



Firehose : The First Year

Michael S. Noble
The Broad Institute of MIT & Harvard

1st TCGA Symposium
Washington, D.C.

November 17, 2011



Acknowledgements

PI: Lynda Chin, Gaddy Getz

Broad

Michael Noble

Douglas Voet

Gordon Saksena

Kristian Cibulskis

Rui Jing

Michael Lawrence

Pei Lin

Aaron McKenna

Andrey Sivachenko

Carrie Sougnez

Petar Stojanov

Lihua Zhou

Lee Lichtenstein

Robert Zupko

Dan DiCara

Raktim Sinha

Belfler-DFCI

Yonghong Xiao

Juinhua Zhang

Spring Liu

Sachet Shukla

Hailei Zhang

Terrence Wu

IGV & GenePattern teams @ Broad

Jill Mesirov

Michael Reich

Peter Carr

Marc-Danie Nazaire

Jim Robinson

Helga Thorvaldsdottir

Harvard

Peter Park

Nils Gehlenborg

Semin Lee

Richard Park

Matthew Meyerson

Todd Golub

Eric Lander



OUTLINE

- I. Why (yet another pipeline)?
- II. What (is Firehose, anyway)?
- III. How (will it help)?
- IV. Insights (gained so far)

I : WHY?

TCGA

ACRONYM: THE CANCER GENOME ATLAS

TCGA

ACRONYM: THE CANCER GENOME ATLAS

SYNONYM: FLOOD (OF DATA & ALGORITHMS)

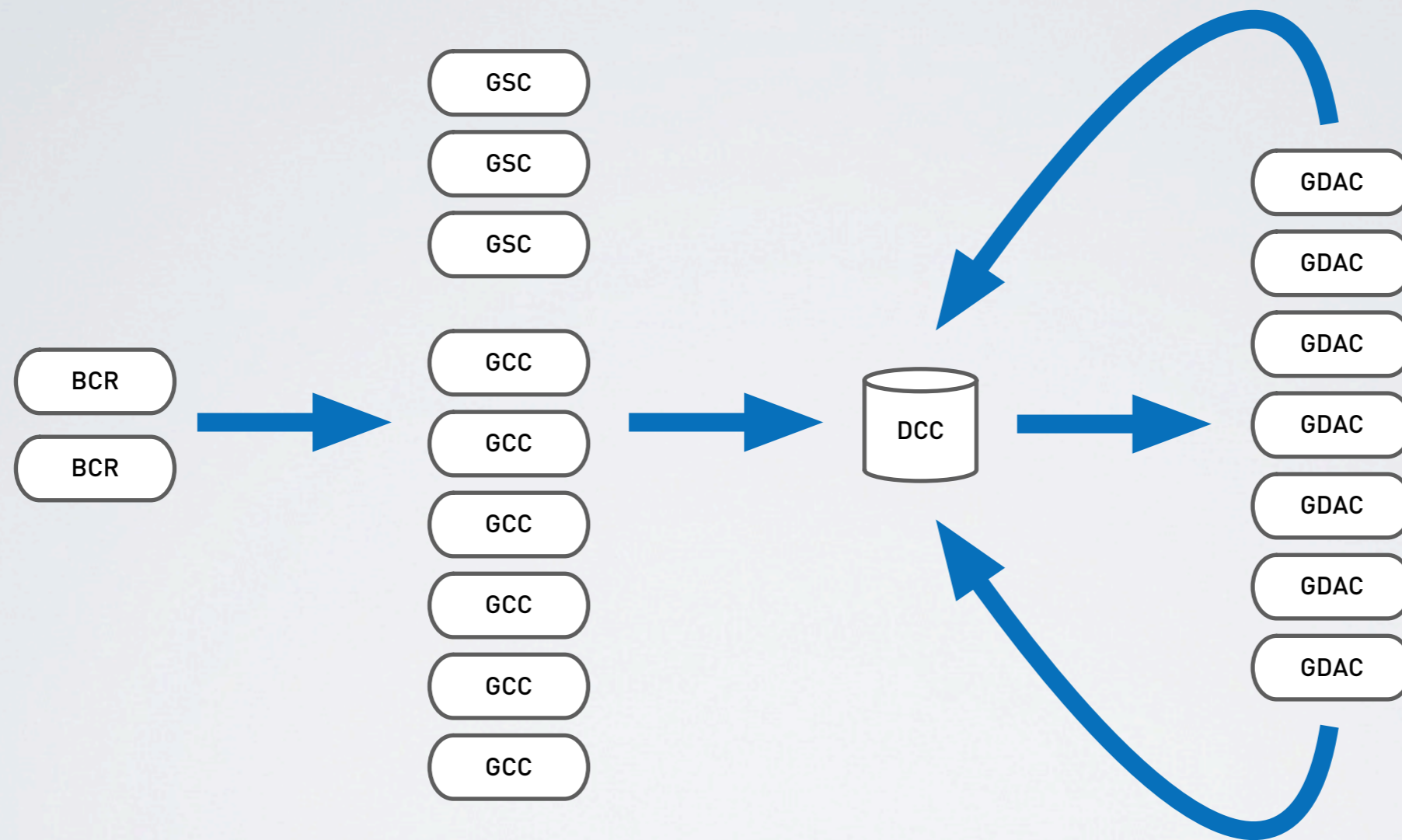
TCGA

SYNONYM: FLOOD (OF DATA & ALGORITHMS)



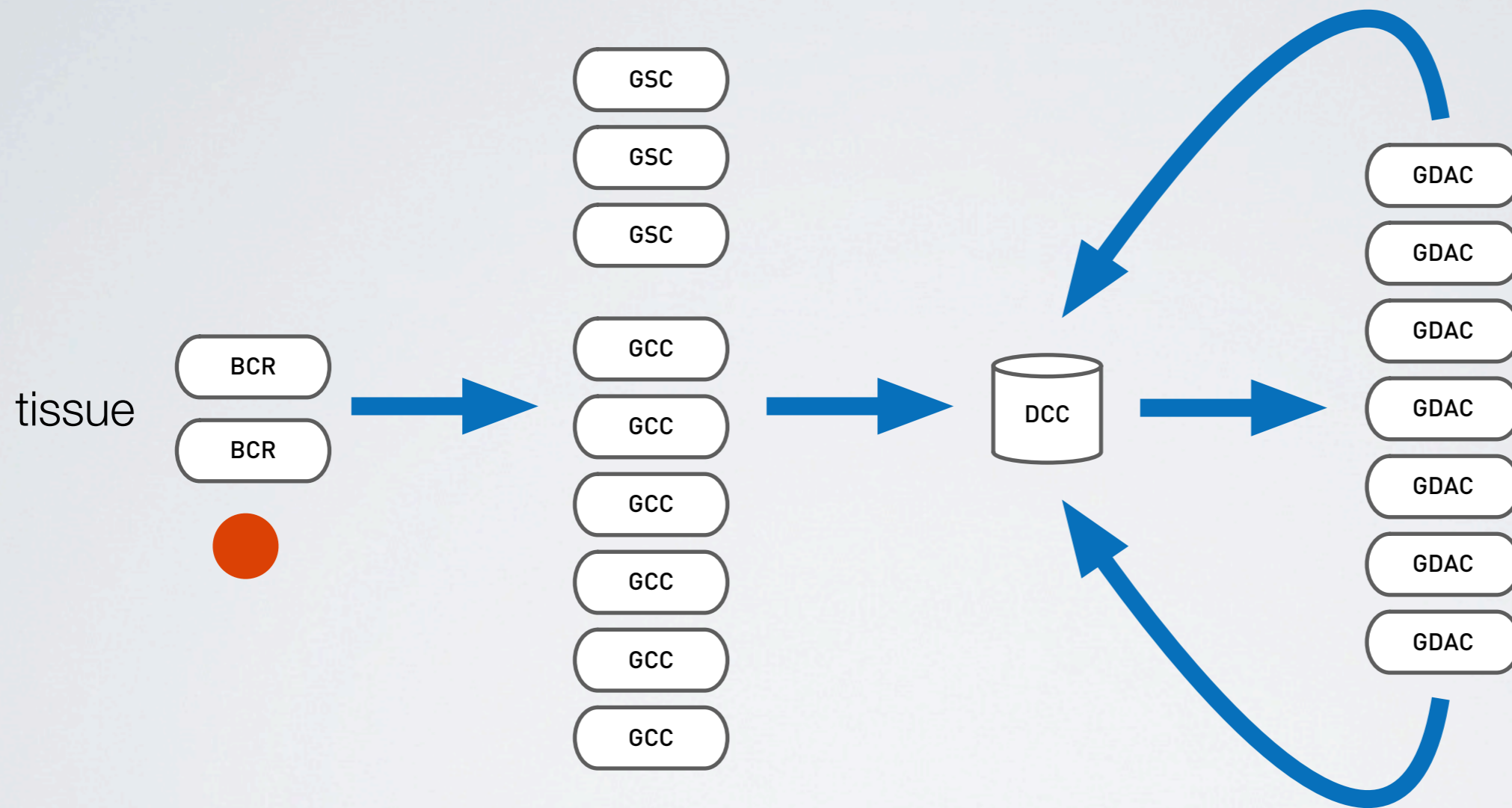
- Thousands of samples: 23 tumor sets + clinical
- Already 5K patient cases, heading to 11K+ total
- Swirling amongst 20 centers nationwide
- **TODAY ... AND EVOLVING DAILY**

