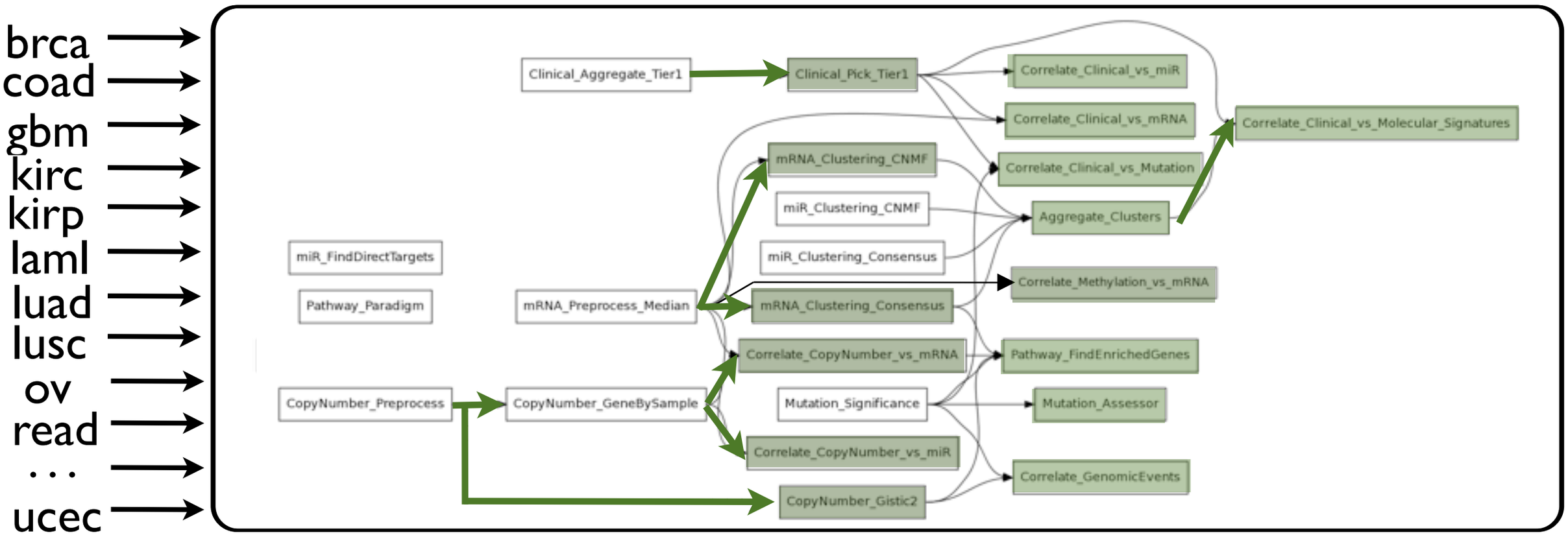
Across datasets

Downstream dependents *correctly read* outputs

With O's correctly wired to I's

Observation 2: Unit Testing Not Enough

Individual researcher invoking THEIR code against THEIR data for THEIR paper, to establish that, in isolation, it runs to completion.



<u>INTEGRATION TESTING</u> must establish that (changes to) codes plays nice with rest of system.

Across datasets	Downstream dependents *correctly read* outputs
With O's correctly wired to I's	And remainder of workflow runs to completion

Observation 3

Versioning & Automation are sacrosanct

Observation 3

Versioning & Automation are sacrosanct

- Otherwise no reproducibility

Observation 3

Versioning & Automation are sacrosanct

- Otherwise no reproducibility
- Or *algorithmic scalability*

Observation 3

Versioning & Automation are sacrosanct

- Otherwise no reproducibility
- Or *algorithmic scalability*
- BOTH code AND data are versioned
- Do not trust: version and verify

} **Babel problem**

Observation 3

Versioning & Automation are sacrosanct

- Otherwise no reproducibility
- Or *algorithmic scalability*
- BOTH code AND data are versioned
- Do not trust: version and verify

} **Babel problem**

- Automation not just of pipelines:

Observation 3

Versioning & Automation are sacrosanct

- Otherwise no reproducibility
- Or *algorithmic scalability*
- BOTH code AND data are versioned
- Do not trust: version and verify

} **Babel problem**

- Automation not just of pipelines:
  - ✓ but also tools used to create them

FH web services Hydrant

Observation 3

Versioning & Automation are sacrosanct

- Otherwise no reproducibility
- Or *algorithmic scalability*
- BOTH code AND data are versioned
- Do not trust: version and verify

} **Babel problem**

- Automation not just of pipelines:
  - ✓ but also tools used to create them — **FH web services Hydrant**
  - ✓ and reports generated from them — **GDAC website**

Observation 3

Versioning & Automation are sacrosanct

- Otherwise no reproducibility
- Or *algorithmic scalability*
- BOTH code AND data are versioned
- Do not trust: version and verify