Tremendous National-Scale Data Coordination & Standards Challenge



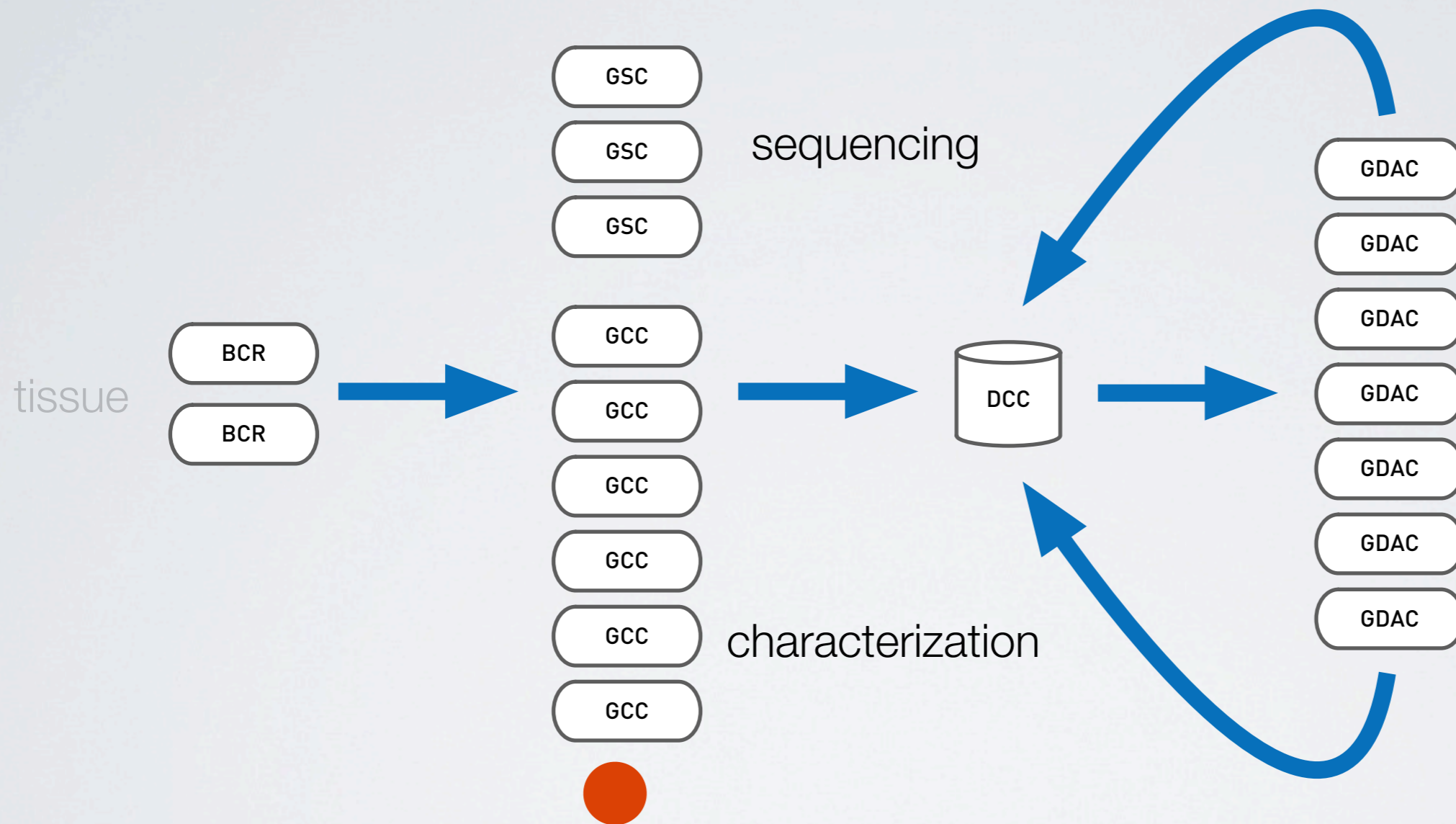
COMPLEX LIFE CYCLE OF A TCGA SAMPLE

Tremendous National-Scale Data Coordination & Standards Challenge



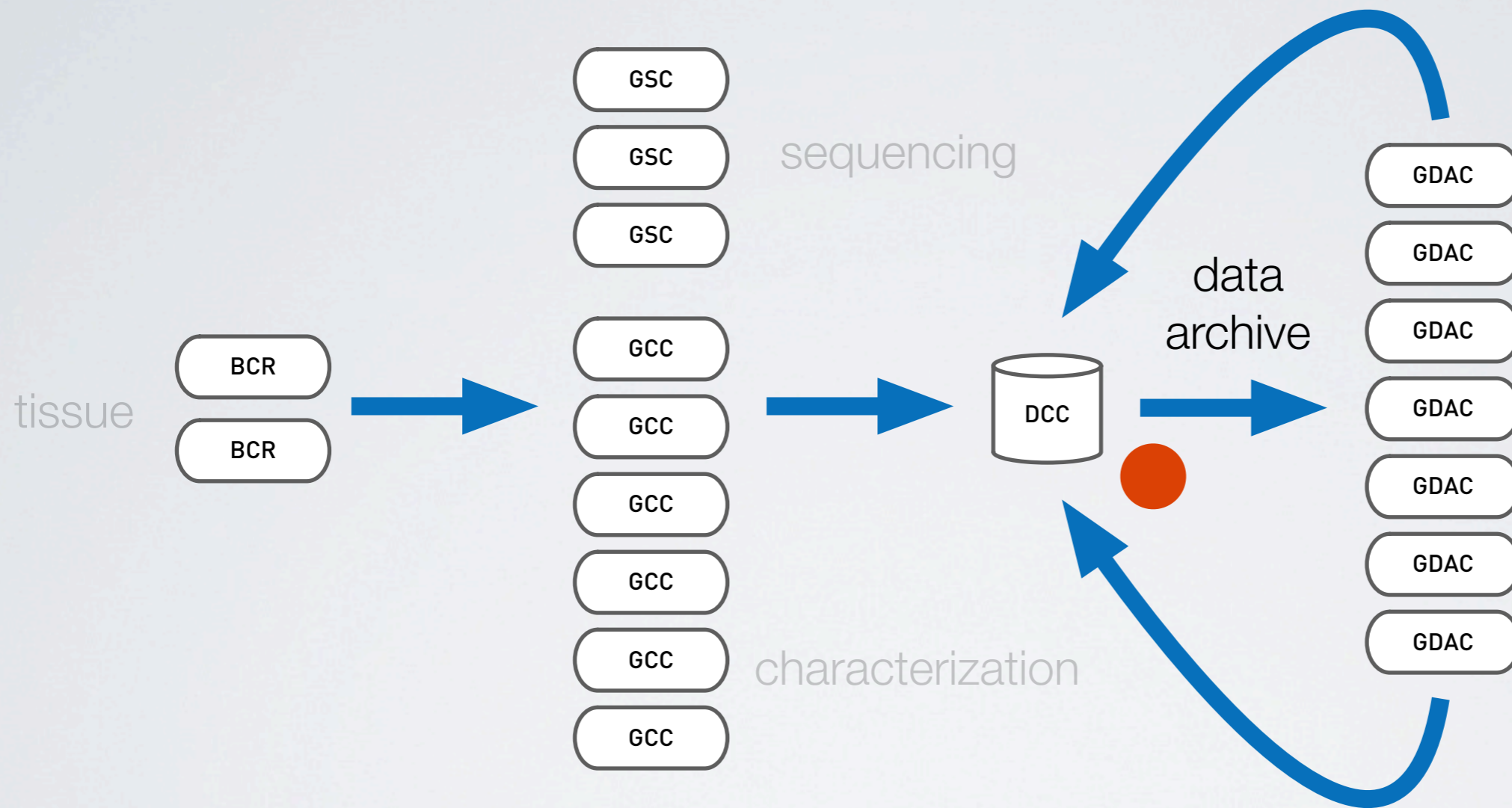
COMPLEX LIFE CYCLE OF A TCGA SAMPLE

Tremendous National-Scale Data Coordination & Standards Challenge



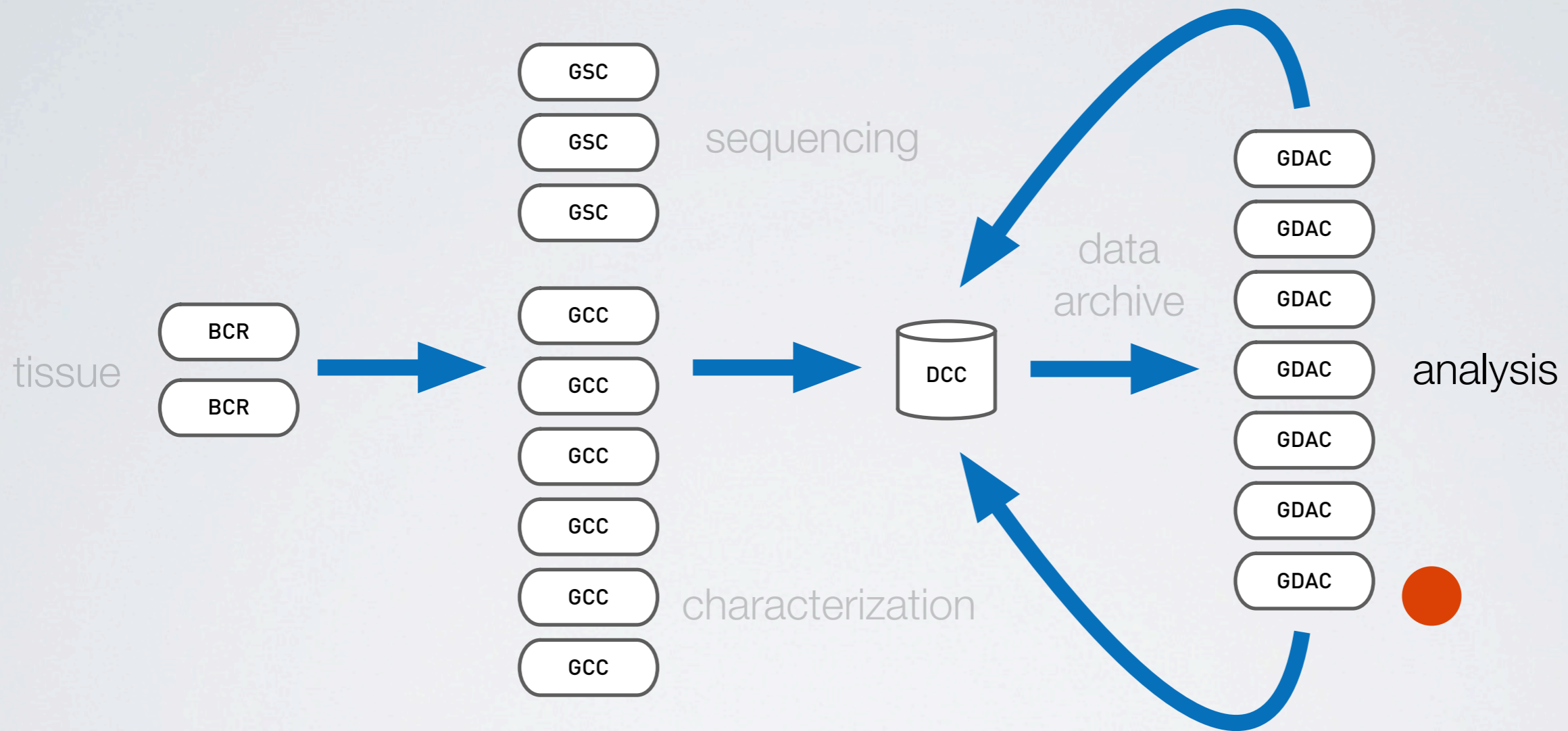
COMPLEX LIFE CYCLE OF A TCGA SAMPLE

Tremendous National-Scale Data Coordination & Standards Challenge



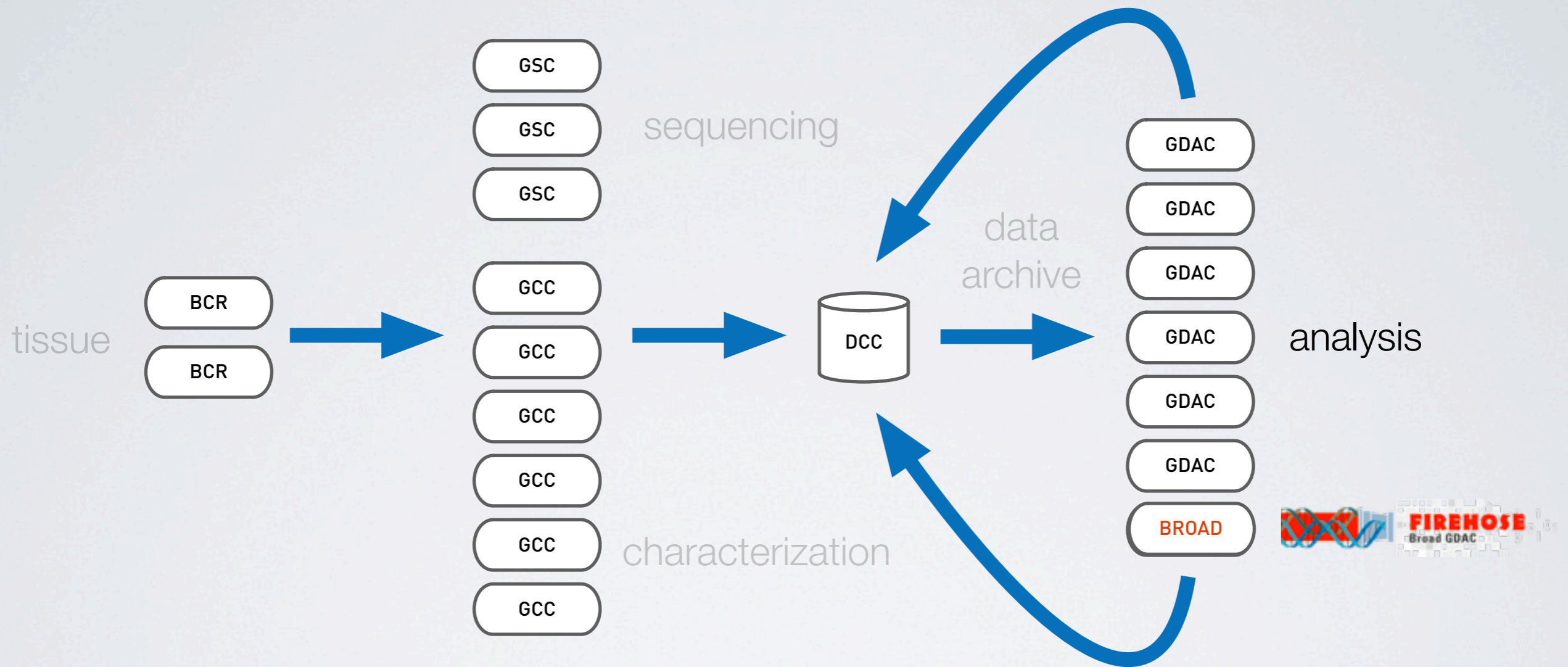
COMPLEX LIFE CYCLE OF A TCGA SAMPLE

Tremendous National-Scale Data Coordination & Standards Challenge



COMPLEX LIFE CYCLE OF A TCGA SAMPLE

Tremendous National-Scale Data Coordination & Standards Challenge



COMPLEX LIFE CYCLE OF A TCGA SAMPLE

MOTIVATION

- At this point you have a broad sense of the TCGA centers and data stream
- But how do they come together to answer common biological questions?

MOTIVATION

- At this point you have a broad sense of the TCGA centers and data stream
- But how do they come together to answer common biological questions?
- Such as:

Is my gene of interest altered in this tumor type? How?

Is that alteration significantly above the background rate?

What distinguishes tumors with clinical or molecular feature X?

MOTIVATION

- At this point you have a broad sense of the TCGA centers and data stream
- But how do they come together to answer common biological questions?
- Such as:

Is my gene of interest altered in this tumor type? How?

Is that alteration significantly above the background rate?

What distinguishes tumors with clinical or molecular feature X?

- There is no one-size-fits-all, cookie-cutter method to answer such questions
- But some analyses are common to many questions and can be automated:

MOTIVATION

- At this point you have a broad sense of the TCGA centers and data stream
- But how do they come together to answer common biological questions?
- Such as:

Is my gene of interest altered in this tumor type? How?
Is that alteration significantly above the background rate?
What distinguishes tumors with clinical or molecular feature X?

- There is no one-size-fits-all, cookie-cutter method to answer such questions
- But some analyses are common to many questions and can be automated:
 - ▶ Mutation calling, classifying, summarizing and significance-testing
 - ▶ Copy number alteration detection and significance-testing
 - ▶ Expression- and methylation-based clustering
 - ▶ Associating genomic data with common clinical, treatment or survival groups

- These common results then become building blocks for higher-level analysis

- These common results then become building blocks for higher-level analysis
- So that downstream users do not have to repeat each time
- Nor perform ad-hoc reinvention of methods

- These common results then become building blocks for higher-level analysis
- So that downstream users do not have to repeat each time
- Nor perform ad-hoc reinvention of methods
- Nor download all low-level data from which they were generated
- ... just to utilize a lower-level analysis result for higher-level, integrative questions

- These common results then become building blocks for higher-level analysis
- So that downstream users do not have to repeat each time
- Nor perform ad-hoc reinvention of methods
- Nor download all low-level data from which they were generated
- ... just to utilize a lower-level analysis result for higher-level, integrative questions
- Nor should they institute their own ad-hoc data freeze/versioning scheme
- ... to ensure accuracy & reproducibility of analytic/statistical results

- These common results then become building blocks for higher-level analysis
- So that downstream users do not have to repeat each time
- Nor perform ad-hoc reinvention of methods
- Nor download all low-level data from which they were generated
- ... just to utilize a lower-level analysis result for higher-level, integrative questions
- Nor should they institute their own ad-hoc data freeze/versioning scheme
- ... to ensure accuracy & reproducibility of analytic/statistical results
- Nor institute ad-hoc QC program ... to minimize human error in large-data analyses

- These common results then become building blocks for higher-level analysis
- So that downstream users do not have to repeat each time
- Nor perform ad-hoc reinvention of methods
- Nor download all low-level data from which they were generated
- ... just to utilize a lower-level analysis result for higher-level, integrative questions
- Nor should they institute their own ad-hoc data freeze/versioning scheme
- ... to ensure accuracy & reproducibility of analytic/statistical results
- Nor institute ad-hoc QC program ... to minimize human error in large-data analyses

It is these concerns which Firehose aims to address.

II : WHAT?

WHAT IS FIREHOSE?

WHAT IS FIREHOSE?

2 THINGS ... FROM A USER PERSPECTIVE

WHAT IS FIREHOSE?

1

Pipeline
infrastructure

Written in
Java[script]

Deployed
as Web APP

WHAT IS FIREHOSE?

1

Pipeline
infrastructure

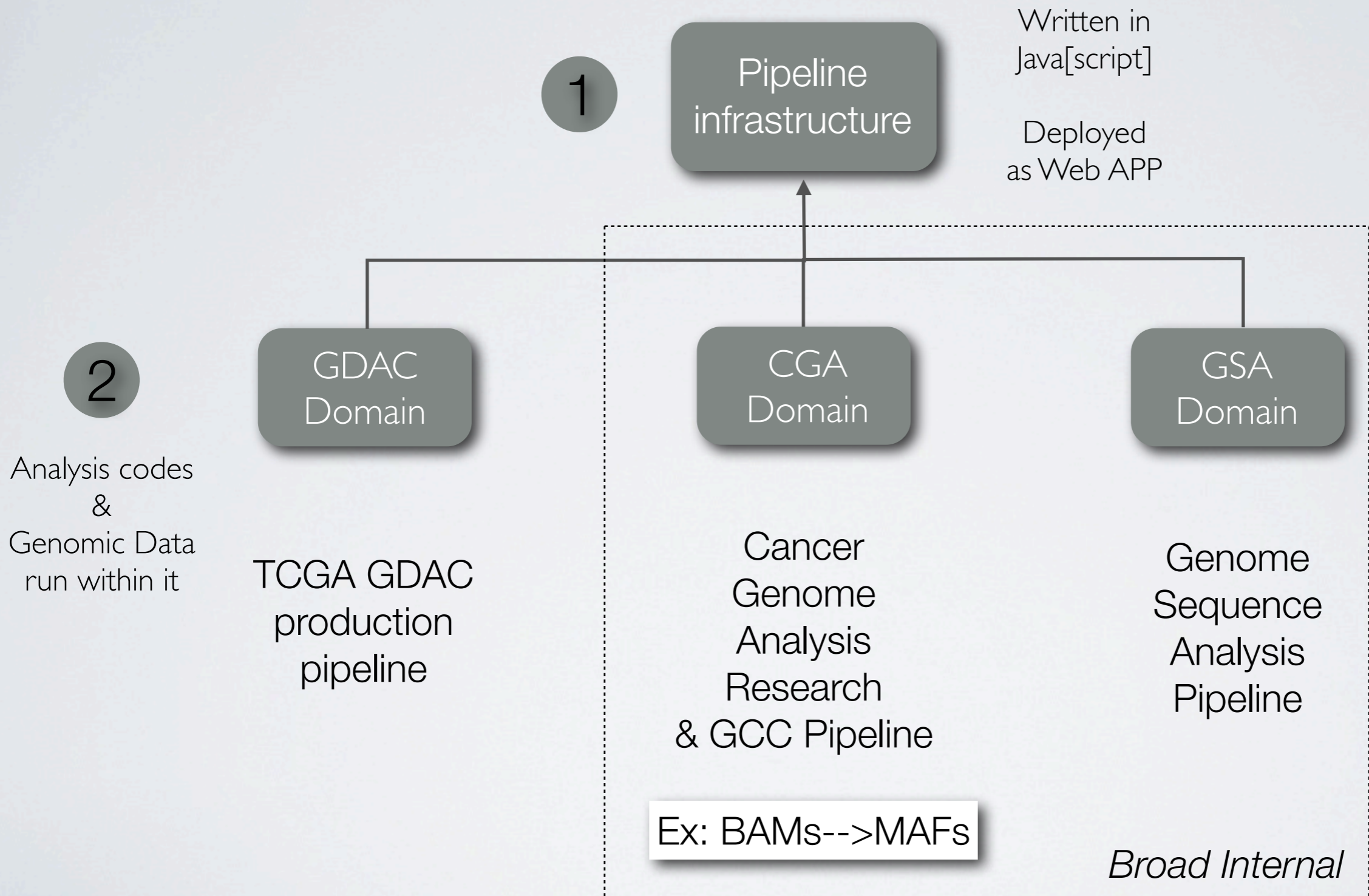
Written in
Java[script]

Deployed
as Web APP

2

Analysis codes
&
Genomic Data
run within it

WHAT IS FIREHOSE?



PROVIDING

- Version control for computational experiments
- Coupled with automated pipeline infrastructure
- Where both analysis code AND data are versioned
- Towards highest possible standards of:

PROVIDING

- Version control for computational experiments
- Coupled with automated pipeline infrastructure
- Where both analysis code AND data are versioned
- Towards highest possible standards of:
 - ▶ Throughput
 - ▶ Transparency → Reproducibility
 - ▶ Scientific Vetting
 - ▶ And ultimately, Reliability

PROVIDING

- Version control for computational experiments
- Coupled with automated pipeline infrastructure
- Where both analysis code AND data are versioned
- Towards highest possible standards of:
 - ▶ Throughput
 - ▶ Transparency → Reproducibility
 - ▶ Scientific Vetting
 - ▶ And ultimately, Reliability

Everything computed as quickly as possible.

... verified as accurately as possible.

... recorded as completely as possible.

Because The Bad Old Days: Manual Experimentation

% create a folder

% download ***data.from.some.where***

% perform local data validation

% run_your_computational_analysis

Because The Bad Old Days: Manual Experimentation

```
% create a folder
```

```
% download data.from.some.where
```

```
% perform local data validation
```

```
% run_your_computational_analysis
```

Then do it again Nov 13, 17, ...

Then forget ... and search, search, search

Then repeat ALL for 19 more tumors

GBM, LUNG, AML, ...

Because The Bad Old Days: Manual Experimentation

% create a folder

% download ***data.from.some.where***

% perform local data validation

% run_your_computational_analysis

Then do it again Nov 13, 17, ...

Then forget ... and search, search, search

Then repeat ALL for 19 more tumors

GBM, LUNG, AML, ...

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

DOESN'T SCALE TO TCGA : OCT 2011 DATA

TumorType	Biospecimen	Any_Level_1	Clinical	CNA	Methylation	mRNA	miR	MAF
BLCA	54	26	26	35	0	0	0	0
BRCA	844	662	810	703	316	533	0	522
CESC	75	23	6	36	0	0	0	0
COAD	423	202	404	375	167	155	0	158
COADREAD	591	276	555	520	236	224	0	227
DLBC	10	0	0	0	0	0	0	0
GBM	600	550	534	537	288	543	491	276
HNSC	241	97	160	127	0	0	0	0
KIRC	502	475	497	489	219	72	0	0
KIRP	107	43	49	43	36	16	0	0
LAML	202	0	0	0	188	0	178	0
LGG	80	30	63	58	0	27	0	0
LIHC	59	38	0	53	0	0	0	0
LNNH	2	0	0	0	0	0	0	0
LUAD	270	85	195	172	128	33	0	258
LUSC	229	184	210	194	133	155	0	188
OV	592	570	580	519	519	570	566	316
PAAD	14	7	0	14	0	0	0	0
PRAD	101	65	0	82	0	0	0	0
READ	168	74	151	145	69	69	0	69
STAD	149	111	148	149	82	0	0	0
THCA	133	39	0	85	0	0	0	0
UCEC	421	220	341	283	117	54	0	237
Totals	5276	3501	4174	4099	2262	2227	1235	2024
	+ 1423	+ 1154	+ 2055	+ 1615	+ 271	+ 213	+ 76	+ 1168

DOESN'T SCALE TO TCGA : OCT 2011 DATA

TumorType	Biospecimen	Any_Level_1	Clinical	CNA	Methylation	mRNA	miR	MAF
BLCA	54	26	26	35	0	0	0	0
BRCA	844	662	810	703	316	533	0	522
CESC	75	23	6	36	0	0	0	0
COAD	423	202	404	375	167	155	0	158
COADREAD	591	276	555	520	236	224	0	227
DLBC	10	0	0	0	0	0	0	0
GBM	600	550	534	537	288	543	491	276
HNSC	241	97	160	127	0	0	0	0
KIRC	502	475	497	489	219	72	0	0
KIRP	107	43	49	43	36	16	0	0
LAML	202	0	0	0	188	0	178	0
LGG	80	30	63	58	0	27	0	0
LIHC	59	38	0	53	0	0	0	0
LNNH	2	0	0	0	0	0	0	0
LUAD	270	85	195	172	128	33	0	258
LUSC	229	184	210	194	133	155	0	188
OV	592	570	580	519	519	570	566	316
PAAD	14	7	0	14	0	0	0	0
PRAD	101	65	0	82	0	0	0	0
READ	168	74	151	145	69	69	0	69
STAD	149	111	148	149	82	0	0	0
THCA	133	39	0	85	0	0	0	0
UCEC	421	220	341	283	117	54	0	237
Totals	5276	3501	4174	4099	2262	2227	1235	2024
	+ 1423	+ 1154	+ 2055	+ 1615	+ 271	+ 213	+ 76	+ 1168

} Diffs Since April

So Firehose Produces

1. Biologist-friendly reports, companioned with
2. Regular package of standard analyses results (~monthly)

For published, vetted algorithms: GISTIC, MutSig, ...

3. From version-stamped, standardized datasets

Generated at Broad, precursor to automated pipeline

So Firehose Produces

1. Biologist-friendly reports, companioned with
2. Regular package of standard analyses results (~monthly)

For published, vetted algorithms: GISTIC, MutSig, ...

3. From version-stamped, standardized datasets

Generated at Broad, precursor to automated pipeline

These broadly map to 3 use cases, loosely corresponding to computational preference.

Use Case 1: Brief

- Browse reports only
- High Level : capture flavor, not depth
- Quickly gain sense of big picture for tumor type X
- When time is short: think PIs
- Useful for idea creation, hypothesis generation
- Can be offline :
 - ▶ On a plane
 - ▶ Or in tedious meetings

Use Case 2: Hands On

- Perhaps start with reports for perspective, but also
- Explore automated analysis results in more depth
- Load output data files from DCC into R, Matlab, etc
- Low-hanging point-of-reference for your custom analyses

Use Case 2: Hands On

- Perhaps start with reports for perspective, but also
- Explore automated analysis results in more depth
- Load output data files from DCC into R, Matlab, etc
- Low-hanging point-of-reference for your custom analyses

“Oh, that’s interesting, maybe my code has found something here ... I wonder if this is seen in the Firehose results, too?”

Use Case 2: Hands On

- Perhaps start with reports for perspective, but also
- Explore automated analysis results in more depth
- Load output data files from DCC into R, Matlab, etc
- Low-hanging point-of-reference for your custom analyses

“Oh, that’s interesting, maybe my code has found something here ... I wonder if this is seen in the Firehose results, too?”

- **Durability of DCC archive fosters citable referencing:**

“We compared our results to TCGA dataset version X generated by Firehose version Y”

Use Case 3: Cutting Edge

- Computational sophisticate
- Maybe doesn't want canned analyses
- Or wants to verify automated pipeline output
- Prefers to reprocess entire analysis sequence
- From scratch, using only lowest-level data

Use Case 3: Cutting Edge

- Computational sophisticate
- Maybe doesn't want canned analyses
- Or wants to verify automated pipeline output
- Prefers to reprocess entire analysis sequence
- From scratch, using only lowest-level data

Standardized, versioned data quite useful here

- ▶ Avoid hard/tedious work of aggregating & normalizing data by hand from 19 centers
- ▶ Fosters concordant views of data: my result may differ from yours because I used v3 of TCGA dataset, but you used v2



Operational ~11 months

*Reproduce ~90% of
2-3 years TCGA pilot
analyses results in
2-3 days*

ARTICLES

Comprehensive genomic characterization defines human glioblastoma genes and core pathways

The Cancer Genome Atlas Research Network*

Human cancer cells typically harbour multiple chromosomal aberrations, nucleotide substitutions and epigenetic modifications that drive malignant transformation. The Cancer Genome Atlas (TCGA) pilot project aims to assess the value of large-scale multi-omic data to the research community. We report the first TCGA pilot project results for glioblastoma, including whole-genome copy number, DNA methylation, gene expression and DNA methylation aberrations. We identify a network view of glioblastoma genes, including TP53, uncovers a network view of glioblastoma genes, including TP53, and a network view of glioblastoma genes, including TP53. We also identify a network view of glioblastoma genes, including TP53. Together, these findings establish the feasibility and power of TCGA, demonstrating that it can rapidly expand knowledge of the molecular basis of cancer.

GBM 2008



Operational ~11 months

ARTICLE

doi:10.1038/nature10166

Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*

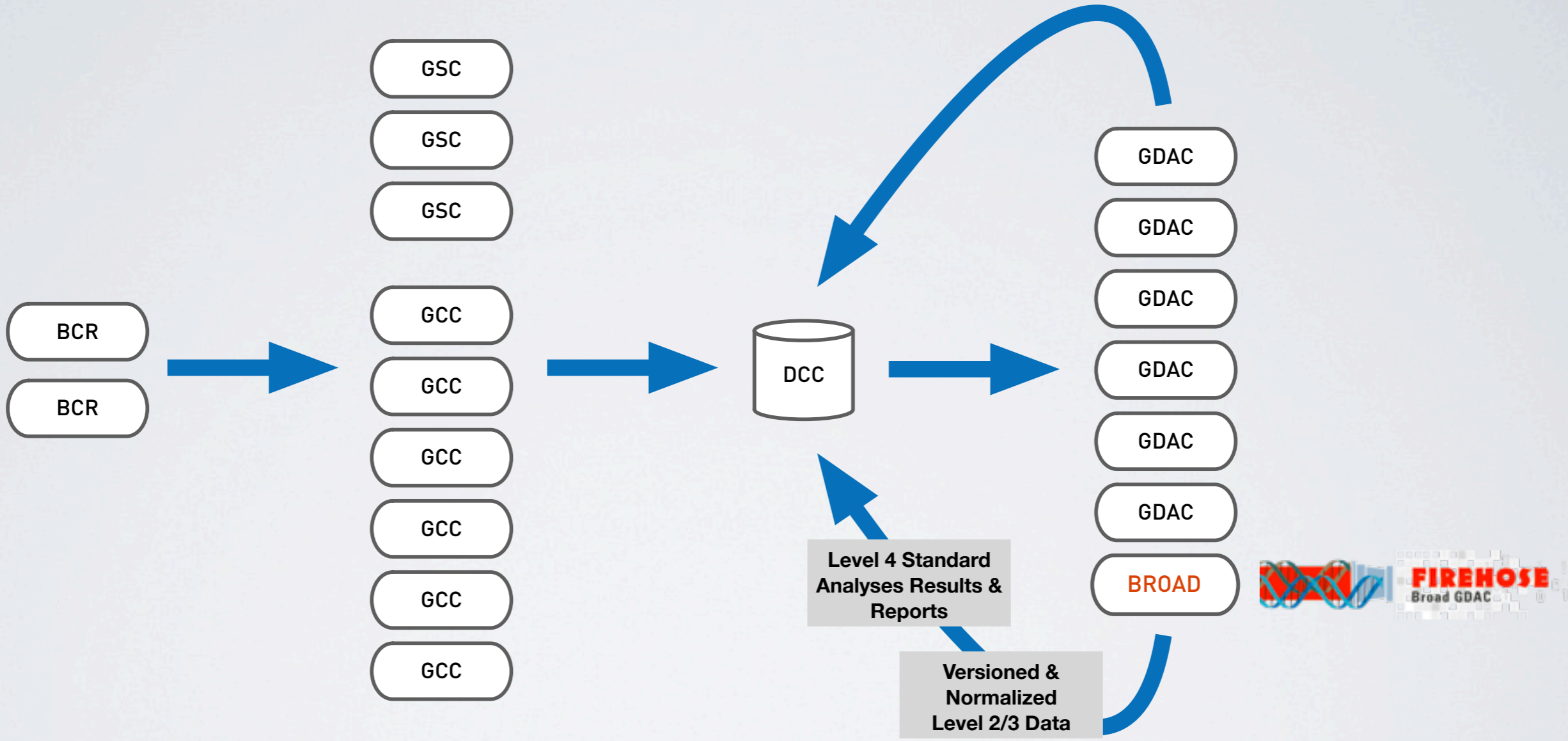
A catalogue of molecular aberrations that cause ovarian cancer is critical for developing and deploying therapies that will improve patients' lives. The Cancer Genome Atlas project has analysed messenger RNA expression, microRNA expression, promoter methylation, DNA sequences of cancer is characterized by somatic mutations, copy number aberrations, cancer transcriptome signature associated with BRCA1/2 (BRCA1 or BRCA2) and CCNE1. We identify a network view of ovarian cancer genes, including BRCA1/2, and a network view of ovarian cancer genes, including BRCA1/2. Together, these findings establish the feasibility and power of TCGA, demonstrating that it can rapidly expand knowledge of the molecular basis of cancer.

OV 6/2011

Reproduce ~90% of 2-3 years TCGA pilot analyses results in 2-3 days

III : How?

FIREHOSE ROLES IN TCGA



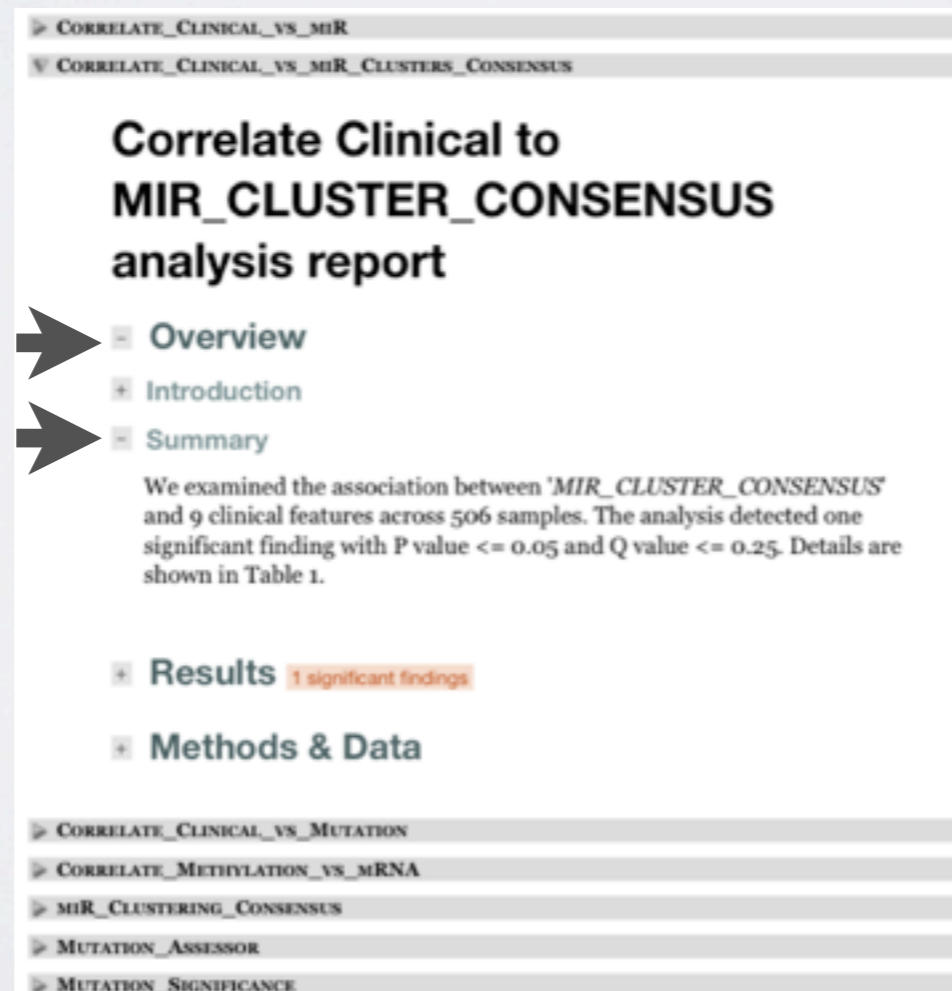
ROLE 1: MONTHLY ANALYSIS RUNS

- APPROX 20 PIPELINES, MANY TAKEN FROM TCGA PILOT
- RUN EN MASSE: AGAINST ALL AVAILABLE TCGA DATA
- WITH EASILY COMPREHENDED SUMMARY REPORTS
- LIKE DRAFT RESULTS SECTION ... SANS PUBLICATION DELAY

ROLE 1: MONTHLY ANALYSIS RUNS

- APPROX 20 PIPELINES, MANY TAKEN FROM TCGA PILOT
- RUN EN MASSE: AGAINST ALL AVAILABLE TCGA DATA
- WITH EASILY COMPREHENDED SUMMARY REPORTS
- LIKE DRAFT RESULTS SECTION ... SANS PUBLICATION DELAY

Nozzle : Analyst &
Biologist-Friendly
Reports



➤ CORRELATE_CLINICAL_VS_MIR

▼ CORRELATE_CLINICAL_VS_MIR_CLUSTERS_CONSENSUS

Correlate Clinical to MIR_CLUSTER_CONSENSUS analysis report

➤ Overview

✦ Introduction

➤ Summary

We examined the association between 'MIR_CLUSTER_CONSENSUS' and 9 clinical features across 506 samples. The analysis detected one significant finding with P value ≤ 0.05 and Q value ≤ 0.25 . Details are shown in Table 1.

✦ Results **1 significant findings**

✦ Methods & Data

➤ CORRELATE_CLINICAL_VS_MUTATION

➤ CORRELATE_METHYLATION_VS_MRNA

➤ MIR_CLUSTERING_CONSENSUS

➤ MUTATION_ASSESSOR

➤ MUTATION_SIGNIFICANCE

ROLE 1: MONTHLY ANALYSIS RUNS

- APPROX 20 PIPELINES, MANY TAKEN FROM TCGA PILOT
- RUN EN MASSE: AGAINST ALL AVAILABLE TCGA DATA
- WITH EASILY COMPREHENDED SUMMARY REPORTS
- LIKE DRAFT RESULTS SECTION ... SANS PUBLICATION DELAY

Nozzle : Analyst &
Biologist-Friendly
Reports

The screenshot shows a report interface with a sidebar on the left containing a list of pipeline names: CORRELATE_CLINICAL_VS_MiR, CORRELATE_CLINICAL_VS_MiR_CLUSTERS_CONSENSUS, CORRELATE_CLINICAL_VS_MUTATION, CORRELATE_METHYLATION_VS_MRNA, MiR_CLUSTERING_CONSENSUS, MUTATION_ASSESSOR, and MUTATION_SIGNIFICANCE. The main content area displays the report for 'CORRELATE_CLINICAL_VS_MiR_CLUSTERS_CONSENSUS'. The report title is 'Correlate Clinical to MIR_CLUSTER_CONSENSUS analysis report'. The sidebar has expandable sections: Overview, Introduction, Summary, Results (highlighted with '1 significant findings'), and Methods & Data. The Summary section contains the text: 'We examined the association between 'MIR_CLUSTER_CONSENSUS' and 9 clinical features across 506 samples. The analysis detected one significant finding with P value <= 0.05 and Q value <= 0.25. Details are shown in Table 1.' Arrows point from the text 'Nozzle : Analyst & Biologist-Friendly Reports' to the Overview and Summary sections, and from the text 'don't miss needle in haystack' to the Results section.

- Standard visual format for ALL pipelines
- Intelligent Scoping:
 - drill from overview to details
 - Significant results “bubble up”
- **don't miss needle in haystack**

Firehose Reports | At-a-Glance

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

Navigation: Navigate to previous or next report or to the overview page.

Layout: In auto width mode the report is automatically fit to the width of the browser window.

Interactions: Expand or collapse all sections of the report. Load a printable version of the report. Tell us about a problem with the report or the results by sending an email directly to our tracking system.

Content: Contact the report maintainer by email. Red markers indicate statistically significant results in this section. Red boxes indicate statistically significant results. Tables can be sorted by clicking on a column header. Underlined results have associated supplementary information. Click to open the supplementary results panel. Dark blue indicates the result for which supplementary information is shown.

Download Results: This is an experimental feature. The full results of the analysis summarized in this report can be downloaded from the TCGA Data Coordination Center.

Table 1: Amplifications

Cytoband	Q value	Residual Q value	Wide Peak Boundaries	# Genes in Wide Peak
<u>7p11.2</u>	0	0	chr7:54954372-54968011	0 [EGFR]
12q14.1	5.1922e-139	6.202e-113	chr12:36411663-36442647	5
4q12	6.7649e-85	6.7649e-85	chr4:54727006-54861623	1
13q32.1	1.3248e-57	1.7421e-57	chr13:202664385-202815140	2
12q15	3.8163e-70	4.0392e-31	chr12:67457108-67551544	2
3p26.33	4.5642e-09	4.5642e-09	chr3:182584087-183044402	2
7q31.2	9.9818e-09	1.7005e-08	chr7:116103324-116267511	1
12p13.32	2.4873e-08	2.4873e-08	chr12:38391333-4302336	3
13q44	2.0116e-07	4.0275e-07	chr13:241495233-242804011	6
7q21.2	1.2098e-06	2.7782e-06	chr7:39366270-42280411	5
13q32.1	1.7964e-05	1.7964e-05	chr13:13735235-14250524	2
2p24.3	4.3245e-05	4.5245e-05	chr2:15933362-16304271	2
13q34	0.03487	0.03487	chr13:108563148-109682638	3
19q12	0.069145	0.069145	chr19:34867390-35007574	2

Table 2: Deletions

Genes in Wide Peak

Genes
CDK4
CTP27B1
TSPAN31
MARCK19
AGAP2

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

Navigation:

- Navigate to previous or next report or to the overview page.
- Expand or collapse all sections of the report.
- In auto width mode the report is automatically fit to the width of the browser window.
- Load a printable version of the report.
- Tell us about a problem with the report or the results by sending an email directly to our tracking system.

Content Features:

- Contact the report maintainer by email.
- Red markers indicate statistically significant results in this section.
- Red boxes indicate statistically significant results.
- Click figures to enlarge. Click again to scale down.
- Get the complete set of results as a text file.
- Tables can be sorted by clicking on a column header.
- Underlined results have associated supplementary information. Click to open the supplementary results panel. Dark blue indicates the result for which supplementary information is shown.

Report Content:

Glioblastoma Multifforme: Copy number analysis (GISTIC2)
 Maintained by Dan DiCara (Broad Institute)

Overview

- Introduction
- Summary

There were 501 tumor samples used in this analysis: 23 significant arm-level results, 14 significant focal amplifications, and 52 significant focal deletions were found.

Results

- Focal results

Table 1. Amplifications Table - 14 significant amplifications found.

Cytoband	Q value	Residual Q value	Wide Peak Boundaries	# Genes in Wide Peak
<u>7p11.2</u>	0	0	chr7:54954372-54968011	0 [EGFR]
<u>12q14.1</u>	5.19228	6.2028-113	chr12:56411663-56442647	5
<u>4q12</u>	6.76498	6.76498-85	chr4:54727006-54861623	1
<u>13q32.1</u>	1.32488	1.74218-57	chr13:202664385-202815140	2
<u>12q15</u>	3.81638	4.03928-31	chr12:67457108-67551544	2
<u>3p26.33</u>	4.56428	4.56428-09	chr3:182584087-183044402	2
<u>7q31.2</u>	9.98188	1.70058-08	chr7:116103324-116267511	1
<u>12p13.32</u>	2.48738	2.48738-08	chr12:38391333-4302336	3
<u>13q44</u>	2.01188	4.02758-07	chr13:441495233-242804011	6
<u>7q21.2</u>	1.20988	2.77828-06	chr7:9266270-9268284	5
<u>13q32.1</u>	1.79648	1.79648-05	chr13:13735235-14250524	2
<u>2p24.3</u>	4.32488	4.52488-05	chr2:15933362-16304271	2
<u>13q34</u>	0.03487	0.03487	chr13:108563148-109682638	3
<u>19q12</u>	0.069145	0.069145	chr19:34867390-35007574	2

Table 2. Deletions Table - 52 significant deletions found.

Genes in Wide Peak

This is the comprehensive list of genes in the wide peak for 12q14.1.

Table S1. Genes in bold are cancer genes as defined by The Sanger Institute's Cancer Gene Census [7].

Genes
CDK4
CTP27B1
TSPAN31
MARCI19
AGAP2

Download Results

This is an experimental feature. The full results of the analysis summarized in this report can be downloaded from the TCGA Data Coordination Center.