} **Babel problem**

- Automation not just of pipelines:
  - ✓ but also tools used to create them — **FH web services Hydrant**
  - ✓ and reports generated from them — **GDAC website**
  - ✓ and data sources which feed them — **DCC, dbGAP**

GUIs alone ARE NOT GOOD ENOUGH for these latter tasks

Because *PROCESS SCALABILITY* matters too

Observation 4: A- not good enough

Suppose all TCGA moving parts run 90% efficient

Observation 4: A- not good enough

Suppose all TCGA moving parts run 90% efficient

After just 4 steps in life of TCGA sample:	$.9^4 = 66\%$ overall efficiency
Assume A = 95%	$.95^4 = 81\%$
And $A^+ = 99\%$	$.99^4 = 96\%$

Observation 5

Given that TCGA arguably largest/richest cancer data ever assembled

Observation 5

Given that TCGA arguably largest/richest cancer data ever assembled

Novel discoveries lurk in Firehose outputs

Observation 5

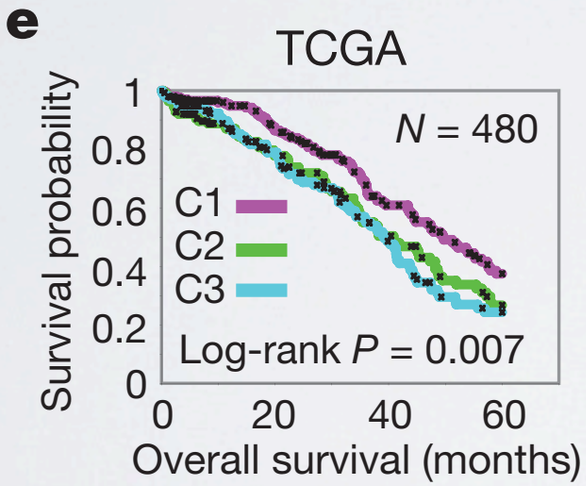
Given that TCGA arguably largest/richest cancer data ever assembled

d

miRNA cluster	Gene cluster D	I	M	P
C1	55	48	15	89
C2	40	21	51	29
C3	39	37	43	20

CNMF clustering of Ovarian miR expression yielded 3 subtypes

e



One of which correlated to significantly longer survivability

Integrated genomic analyses of ovarian carcinoma
TCGA Network, Nature, 2011

Novel discoveries lurk in Firehose outputs

Observation 5

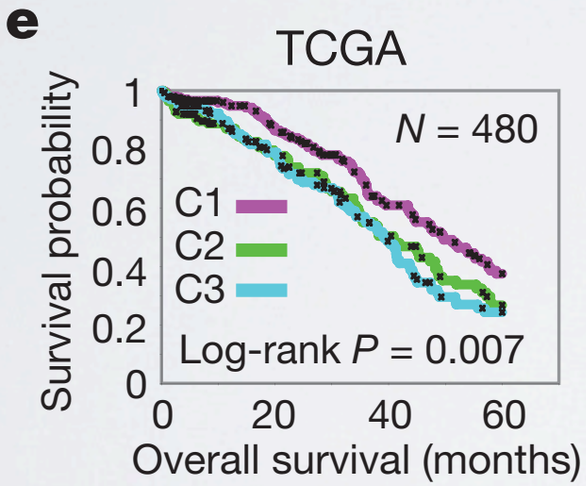
Given that TCGA arguably largest/richest cancer data ever assembled

d

miRNA cluster	Gene cluster D	I	M	P
C1	55	48	15	89
C2	40	21	51	29
C3	39	37	43	20

CNMF clustering of Ovarian miR expression yielded 3 subtypes

e



One of which correlated to significantly longer survivability

Integrated genomic analyses of ovarian carcinoma
TCGA Network, Nature, 2011

Novel discoveries lurk in Firehose outputs

∴ Firehose for active research: low-hanging results waiting to be plucked

Points to Potential Clinical Gold Mine ...

Firehose automatically mines entire suite of clinical params to identify statistically significant relationships with every TCGA datatype or aggregate (e.g. clusters)

Points to Potential Clinical Gold Mine ...