- Analysis Results (MD5 checksum)
- Auxiliary Data (MD5 checksum)
- MAGE-TAB File (MD5 checksum)

References

Organized like a paper

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

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

Navigation and Convenience:

- Navigate to previous or next report or to the overview page.
- Expand or collapse all sections of the report.
- In auto width mode the report is automatically fit to the width of the browser window.
- Load a printable version of the report.
- Tell us about a problem with the report or the results by sending an email directly to our tracking system.
- Contact the report maintainer by email.
- Click "X" to hide the supplementary results panel.
- Download all result files associated with the analysis presented in the report from the TCGA DCC.

Data and Results:

- Red markers indicate statistically significant results in this section.
- Red boxes indicate statistically significant results.
- Tables can be sorted by clicking on a column header.
- Underlined results have associated supplementary information. Click to open the supplementary results panel. Dark blue indicates the result for which supplementary information is shown.

Table 1: Amplifications Table - 14 significant amplifications found.

Cytoband	Q value	Residual Q value	Wide Peak Boundaries	# Genes in Wide Peak
7p11.2	0	0	chr7:54954372-54968011	0 [EGFR]
12q14.1	5.1922e-09	6.202e-113	chr12:56411663-56442647	5
4q12	6.7649e-85	6.7649e-85	chr4:54727006-54861623	1
13q32.1	1.3248e-57	1.7421e-57	chr13:202664385-202815140	2
12q15	3.8163e-70	4.0392e-31	chr12:67457108-67551544	2
3p26.33	4.5642e-09	4.5642e-09	chr3:182584087-183044402	2
7q31.2	9.9818e-09	1.7005e-08	chr7:116103324-116267511	1
12p13.32	2.4873e-08	2.4873e-08	chr12:38391333-4302336	3
13q44	2.0116e-07	4.0275e-07	chr13:241495233-242804011	6
7q21.2	1.2098e-06	2.7782e-06	chr7:9266270-9268284	5
13p06.21	1.7964e-05	1.7964e-05	chr13:13735235-14250524	2
2p24.3	4.3245e-05	4.3245e-05	chr2:15933362-16304271	2
13q34	0.03487	0.03487	chr13:108563148-109682638	3
19q12	0.059145	0.059145	chr19:34867390-35007574	2

Table 2: Deletions Table - 52 significant deletions found.

Genes in Wide Peak

Table S1. Genes in bold are cancer genes as defined by The Sanger Institute's Cancer Gene Census [7].

Genes
CDK4
CTP27B1
TSPAN31
MARCKS19
AGAP2

Organized like a paper

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

With Browser Convenience

- Dynamic zooming
- And navigation
- View partial or full data
- Easily printable
- Built-in bug reporting
- No HTML coding: just R

Firehose Reports: Example 1

ARTICLE doi:10.1038/nature10166

Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*

UP EXPAND ALL COLLAPSE ALL SET AUTO WIDTH PRINT REPORT REPORT A PROBLEM

Ovarian Serous Cystadenocarcinoma: Mutation Analysis (MutSig)

Maintained by [Estat Stojanov](#) (Broad Institute)

- Overview
 - Introduction
 - Summary
- Results
 - Breakdown of Mutations by Type
 - Breakdown of Mutation Rates by Category Type
 - Target Coverage for Each Individual
 - Distribution of Mutation Counts, Coverage, and Mutation Rates Across Samples
 - Significantly Mutated Genes

Table 3. A Ranked List of Significantly Mutated Genes. Number of significant genes found: 9. Number of genes displayed: 35

rank	gene	description	N	n	n1	n2	n3	n4	n5	p	q
1	TP53	tumor protein p53	384444	292	48	32	37	63	112	<1.00e-11	<1.89e-07
2	BRCA1	breast cancer 1, early onset	1728968	9	0	0	1	0	8	1.33e-05	0.013
3	NF1	neurofibromin 1 (neurofibromatosis, von Recklinghausen disease, Watson disease)	2512245	13	1	0	1	3	8	2.43e-06	0.015
4	FAT3	FAT tumor suppressor homolog 3 (Drosophila)	3559809	19	4	2	3	9	1	0.000013	0.053
5	GABRA6	gamma-aminobutyric acid (GABA) A receptor, alpha 6	423382	6	1	3	1	1	0	0.000023	0.087
6	CDK12		1295954	9	0	0	1	3	5	0.000035	0.092
7	CSMD3	CU3 and Sushi multiple domains 3	3473921	19	1	2	7	8	1	0.000037	0.092
8	RB1	retinoblastoma 1 (including osteosarcoma)	791208	6	0	0	1	0	5	0.000039	0.092
9	BRCA2	breast cancer 2, early onset	2762828	10	1	0	0	2	7	0.000054	0.11
10	OR5D16	olfactory receptor, family 5, subfamily D, member 16	295338	4	2	0	1	1	0	0.00015	0.29
11	TNRC10	tumor necrosis receptor...	314800	7	0	0	2	1	0	0.00017	0.30

Table 2 | Significantly mutated genes in HGS-OvCa

Gene	No. of mutations	No. validated	No. unvalidated
<i>TP53</i>	302	294	8
<i>BRCA1</i>	11	10	1
<i>CSMD3</i>	19	19	0
<i>NF1</i>	13	13	0
<i>CDK12</i>	9	9	0
<i>FAT3</i>	19	18	1
<i>GABRA6</i>	6	6	0
<i>BRCA2</i>	10	10	0
<i>RB1</i>	6	6	0

Validated mutations are those that have been confirmed with an independent assay. Most of them are validated using a second independent whole-genome-amplification sample from the same tumour. Unvalidated mutations have not been independently confirmed but have a high likelihood to be true mutations. An extra 25 mutations in *TP53* were observed by hand curation.

Mutation Significance

Firehose Reports: Example 2

Cell
PRESS

Cancer Cell
Article

Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*

EXPAND ALL COLLAPSE ALL SET AUTO WIDTH PRINT REPORT REPORT A PROBLEM

Glioblastoma Multiforme: Clustering of mRNA expression: consensus NMF

Maintained by Robert Zapko (Proval Institute)

- Overview
- Introduction
- Summary
- Results

The most robust consensus NMF clustering of 490 samples using the 7500 most variable genes was identified for $k = 4$ clusters. We computed the clustering for $k = 3$ to $k = 8$ and used the asphenetic correlation coefficient to determine the best solution.

Gene expression patterns of molecular subtypes

Consensus and correlation matrix

Figure 2. The consensus matrix after clustering shows 4 clusters with limited overlap between clusters.

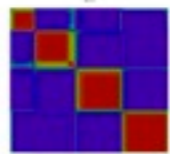


Figure 3. The correlation matrix also shows 4 clusters.

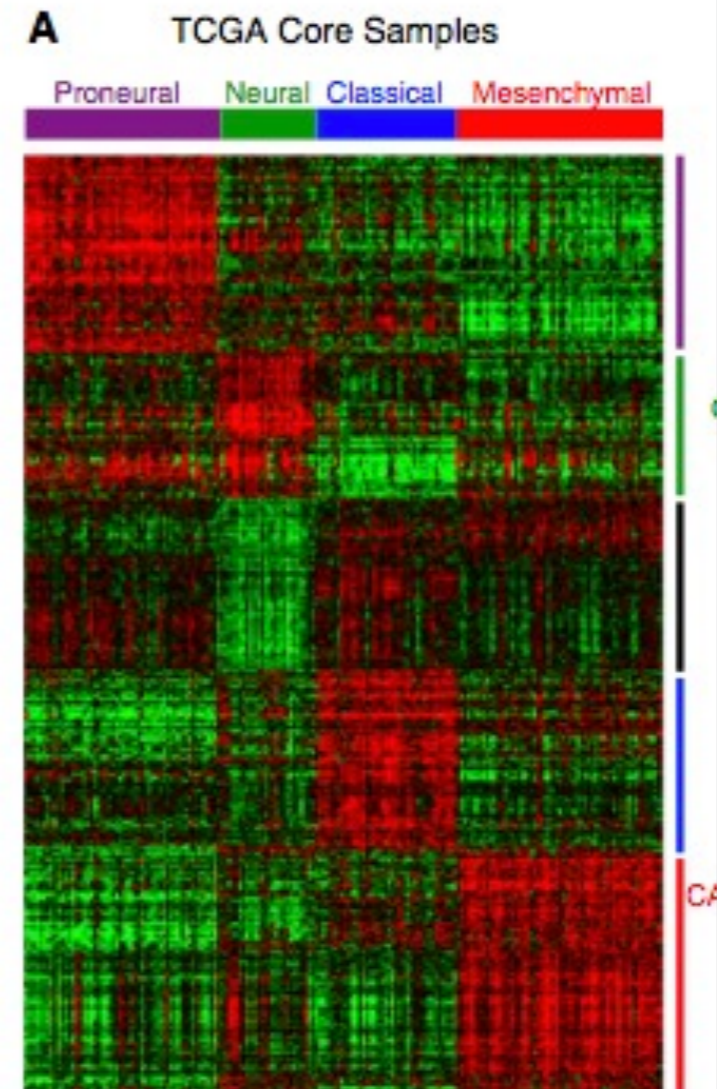
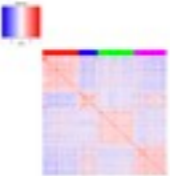


Figure 2. Gene Expression Data Identify Four Gene
(A) Using the predictive 840 gene list, samples were ordered samples.

Gene Expression Clustering

Firehose Reports: Example 3

ARTICLE

doi:10.1038/nature10166

Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*

Genome Browser: Ovarian Serous Cystadenocarcinoma: Copy number analysis (GISTIC2)

Minimised by Das, DGAs (Profil Institute)

- Overview
- Introduction
- Summary
- Results **113 significant findings**
- Focal results **64 significant findings**

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

Figure 2. Genomic positions of deleted regions: the X-axis represents the normalized deletion signal (blue) and significance by Q value (bottom). The green line represents the significance cutoff at Q value=0.25.

Table 1. Amplifications: Table - 39 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.

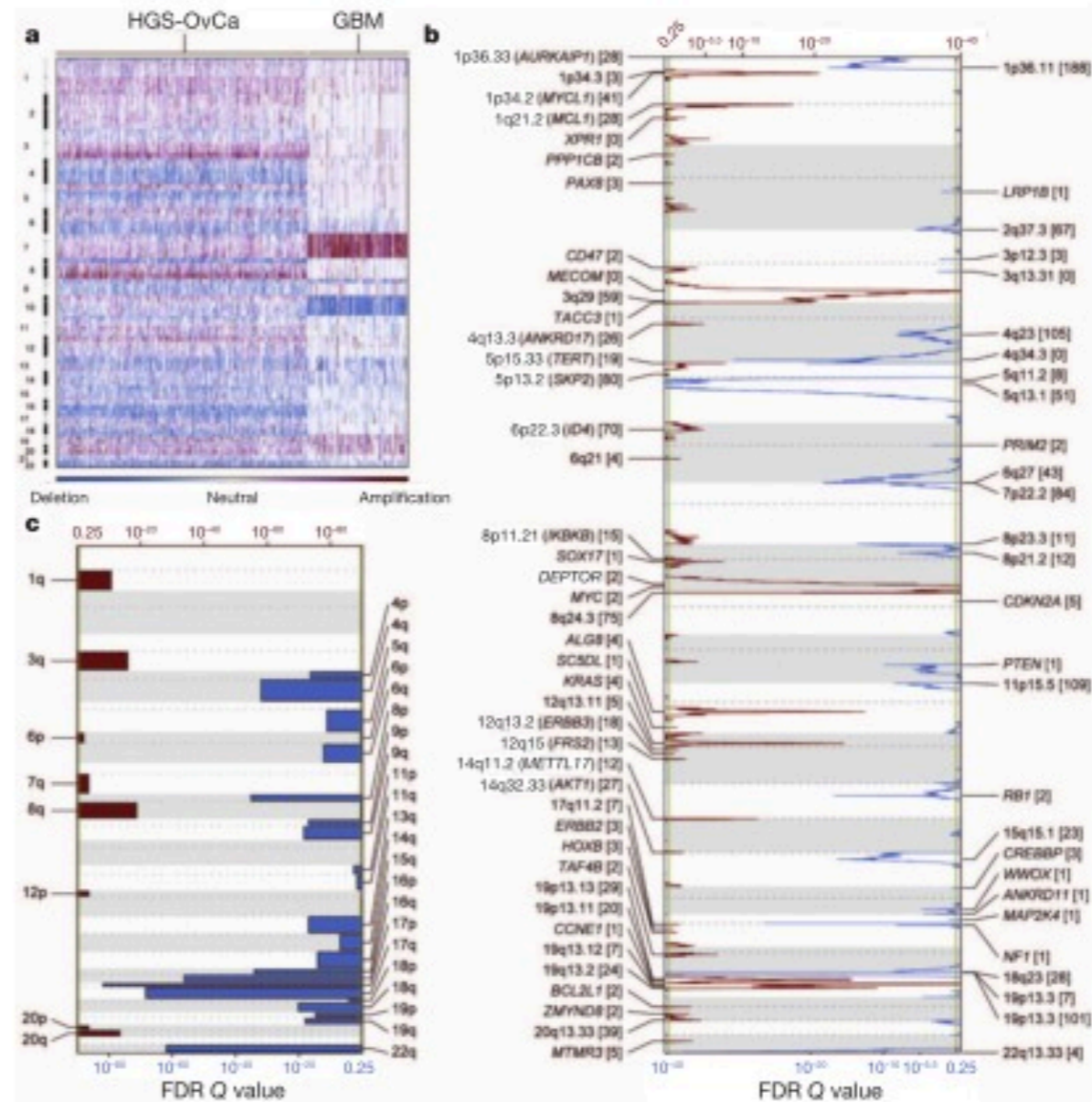


Figure 1 | Genome copy number abnormalities. **a**, Copy number profiles of 89 HGS-OvCa, compared with profiles of 197 glioblastoma multiforme (GBM) tumors. **b**, Genomic positions of significant amplified and deleted regions, well-localized regions with fewer genes, and regions with known cancer genes or genes identified in previous studies. **c**, Heatmap of FDR Q values for chromosomes 1-22. The number of genes identified in each region is shown in brackets.

Copy Number Alterations

FINE PRINT

These results are offered to the community as an additional reference point, enabling a wide range of cancer biologists, clinical investigators, and genome & computational scientists to easily incorporate TCGA into the backdrop of ongoing research.

FINE PRINT

These results are offered to the community as an additional reference point, enabling a wide range of cancer biologists, clinical investigators, and genome & computational scientists to easily incorporate TCGA into the backdrop of ongoing research.

STARTING POINT : NOT FINAL WORD

FINE PRINT

These results are offered to the community as an additional reference point, enabling a wide range of cancer biologists, clinical investigators, and genome & computational scientists to easily incorporate TCGA into the backdrop of ongoing research.

STARTING POINT : NOT FINAL WORD

- Aim is to enable readers (like bench bios, clinical trialists)
- To quickly take pulse of pipeline for given tumor type(s)
- With just a few glances at common representational figures
- Not deep head-scratching