Firehose automatically mines entire suite of clinical params to identify statistically significant relationships with every TCGA datatype or aggregate (e.g. clusters)

The results, which e.g. include survival curves (when possible) for every TCGA disease, are posted openly on the Broad GDAC site in the form of biologist-friendly HTML reports

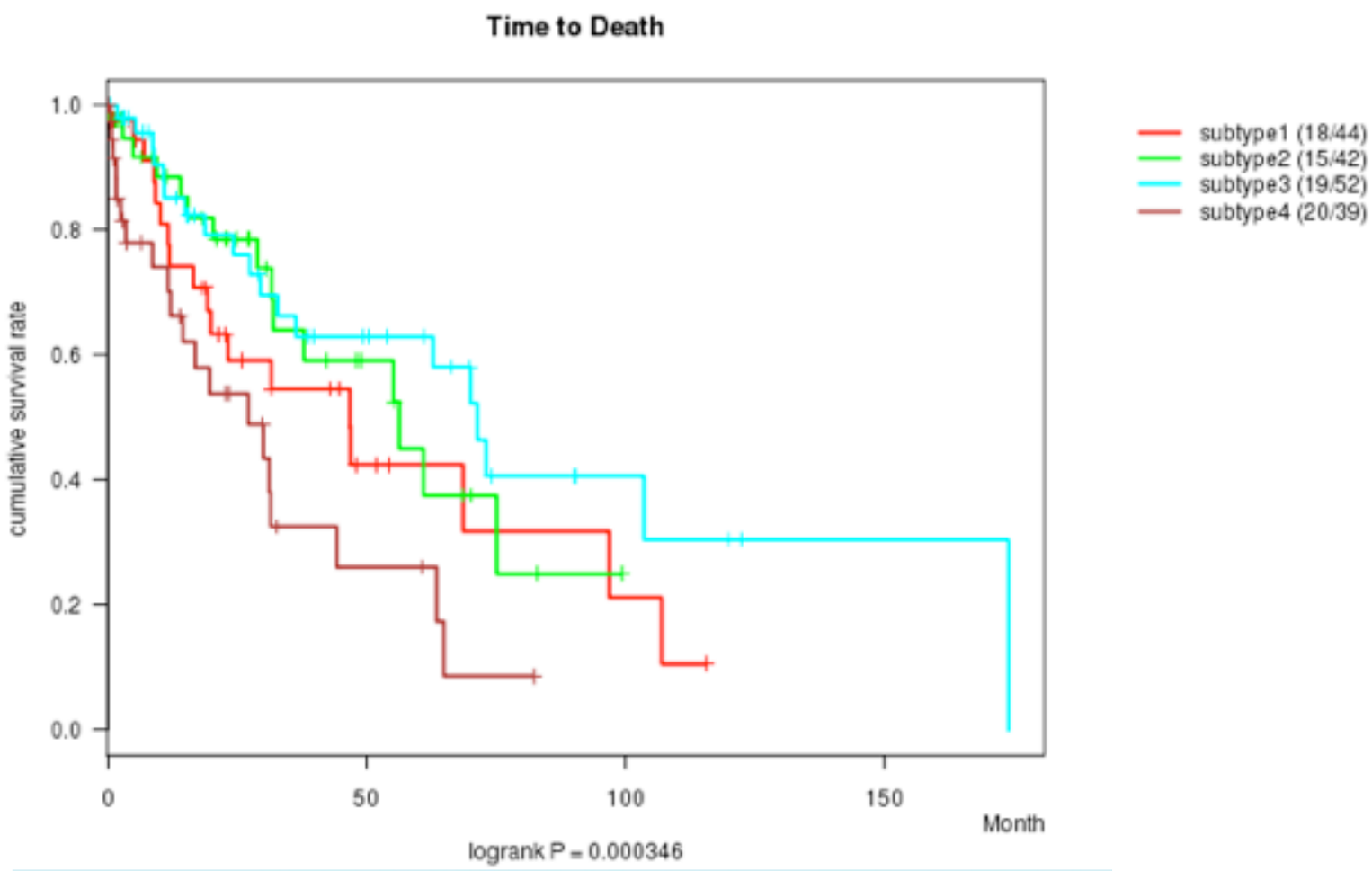
Points to Potential Clinical Gold Mine ...

Firehose automatically mines entire suite of clinical params to identify statistically significant relationships with every TCGA datatype or aggregate (e.g. clusters)

The results, which e.g. include survival curves (when possible) for every TCGA disease, are posted openly on the Broad GDAC site in the form of biologist-friendly HTML reports

Since automation is “free,” these don’t have to be 100% to establish potentially interesting signposts

Lung Squamous



	nPatients	nDeath	Duration Range (Median), Month
ALL	**177**	**72**	**0.0 - 173.8 (16.6)**
subtype1	44	18	0.2 - 115.6 (14.3)
subtype2	42	15	0.2 - 99.2 (23.0)
subtype3	52	19	0.0 - 173.8 (17.8)
subtype4	39	20	0.1 - 82.2 (8.8)

2012_09_13 Analyses

'RPPA cHierClus subtypes' versus *'Time to Death'*

P value = *0.000346* (logrank test)

Much more low-hanging fruit, lurking in wait for set of willing eyes

Summary

FIREHOSE
Broad GDAC

Simplifying & Systematizing Science at Unprecedented Scales & Complexity

Fin