Flow of Standard Analyses Runs

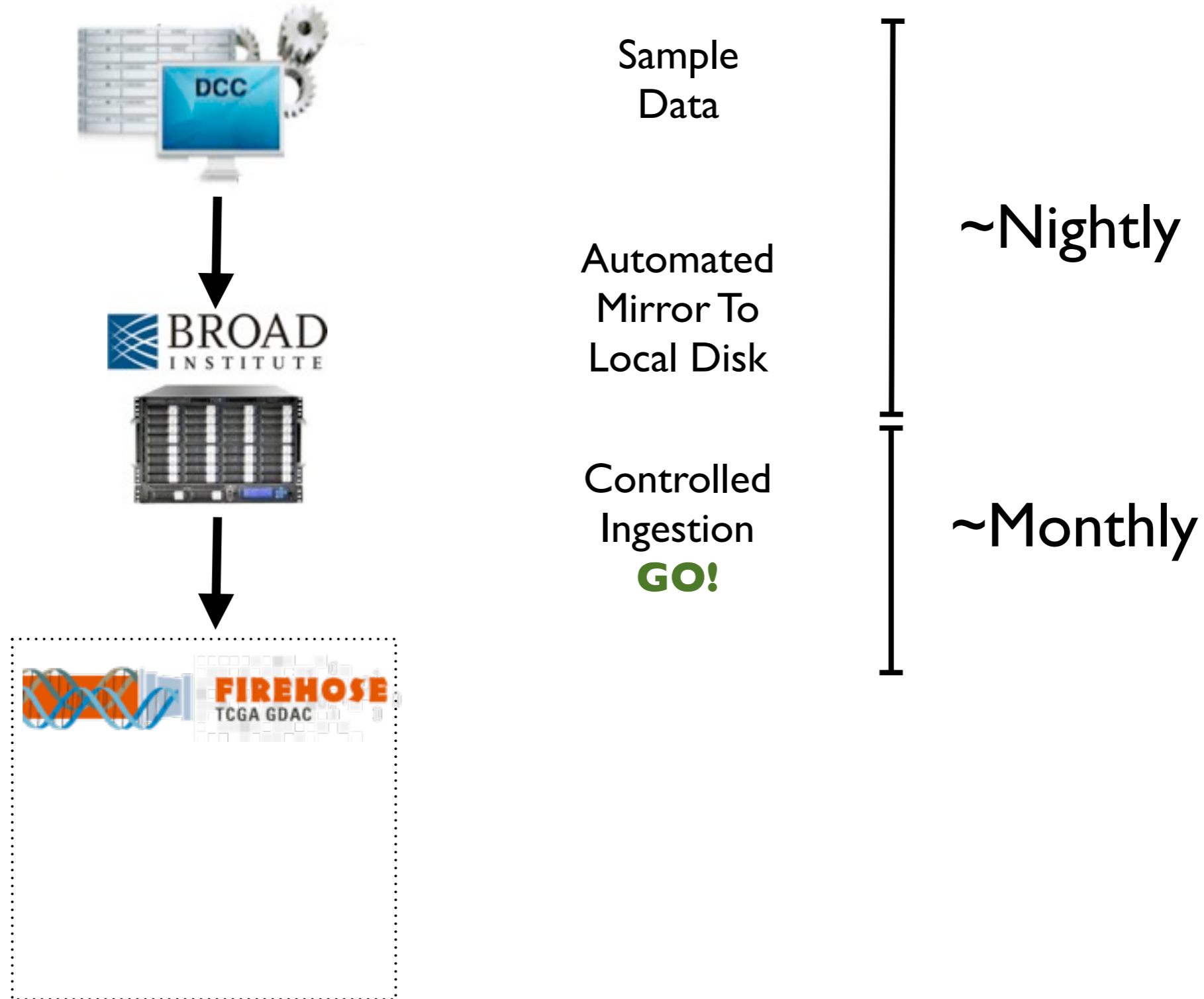


Sample
Data

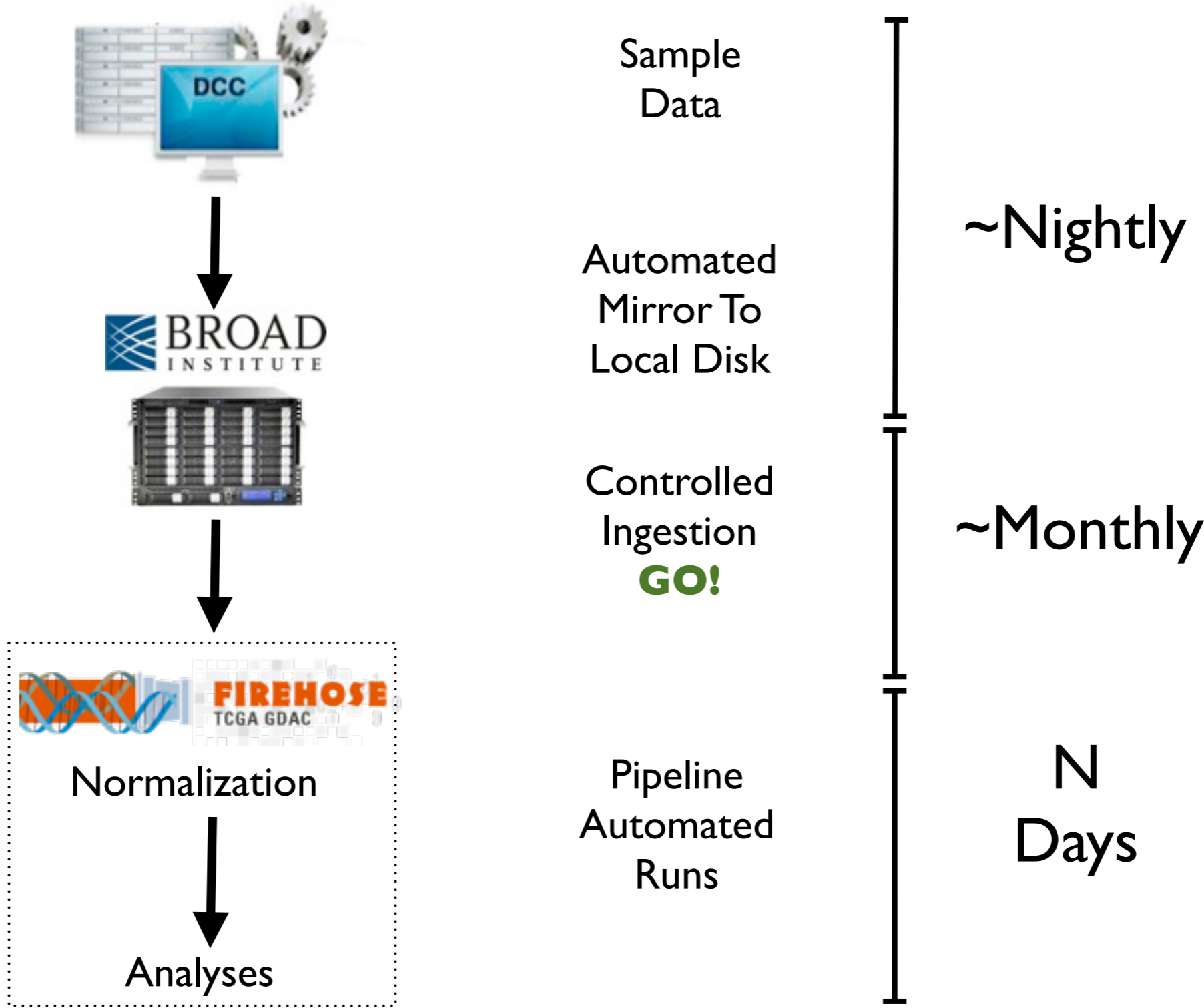
Automated
Mirror To
Local Disk

~Nightly

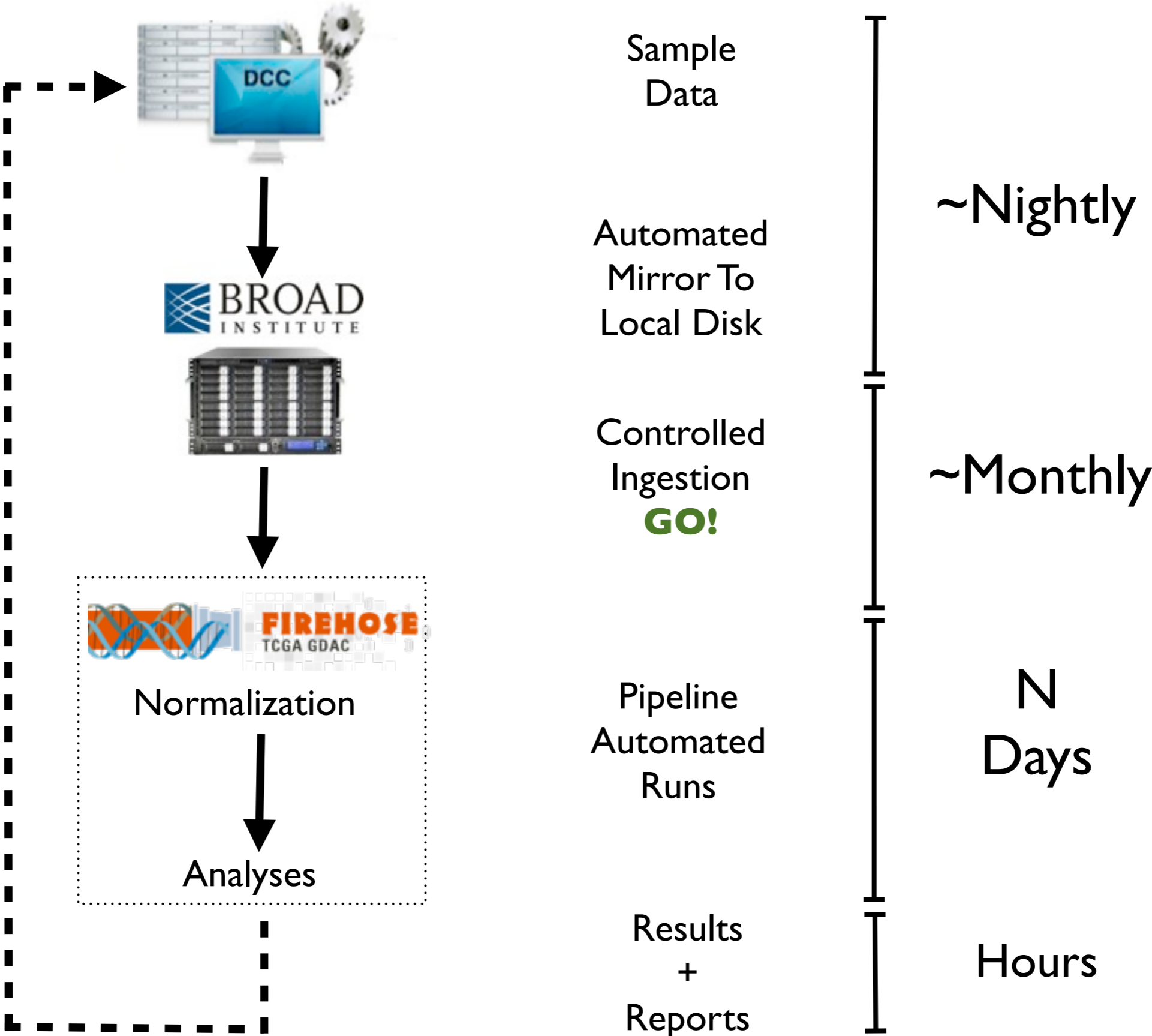
Flow of Standard Analyses Runs



Flow of Standard Analyses Runs



Flow of Standard Analyses Runs



BUT WHILE DOING THIS WE CONSTANTLY SEE

THE BABEL PROBLEM

BUT WHILE DOING THIS WE CONSTANTLY SEE

THE BABEL PROBLEM

RARELY IS THERE AGREEMENT ON CENTRAL QUESTION:

BUT WHILE DOING THIS WE CONSTANTLY SEE

THE BABEL PROBLEM

RARELY IS THERE AGREEMENT ON CENTRAL QUESTION:

HOW MUCH DATA DO WE HAVE?



ROLE 2: VERSIONED DATA RUNS



ROLE 2: VERSIONED DATA RUNS

- BI-WEEKLY OUTPUT OF OUR DATA STANDARDIZER
- WHICH PREPARES TCGA INPUTS FOR AUTOMATIC CONSUMPTION



ROLE 2: VERSIONED DATA RUNS

- BI-WEEKLY OUTPUT OF OUR DATA STANDARDIZER
- WHICH PREPARES TCGA INPUTS FOR AUTOMATIC CONSUMPTION
 - ✓ **Partition:** to one sample per file



ROLE 2: VERSIONED DATA RUNS

- BI-WEEKLY OUTPUT OF OUR DATA STANDARDIZER
- WHICH PREPARES TCGA INPUTS FOR AUTOMATIC CONSUMPTION
 - ✓ **Partition:** to one sample per file
 - ✓ **Cleanup:** remove variations that are problematic for automation



ROLE 2: VERSIONED DATA RUNS

- BI-WEEKLY OUTPUT OF OUR DATA STANDARDIZER
- WHICH PREPARES TCGA INPUTS FOR AUTOMATIC CONSUMPTION
 - ✓ **Partition:** to one sample per file
 - ✓ **Cleanup:** remove variations that are problematic for automation
 - ✓ **Selection:** filtered (by DNU list) samples merged ...



ROLE 2: VERSIONED DATA RUNS

- BI-WEEKLY OUTPUT OF OUR DATA STANDARDIZER
- WHICH PREPARES TCGA INPUTS FOR AUTOMATIC CONSUMPTION
 - ✓ **Partition:** to one sample per file
 - ✓ **Cleanup:** remove variations that are problematic for automation
 - ✓ **Selection:** filtered (by DNU list) samples merged ...
- WE USE THESE NORMED DATA FOR STANDARD ANALYSES
- AND HAVE BEGUN TO PROVIDE TO ENTIRE TCGA



ROLE 2: VERSIONED DATA RUNS

- BI-WEEKLY OUTPUT OF OUR DATA STANDARDIZER
- WHICH PREPARES TCGA INPUTS FOR AUTOMATIC CONSUMPTION
 - ✓ **Partition:** to one sample per file
 - ✓ **Cleanup:** remove variations that are problematic for automation
 - ✓ **Selection:** filtered (by DNU list) samples merged ...
- WE USE THESE NORMED DATA FOR STANDARD ANALYSES
- AND HAVE BEGUN TO PROVIDE TO ENTIRE TCGA

Fostering TCGA-wide **Standard View** of the data stream

BABEL PROBLEM IN ACTION : OVARIAN

rank	gene	p	q
1	TP53	<1.00e-11	<4.72e-08
2	LOC200030	<1.00e-11	<4.72e-08
3	NBPF16	<1.00e-11	<4.72e-08
4	CSNK2B	<1.00e-11	<4.72e-08
5	ACYP1	1.26e-10	4.76e-07
6	PDE8B	3.27e-10	1.03e-06
7	OR2W3	9.79e-10	2.64e-06
8	ACSBG2	1.17e-09	2.75e-06
9	DNAJC25-GNG1	1.91e-09	4.00e-06

383 MUTATION SAMPLES
IN FIREHOSE MAY 2011

~70 CONTAMINATED

BABEL PROBLEM IN ACTION : OVARIAN

rank	gene	p	q
1	TP53	<1.00e-11	<4.72e-08
2	LOC200030	<1.00e-11	<4.72e-08
3	NBPF16	<1.00e-11	<4.72e-08
4	CSNK2B	<1.00e-11	<4.72e-08
5	ACYP1	1.26e-10	4.76e-07
6	PDE8B	3.27e-10	1.03e-06
7	OR2W3	9.79e-10	2.64e-06
8	ACSBG2	1.17e-09	2.75e-06
9	DNAJC25-GNG1	1.91e-09	4.00e-06

383 MUTATION SAMPLES
IN FIREHOSE MAY 2011

~70 CONTAMINATED

Table 2 | Significantly mutated genes in HGS-OvCa

Gene	No. of mutations	No. validated	No. unvalidated
<i>TP53</i>	302	294	8
<i>BRCA1</i>	11	10	1
<i>CSMD3</i>	19	19	0
<i>NF1</i>	13	13	0
<i>CDK12</i>	9	9	0
<i>FAT3</i>	19	18	1
<i>GABRA6</i>	6	6	0
<i>BRCA2</i>	10	10	0
<i>RB1</i>	6	6	0

316 MUTATION SAMPLES
JUNE 2010 MANUSCRIPT (ABOVE)

BABEL PROBLEM IN ACTION : OVARIAN

rank	gene	p	q
1	TP53	<1.00e-11	<4.72e-08
2	LOC200030	<1.00e-11	<4.72e-08
3	NBPF16	<1.00e-11	<4.72e-08
4	CSNK2B	<1.00e-11	<4.72e-08
5	ACYP1	1.26e-10	4.76e-07
6	PDE8B	3.27e-10	1.03e-06
7	OR2W3	9.79e-10	2.64e-06
8	ACSBG2	1.17e-09	2.75e-06
9	DNAJC25-GNG1	1.91e-09	4.00e-06

383 MUTATION SAMPLES
IN FIREHOSE MAY 2011
~70 CONTAMINATED

Table 2 | Significantly mutated genes in HGS-OvCa

Gene	No. of mutations	No. validated	No. unvalidated
<i>TP53</i>	302	294	8
<i>BRCA1</i>	11	10	1
<i>CSMD3</i>	19	19	0
<i>NF1</i>	13	13	0
<i>CDK12</i>	9	9	0
<i>FAT3</i>	19	18	1
<i>GABRA6</i>	6	6	0
<i>BRCA2</i>	10	10	0
<i>RB1</i>	6	6	0

316 MUTATION SAMPLES
JUNE 2010 MANUSCRIPT (ABOVE)
MANUALLY REDACTED FIREHOSE RUN (BELOW)

Significantly Mutated Genes

Table 3. A Ranked List of Significantly Mutated Genes. Number of significant genes found: 9. Number of genes displayed: 35

rank	gene	description	N	n	n1	n2	n3	n4	n5	p	q
1	TP53	tumor protein p53	384444	292	48	32	37	63	112	<1.00e-11	<1.89e-07
2	BRCA1	breast cancer 1, early onset	1728968	9	0	0	1	0	8	1.33e-06	0.013
3	NF1	neurofibromin 1 (neurofibromatosis, von Recklinghausen disease, Watson disease)	2512246	13	1	0	1	3	8	2.43e-06	0.015
4	FAT3	FAT tumor suppressor homolog 3 (Drosophila)	3559809	19	4	2	3	9	1	0.000013	0.063
5	GABRA6	gamma-aminobutyric acid (GABA) A receptor, alpha 6	423382	6	1	3	1	1	0	0.000023	0.087
6	CDK12		1295984	9	0	0	1	3	5	0.000035	0.092
7	CSMD3	CUB and Sushi multiple domains 3	3473121	19	1	2	7	8	1	0.000037	0.092
8	RB1	retinoblastoma 1 (including osteosarcoma)	791208	6	0	0	1	0	5	0.000039	0.092
9	BRCA2	breast cancer 2, early onset	2762828	10	1	0	0	2	7	0.000054	0.11

COULD YOU AVOID BABEL PROBLEM ON YOUR OWN?

Certainly. But do you want to? Is that wise?

COULD YOU AVOID BABEL PROBLEM ON YOUR OWN?

Certainly. But do you want to? Is that wise?

Scores of scientists re-validating
their data across TCGA would
curtail their collective scientific reach.

COULD YOU AVOID BABEL PROBLEM ON YOUR OWN?

Certainly. But do you want to? Is that wise?

Scores of scientists re-validating
their data across TCGA would
curtail their collective scientific reach.

Better to at least try to minimize duplication, no?



ROLE 3: TARGETED AWG RUNS



ROLE 3: TARGETED AWG RUNS

- Analysis Targets Of Opportunity:



ROLE 3: TARGETED AWG RUNS

- Analysis Targets Of Opportunity:

e.g. for coordinated activity like AWG workshops



ROLE 3: TARGETED AWG RUNS

- Analysis Targets Of Opportunity:
 - e.g. for coordinated activity like AWG workshops
- Example: 2 runs performed in April 2011



ROLE 3: TARGETED AWG RUNS

- Analysis Targets Of Opportunity:
 - e.g. for coordinated activity like AWG workshops
- Example: 2 runs performed in April 2011
 - Standard analyses run



ROLE 3: TARGETED AWG RUNS

- Analysis Targets Of Opportunity:
 - e.g. for coordinated activity like AWG workshops
- Example: 2 runs performed in April 2011
 - Standard analyses run
 - TOO for May 2 LUNG workshop in NC



ROLE 3: TARGETED AWG RUNS

- Analysis Targets Of Opportunity:
e.g. for coordinated activity like AWG workshops
- Example: 2 runs performed in April 2011
 - Standard analyses run
 - TOO for May 2 LUNG workshop in NC

Broad GDAC Analysis Summary lung_awg_2011_05_02 Run

Tables of Ingested Data: [HTML](#) [PNG](#) [TSV](#)

Tumor Type	# Completed	Percentage
LUSC	19	79%
LUAD	19	79%

Excerpted GISTIC report [LUAD](#) [LUSC](#)

Excerpted MutSig report [LUAD](#) [LUSC](#)

[Broad Institute VPN](#)

[All LUAD Reports \(needs VPN + FH login\)](#)

[All LUSC Reports \(needs VPN + FH login\)](#)

[Excerpted Nozzle LUAD & LUSC Reports](#)

Peek Behind The Mirror

```
% cd <DCC>/tcga4yeo/tumor && ds
```

blca has size	26G	lihc has size	66G
brca has size	866G	luad has size	163G
cesc has size	17G	lusc has size	224G
coad has size	402G	<u>ov has size</u>	<u>1.6T</u>
<u>gbm has size</u>	<u>1.8T</u>	paad has size	5.3G
hnsc has size	73G	prad has size	66G
kirc has size	453G	read has size	153G
kirp has size	64G	stad has size	84G
laml has size	30G	thca has size	61G
lgg has size	61G	ucec has size	262G

Sept 2011: ~6.4 T total ... CEL, mage-tab, MAF, XML ...

Putting New Codes In

- Source code not private (published/open/available)
- Tested on TCGA data, preferably multiple tumors
- Provides programmatic access to version info
- Runnable from Unix
- Drivable by command line args
- Meaning essentially any language is OK, even proprietary runtimes (but only MatLab so far)
- Library ok, but need executable wrapper
- Then contact us

Coming in 2012: Public FH Release with Task Registry

Accessing Results

Q: How or where can I access the results of a run?

A: In one of two ways:

- Both analyses and standardized data are stored in the [Broad repository of the TCGA Data Coordination Center \(DCC\)](#). After signing in (TCGA credentials required), you should see something like

Index of /tcgafiles/ftp_auth/distro_ftpusers/tcga4yeo/other/gdacs/gdacbroad

Name	Last modified	Size
Parent Directory		-
LATEST_RUN	08-Oct-2011 15:34	40
README.txt	04-Feb-2011 13:33	411
blca/	08-Oct-2011 11:03	-
brca/	08-Oct-2011 10:56	-
cesc/	08-Oct-2011 11:03	-
coad/	08-Oct-2011 10:56	-
coadread/	08-Oct-2011 11:01	-
full/	08-Oct-2011 10:56	-
gbm/	08-Oct-2011 10:56	-
hnscc/	08-Oct-2011 11:03	-
kirc/	08-Oct-2011 10:57	-
kirp/	08-Oct-2011 10:58	-
lanl/	08-Oct-2011 10:58	-
lgg/	08-Oct-2011 11:03	-
lihc/	08-Oct-2011 11:03	-
luad/	08-Oct-2011 10:58	-
lusc/	08-Oct-2011 10:58	-
ov/	08-Oct-2011 10:58	-
paad/	08-Oct-2011 11:03	-
prad/	08-Oct-2011 11:03	-
read/	08-Oct-2011 11:01	-
reports/	12-Oct-2011 14:12	-
stad/	08-Oct-2011 11:01	-
thca/	08-Oct-2011 11:03	-
ucec/	08-Oct-2011 11:01	-

★ **New!**
See
Jim Robinson &
Raktim Sinha
For Details

from which you may simply navigate to the tumor type and run date of interest.

- Standardized data packages can also be viewed directly within your [local IGV installation](#), without signing in to the DCC, by following [the instructions given here](#).

Quicklook Visualization in IGV

The screenshot displays the IGV interface with the 'Available Datasets' dialog box open. The dialog lists the following datasets and their sub-items:

- Available Datasets
 - Broad Firehose Run: 2011_09_21
 - Broad Firehose Run: 2011_07_28
 - COAD
 - CESC
 - PAAD
 - STAD
 - LUSC
 - READ
 - THCA
 - LIHC
 - CBM
 - CopyNumber: [genome_wide_snp_6__broad]
 - Expression: [agilentg4502a_07_2]
 - Methylation: [humanmethylation27__jhu_usc]
 - Mutation
 - LGG
 - LAML
 - COADREAD
 - BLCA
 - KIRC
 - UCEC
 - PRAD
 - LUAD
 - OV
 - BRCA
 - HNSC
 - KIRP
 - Broad Firehose Run: 2011_05_25

The background interface shows the 'Human hg18' genome browser with a 'RefSeq genes' track and a blue signal track. The status bar at the bottom indicates 'Loading ...' and '84M of 253M'.

Quicklook Visualization in IGV

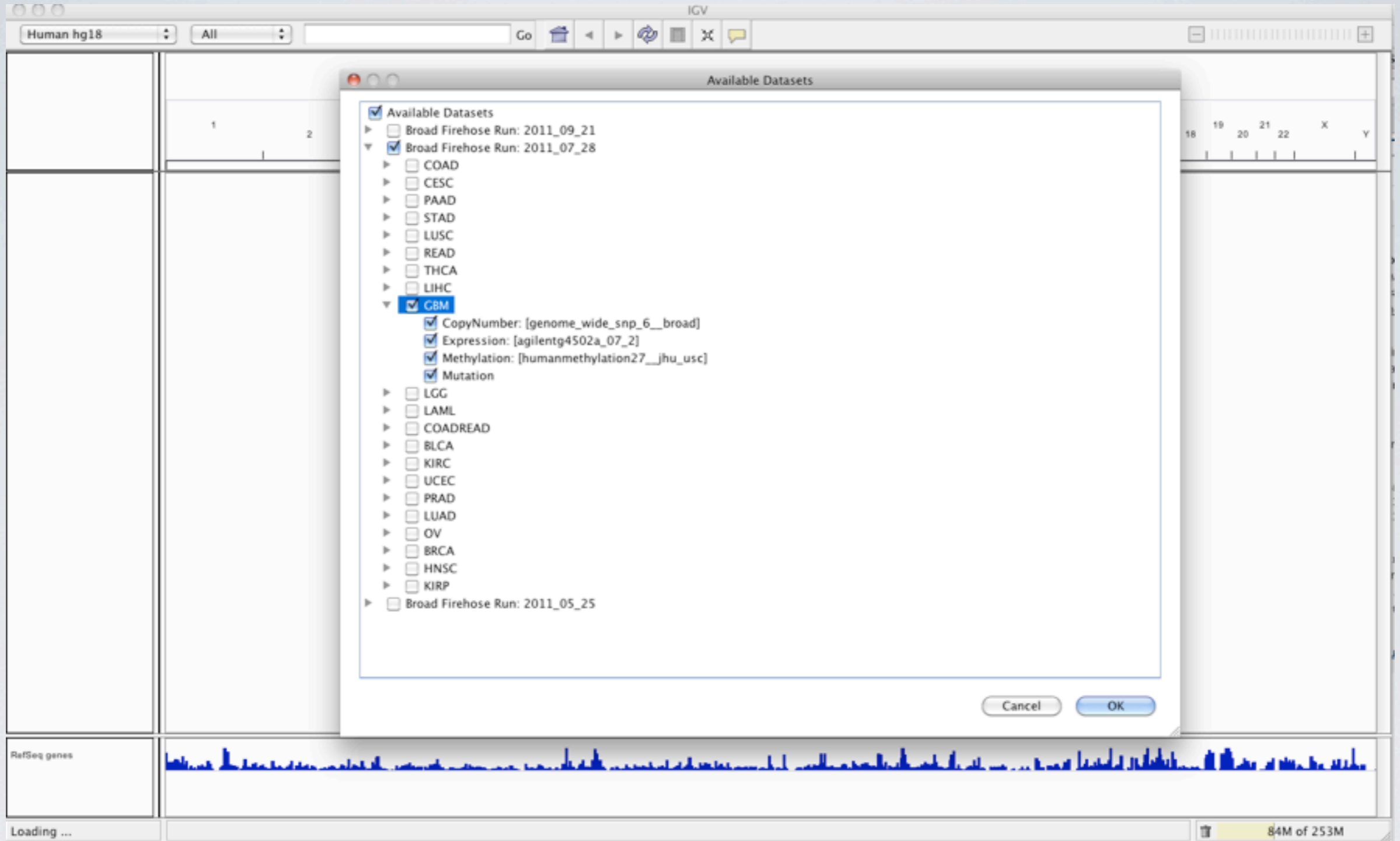
The screenshot displays the IGV interface with the 'Available Datasets' dialog box open. The dialog lists the following datasets and their sub-tracks:

- Available Datasets
 - Broad Firehose Run: 2011_09_21
 - Broad Firehose Run: 2011_07_28
 - COAD
 - CESC
 - PAAD
 - STAD
 - LUSC
 - READ
 - THCA
 - LIHC
 - CBM
 - CopyNumber: [genome_wide_snp_6__broad]
 - Expression: [agilentg4502a_07_2]
 - Methylation: [humanmethylation27__jhu_usc]
 - Mutation
 - LGG
 - LAML
 - COADREAD
 - BLCA
 - KIRC
 - UCEC
 - PRAD
 - LUAD
 - OV
 - BRCA
 - HNSC
 - KIRP
- Broad Firehose Run: 2011_05_25

The background shows a genomic track with a blue signal plot and a 'RefSeq genes' track. The status bar at the bottom indicates 'Loading ...' and '84M of 253M'.

Directly from Broad, no TCGA credentials required

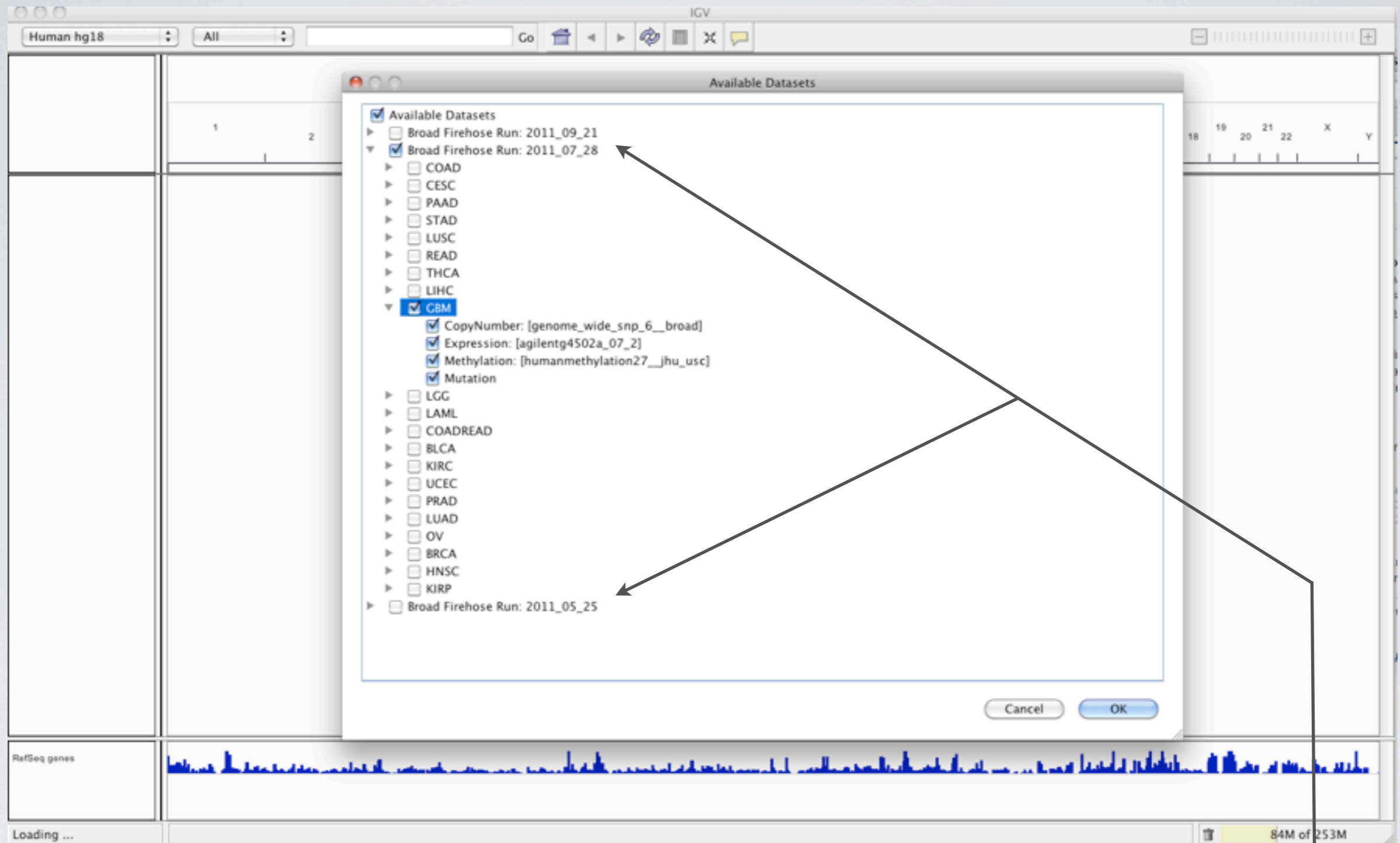
Quicklook Visualization in IGV



Directly from Broad, no TCGA credentials required

<https://confluence.broadinstitute.org/display/GDAC/IGV+Data+Loading>

Quicklook Visualization in IGV



Directly from Broad, no TCGA credentials required

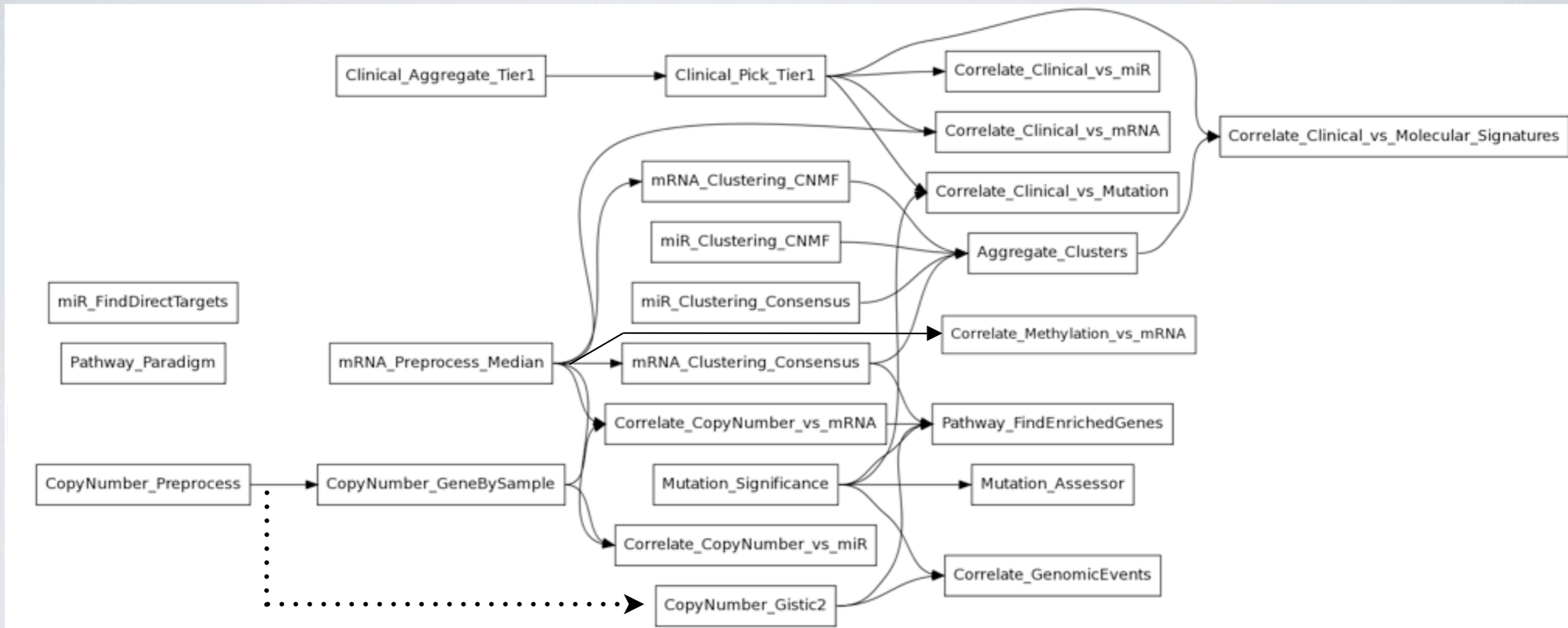
<https://confluence.broadinstitute.org/display/GDAC/IGV+Data+Loading>

Each data package identified by date corresponding to our GDAC runs.

IV : INSIGHTS & CHALLENGES

Insight 1:

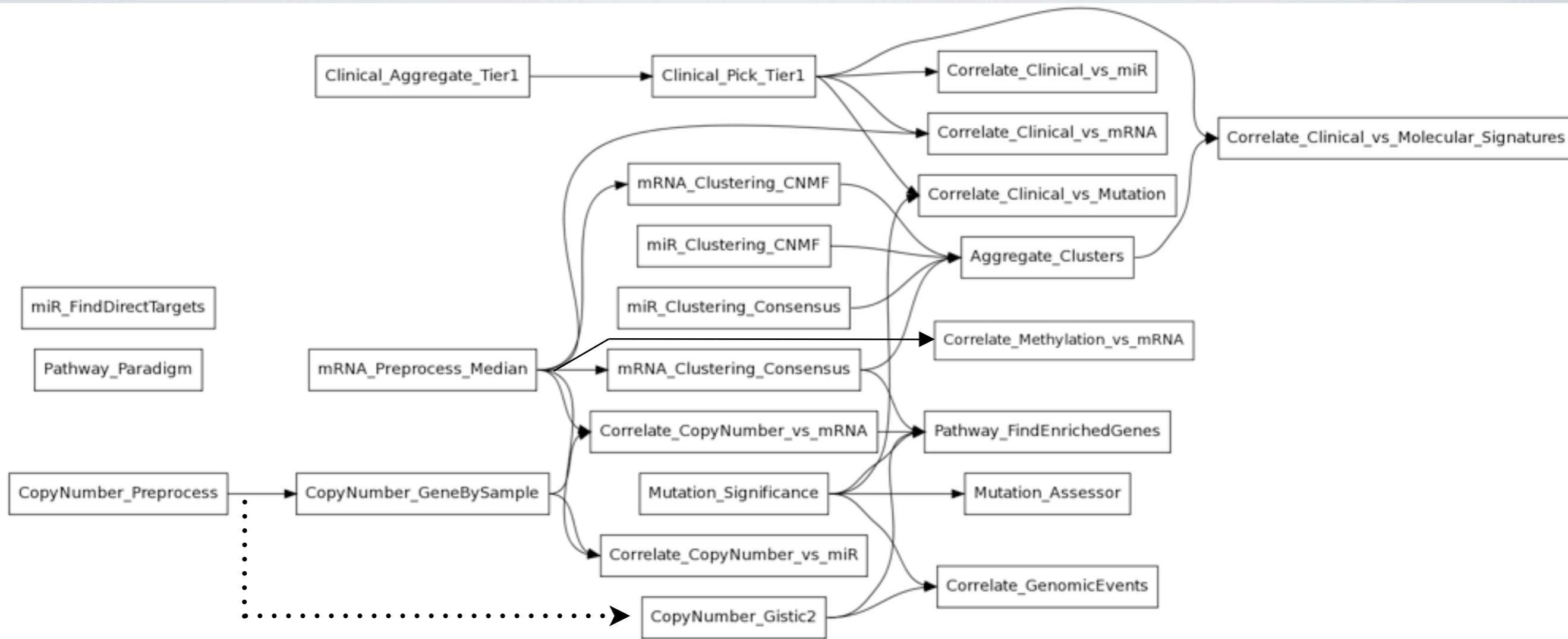
This ...



504 pipes and ~1000 GenePattern modules, per run

Insight 1:

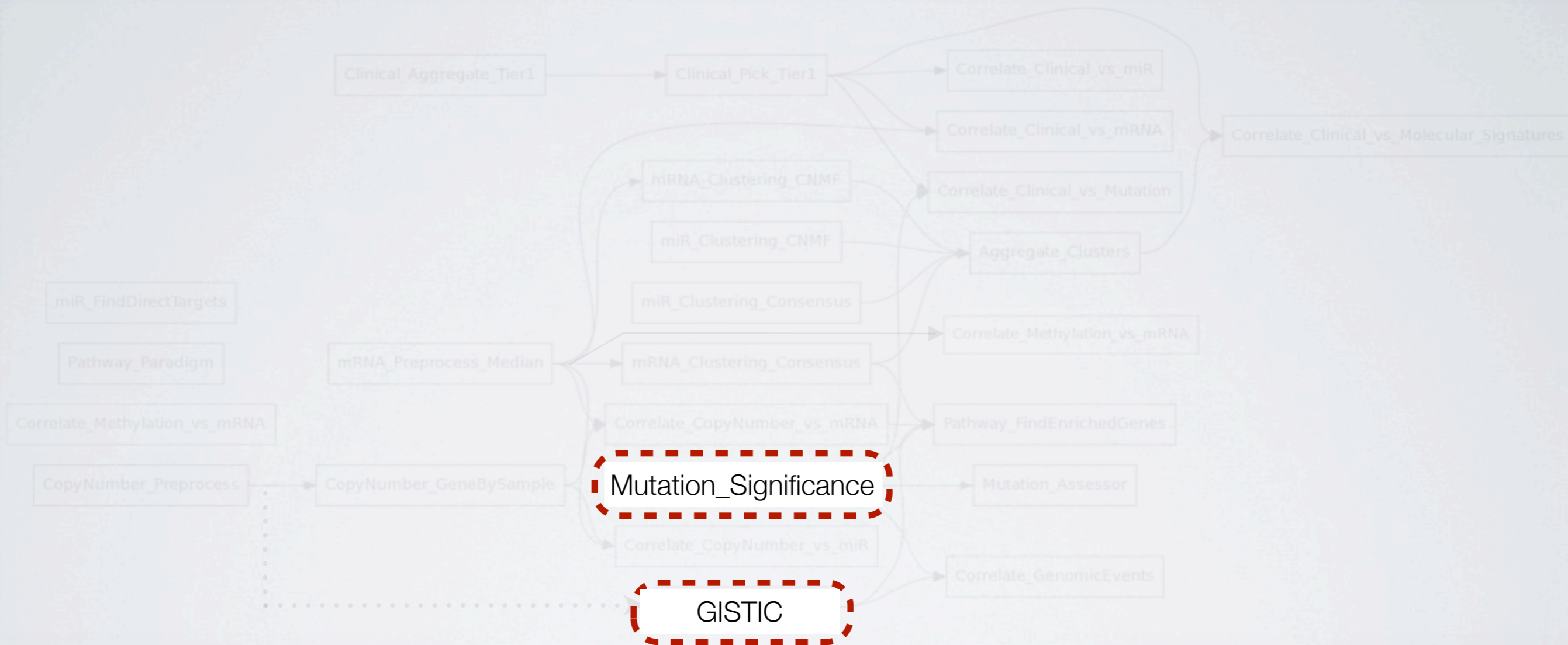
This ... is really a META-pipeline of pipelines



504 pipes and ~1000 GenePattern modules, per run

Insight 1:

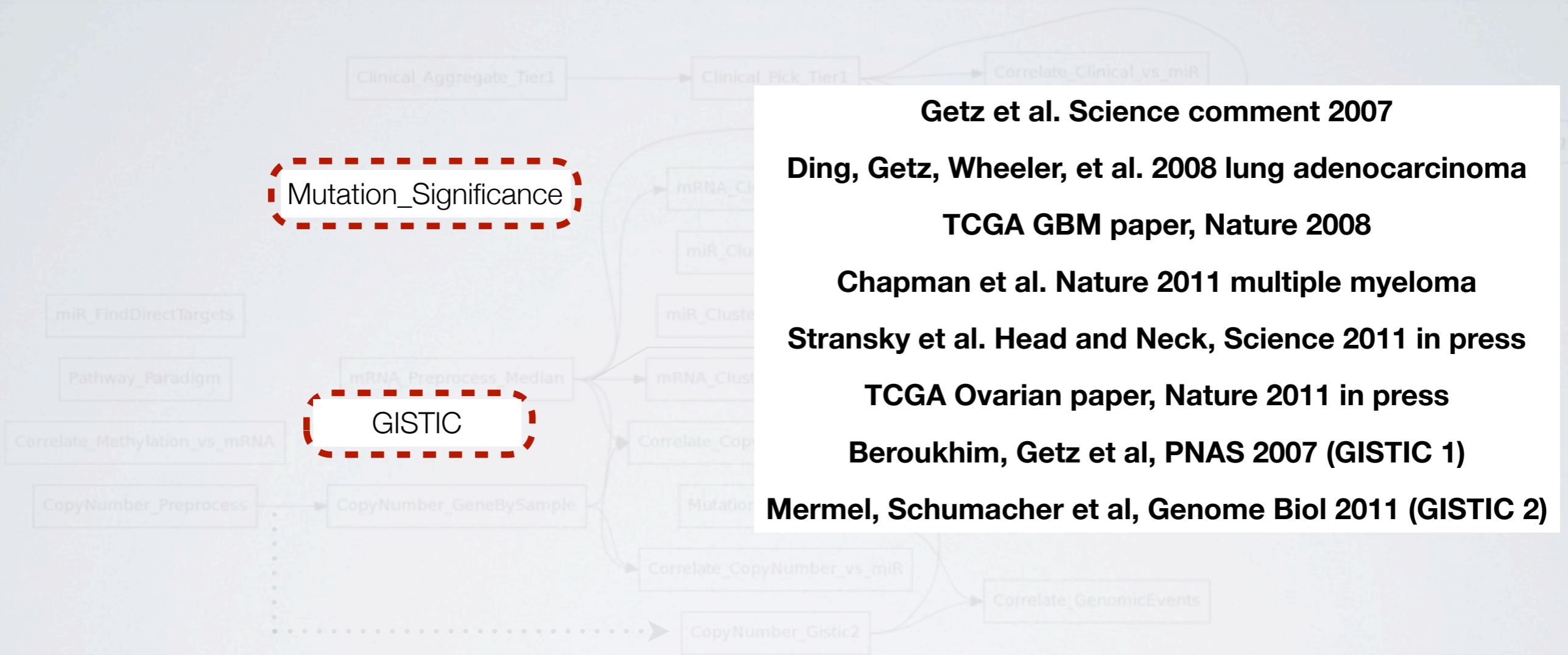
This ... is really a META-pipeline of pipelines



Some of which are themselves complex pipelined codes.

Insight 1:

This ... is really a META-pipeline of pipelines



Some of which are themselves complex pipelined codes.

Continuously evolving through years of publication use.


Like ENIAC, no simple task
to keep it all running

... in part because ...

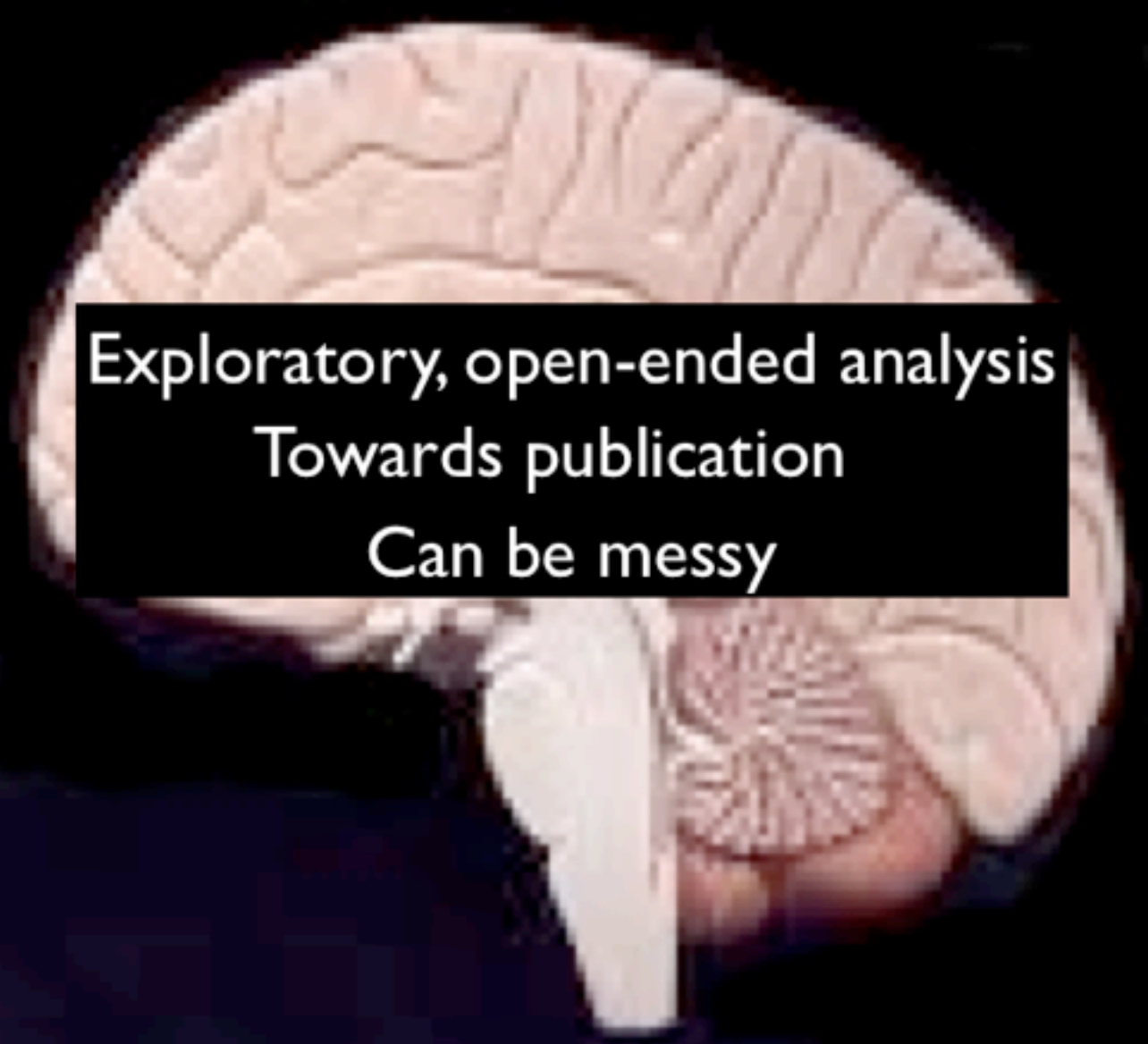
A Tale of Two Coders

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

Insight 2: So Unit Testing Not Enough

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

Insight 2: So Unit Testing Not Enough

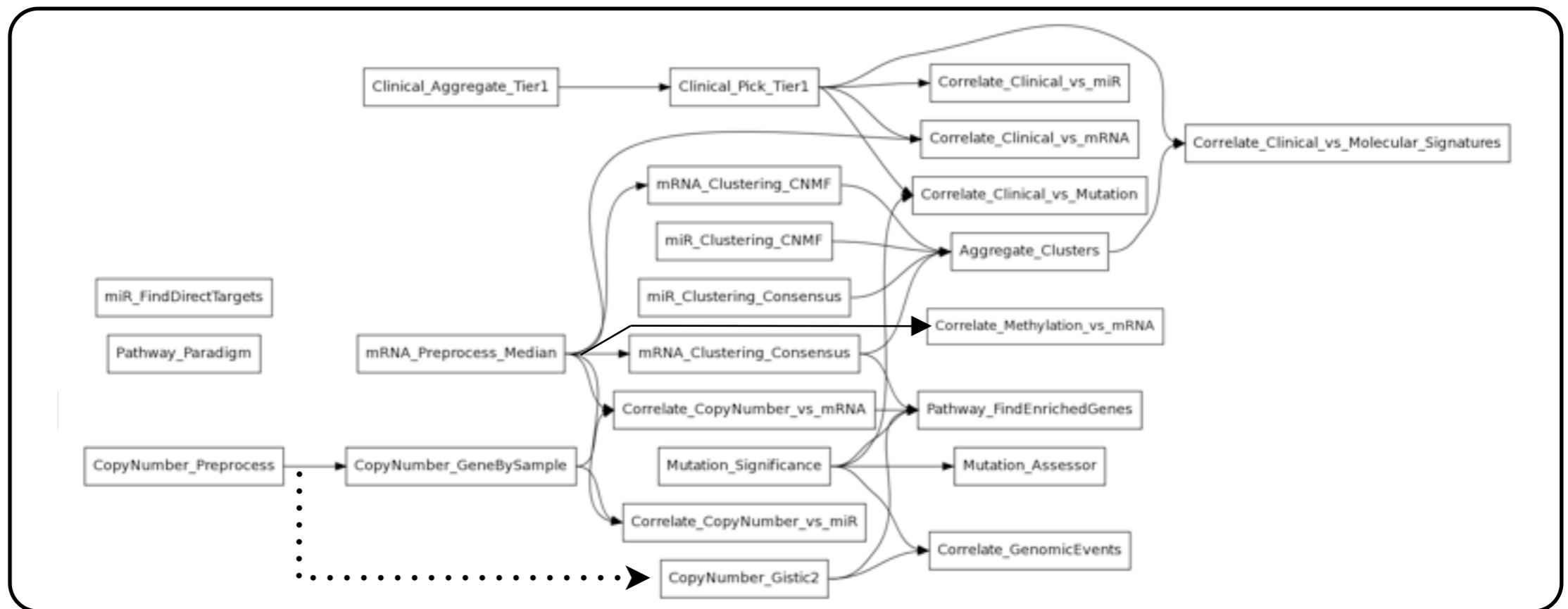
Individual researcher invoking THEIR code against THEIR data for THEIR paper, to establish that, in isolation, it runs to completion.

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

VITAL to maintain production operation of Firehose “data factory”

Insight 2: So Unit Testing Not Enough

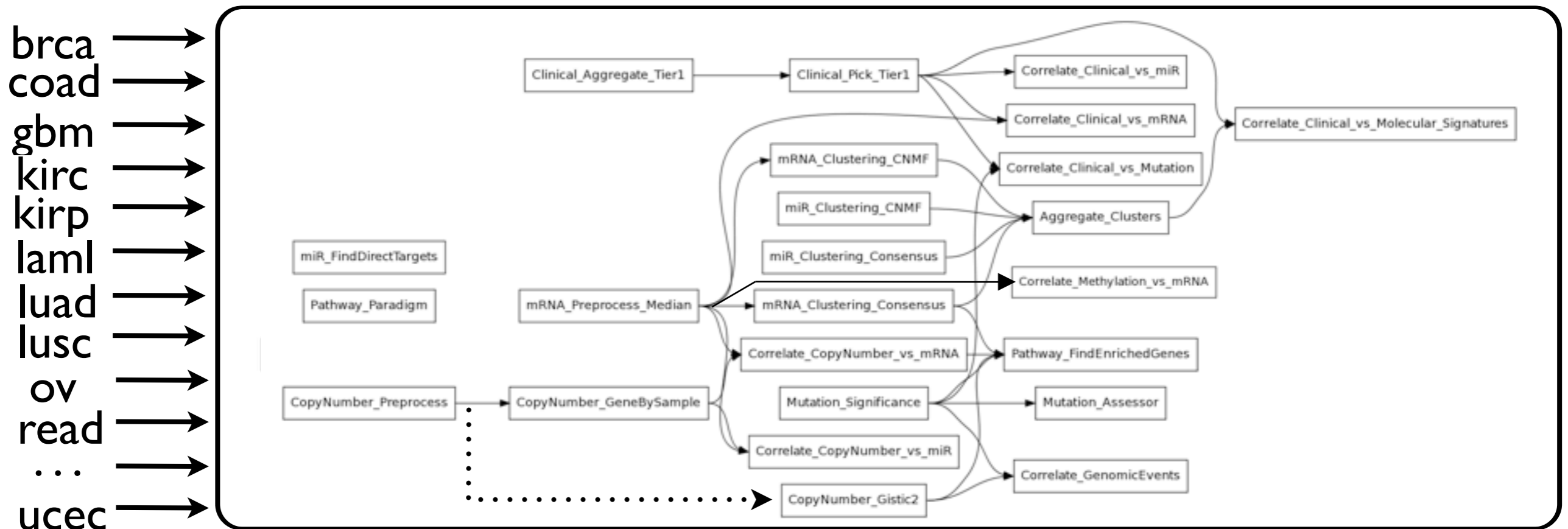
Individual researcher invoking THEIR code against THEIR data for THEIR paper, to establish that, in isolation, it runs to completion.



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

Insight 2: So Unit Testing Not Enough

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

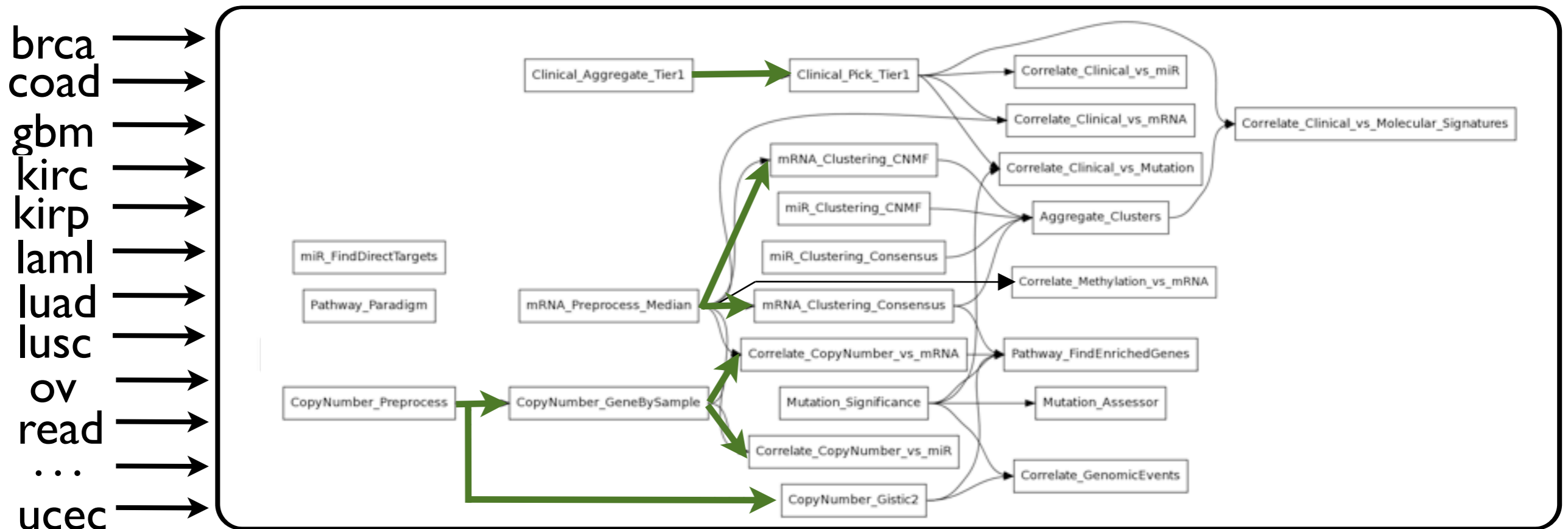


INTEGRATION TESTING must establish that (changes to) codes plays nice with rest of system.

Across datasets

Insight 2: So Unit Testing Not Enough

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

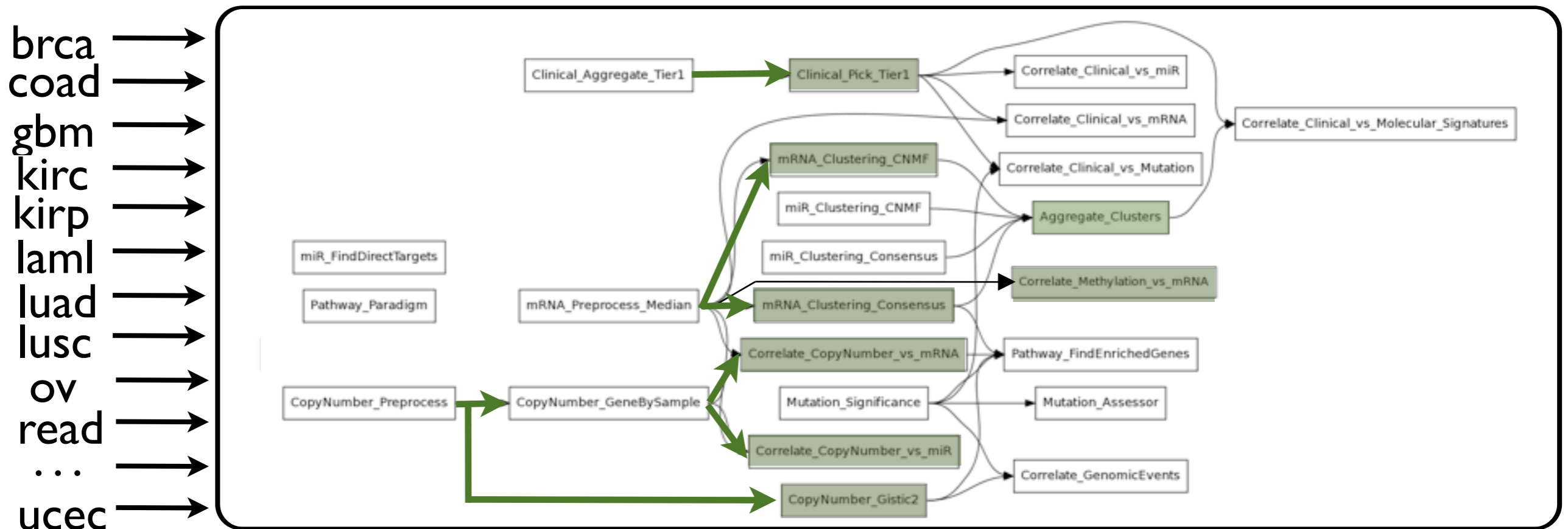


INTEGRATION TESTING must establish that (changes to) codes plays nice with rest of system.

Across datasets
With O's correctly wired to I's

Insight 2: So Unit Testing Not Enough

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

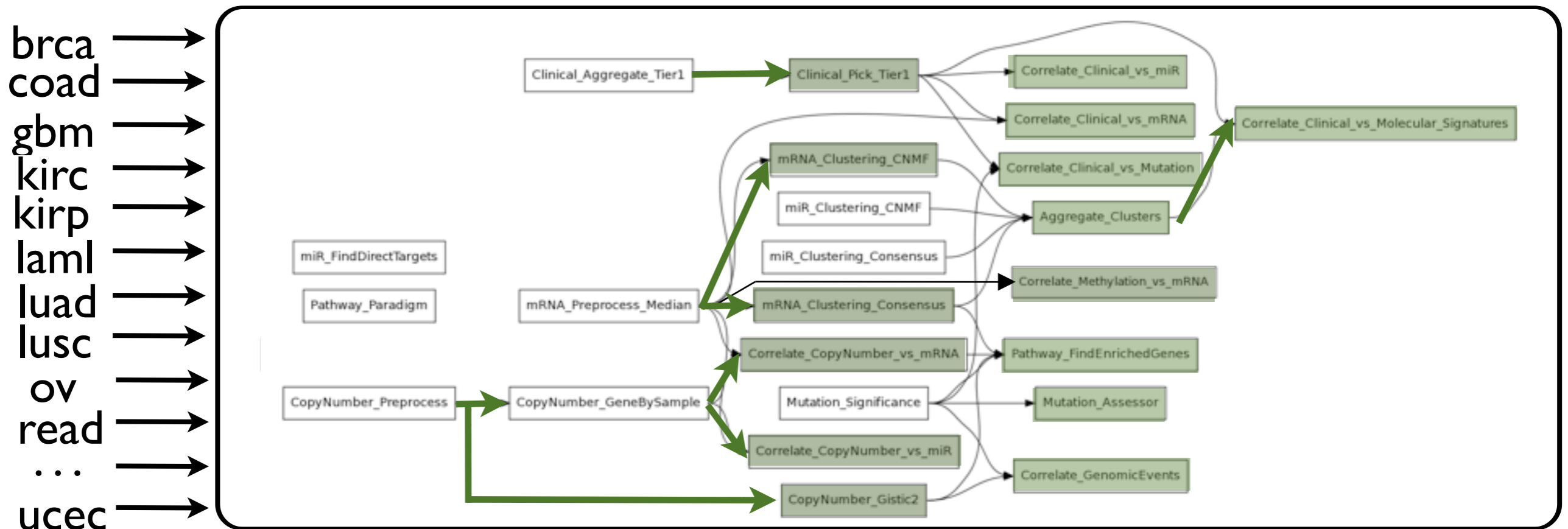


INTEGRATION TESTING must establish that
(changes to) codes plays nice with rest of system.

Across datasets Downstream dependents *correctly read* outputs
With O's correctly wired to I's

Insight 2: So Unit Testing Not Enough

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



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

Across datasets
With O's correctly wired to I's

Downstream dependents *correctly read* outputs
And remainder of workflow runs to completion

Insight 3:

Versioning and Automation are sacrosanct

Insight 3:

Versioning and Automation are sacrosanct

- Otherwise no reproducibility

Insight 3:

Versioning and Automation are sacrosanct

- Otherwise no reproducibility
- Or algorithmic scalability

Insight 3:

Versioning and Automation are sacrosanct

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

Insight 3:

Versioning and Automation are sacrosanct

- Otherwise no reproducibility
 - Or algorithmic scalability
 - BOTH code AND data are versioned
 - Do not trust: version and verify
 - Automation not just of pipelines:
- } Babel problem

Insight 3:

Versioning and Automation are sacrosanct

- Otherwise no reproducibility
- Or algorithmic scalability
- BOTH code AND data are versioned
- Do not trust: version and verify
- Automation not just of pipelines:
 - ✓ but also tools used to create them

} Babel
problem

**FH web services
Hydrant**

Insight 3:

Versioning and Automation are sacrosanct

- Otherwise no reproducibility
 - Or algorithmic scalability
 - BOTH code AND data are versioned
 - Do not trust: version and verify
 - Automation not just of pipelines:
 - ✓ but also tools used to create them
 - ✓ and reports generated from them
- } Babel problem
- FH web services
Hydrant
- GDAC website

Insight 3:

Versioning and Automation are sacrosanct

- Otherwise no reproducibility
 - Or algorithmic scalability
 - BOTH code AND data are versioned
 - Do not trust: version and verify
 - Automation not just of pipelines:
 - ✓ but also tools used to create them
 - ✓ and reports generated from them
 - ✓ and data sources which feed them
- } Babel problem
- FH web services
Hydrant
- GDAC website
- DCC, dbGAP

GUIs alone ARE NOT GOOD ENOUGH for these latter tasks
Because PROCESS SCALABILITY matters too

Insight 4: A- not good enough

Suppose all TCGA moving parts run 90% efficient

Insight 4: A- not good enough

Suppose all TCGA moving parts run 90% efficient

After just 4 steps in life
of TCGA sample:

$$.9^4 = 66\% \text{ overall efficiency}$$

Assume A = 95%

$$.95^4 = 81\%$$

And A⁺ = 99%

$$.99^4 = 96\%$$

Insight 4: A- not good enough

Suppose all TCGA moving parts run 90% efficient

After just 4 steps in life
of TCGA sample:

$$.9^4 = 66\% \text{ overall efficiency}$$

Assume $A = 95\%$

$$.95^4 = 81\%$$

And $A^+ = 99\%$

$$.99^4 = 96\%$$

Average sample
travels at
least $4 \cdot S_i$ steps:

Insight 4: A- not good enough

Suppose all TCGA moving parts run 90% efficient

After just 4 steps in life
of TCGA sample:

$$.9^4 = 66\% \text{ overall efficiency}$$

Assume A = 95%

$$.95^4 = 81\%$$

And A⁺ = 99%

$$.99^4 = 96\%$$

Average sample
travels at
least $4 \cdot S_i$ steps:

BCR → GCC/GSC → DCC → GDAC

Insight 4: A- not good enough

Suppose all TCGA moving parts run 90% efficient

After just 4 steps in life
of TCGA sample:

$$.9^4 = 66\% \text{ overall efficiency}$$

Assume A = 95%

$$.95^4 = 81\%$$

And A⁺ = 99%

$$.99^4 = 96\%$$

Average sample
travels at
least $4 \cdot S_i$ steps:

BCR → GCC/GSC → DCC → GDAC
Minimum $i=4$ centers, S_i steps within each

Insight 5:

Given that TCGA arguably largest/richest cancer data ever assembled

Insight 5:

Given that TCGA arguably largest/richest cancer data ever assembled

Discoveries lurk in our GDAC pipeline outputs

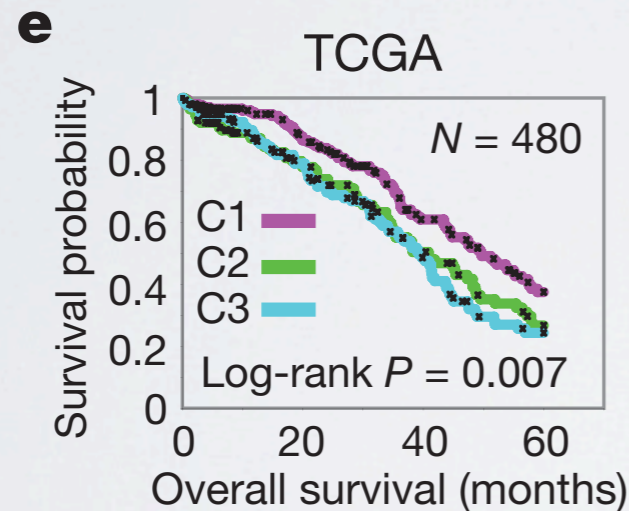
Insight 5:

Given that TCGA arguably largest/richest cancer data ever assembled

d

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

CNMF clustering of OV miR expression yielded 3 subtypes



One of which correlated to significantly longer survivability

***Integrated genomic analyses of ovarian carcinoma
TCGA Network, Nature, in press***

Discoveries lurk in our GDAC pipeline outputs

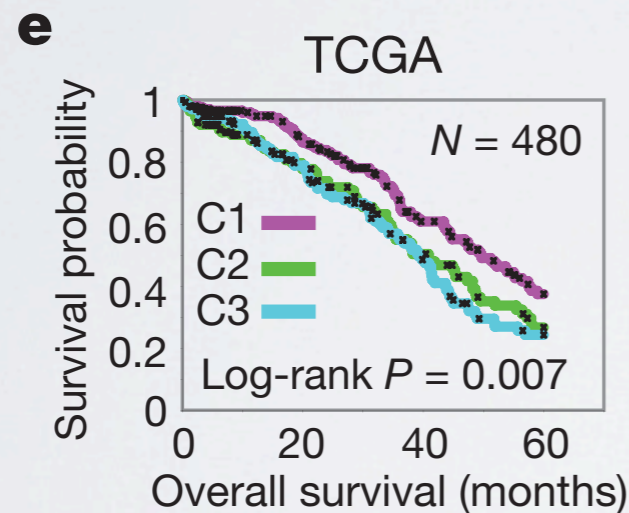
Insight 5:

Given that TCGA arguably largest/richest cancer data ever assembled

d

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

CNMF clustering of OV miR expression yielded 3 subtypes



One of which correlated to significantly longer survivability

***Integrated genomic analyses of ovarian carcinoma
TCGA Network, Nature, in press***

Discoveries lurk in our GDAC pipeline outputs

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

Insight 6:

Cross-tumor studies increasingly valuable

Insight 6:

Cross-tumor studies increasingly valuable

some genes mutated across tumor types

TP53

gene	description	N	n	npat	nsite	nsil	aml	dl	crc	gli	mel	mm	pr	ov	br	hn	nb	luc	S counts	S dust	S funct	S overall
TP53	tumor protein p53	1196851	437	420	231	6	6	8	3	49	1	3	0	283	29	18	0	37	10	10	10	10
PTEN	phosphatase and tensin homolog (m	1115222	61	60	50	0	0	0	0	48	1	0	1	2	3	3	0	3	10	4	10	10
EGFR	epidermal growth factor receptor (i	3827203	56	52	39	7	0	1	0	43	0	1	0	6	3	0	1	1	10	2	5	10
PIK3CA	phosphoinositide-3-kinase, catalyti	3034809	45	44	17	4	0	0	1	4	1	0	0	2	29	2	1	5	10	10	10	10
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral	680341	28	28	13	0	6	1	5	1	0	10	0	2	0	0	0	3	10	4	3	10
NRAS	neuroblastoma RAS viral (v-ras) on	578341	24	24	7	0	1	0	2	2	8	8	0	2	0	0	1	0	10	10	0	10
BRAF	v-raf murine sarcoma viral oncogen	2197251	18	17	8	2	0	0	0	1	12	1	0	2	1	0	0	1	10	10	4	10
FBXW7	F-box and WD repeat domain contai	2484107	21	18	18	0	0	4	1	3	4	0	0	3	1	2	0	3	9	4	3	9
IDH1	isocitrate dehydrogenase 1 (NADP+	1233336	11	11	3	0	2	1	0	8	0	0	0	0	0	0	0	0	8	10	2	8
CDKN2A	cydin-dependent kinase inhibitor 2	500455	14	14	13	0	1	0	0	4	0	0	0	0	1	1	0	7	10	0	1	8
SI	sucrase-isomaltase (alpha-glucosid	5402709	35	35	35	4	1	2	0	2	0	3	0	9	3	1	3	11	10	0	0	7
RB1	retinoblastoma 1 (including osteosa	2545919	24	24	24	1	0	0	0	11	0	1	0	9	0	1	0	2	9	1	0	7
MYD88	myeloid differentiation primary res	754554	7	7	3	0	0	7	0	0	0	0	0	0	0	0	0	0	8	5	0	7

Insight 6:

Cross-tumor studies increasingly valuable

some genes mutated across tumor types

TP53

some not

BRAF melanoma

gene	description	N	n	npat	nsite	nsil	aml	dl	crc	gbr	mel	mm	pr	ov	br	hn	nb	lusc	S counts	S dust	S funct	S overall
TP53	tumor protein p53	1196851	437	420	231	6	6	8	3	49	1	3	0	283	29	18	0	37	10	10	10	10
PTEN	phosphatase and tensin homolog (m	1115222	61	60	50	0	0	0	0	48	1	0	1	2	3	3	0	3	10	4	10	10
EGFR	epidermal growth factor receptor (i	3827203	56	52	39	7	0	1	0	43	0	1	0	6	3	0	1	1	10	2	5	10
PIK3CA	phosphoinositide-3-kinase, catalyti	3034809	45	44	17	4	0	0	1	4	1	0	0	2	29	2	1	5	10	10	10	10
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral	680341	28	28	13	0	6	1	5	1	0	10	0	2	0	0	0	3	10	4	3	10
NRAS	neuroblastoma RAS viral (v-ras) or	578341	24	24	7	0	1	0	2	2	8	8	0	2	0	0	1	0	10	10	0	10
BRAF	v-raf murine sarcoma viral oncoge	2197251	18	17	8	2	0	0	0	1	12	1	0	2	1	0	0	1	10	10	4	10
FBXW7	F-box and WD repeat domain contai	2484107	21	18	18	0	0	4	1	3	4	0	0	3	1	2	0	3	9	4	3	9
IDH1	isocitrate dehydrogenase 1 (NADP+	1233336	11	11	3	0	2	1	0	8	0	0	0	0	0	0	0	0	8	10	2	8
CDKN2A	cyclin-dependent kinase inhibitor 2	500455	14	14	13	0	1	0	0	4	0	0	0	0	1	1	0	7	10	0	1	8
SI	sucrase-isomaltase (alpha-glucosid	5402709	35	35	35	4	1	2	0	2	0	3	0	9	3	1	3	11	10	0	0	7
RB1	retinoblastoma 1 (including osteosa	2545919	24	24	24	1	0	0	0	11	0	1	0	9	0	1	0	2	9	1	0	7
MYD88	myeloid differentiation primary res	754554	7	7	3	0	0	7	0	0	0	0	0	0	0	0	0	0	8	5	0	7

gene	description	N	n	npat	nsite	nsil	aml	dl	crc	gbr	mel	mm	pr	ov	br	hn	nb	lusc	S counts	S dust	S funct	S overall
TP53	tumor protein p53	1196851	437	420	231	6	6	8	3	49	1	3	0	283	29	18	0	37	10	10	10	10
PTEN	phosphatase and tensin homolog (m	1115222	61	60	50	0	0	0	0	48	1	0	1	2	3	3	0	3	10	4	10	10
EGFR	epidermal growth factor receptor (i	3827203	56	52	39	7	0	1	0	43	0	1	0	6	3	0	1	1	10	2	5	10
PIK3CA	phosphoinositide-3-kinase, catalyti	3034809	45	44	17	4	0	0	1	4	1	0	0	2	29	2	1	5	10	10	10	10
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral	680341	28	28	13	0	6	1	5	1	0	10	0	2	0	0	0	3	10	4	3	10
NRAS	neuroblastoma RAS viral (v-ras) or	578341	24	24	7	0	1	0	2	2	8	8	0	2	0	0	1	0	10	10	0	10
BRAF	v-raf murine sarcoma viral oncoge	2197251	18	17	8	2	0	0	0	1	12	1	0	2	1	0	0	1	10	10	4	10
FBXW7	F-box and WD repeat domain contai	2484107	21	18	18	0	0	4	1	3	4	0	0	3	1	2	0	3	9	4	3	9
IDH1	isocitrate dehydrogenase 1 (NADP+	1233336	11	11	3	0	2	1	0	8	0	0	0	0	0	0	0	0	8	10	2	8
CDKN2A	cyclin-dependent kinase inhibitor 2	500455	14	14	13	0	1	0	0	4	0	0	0	0	1	1	0	7	10	0	1	8
SI	sucrase-isomaltase (alpha-glucosid	5402709	35	35	35	4	1	2	0	2	0	3	0	9	3	1	3	11	10	0	0	7
RB1	retinoblastoma 1 (including osteosa	2545919	24	24	24	1	0	0	0	11	0	1	0	9	0	1	0	2	9	1	0	7
MYD88	myeloid differentiation primary res	754554	7	7	3	0	0	7	0	0	0	0	0	0	0	0	0	0	8	5	0	7

MutSig: M. Lawrence, G. Getz, et al

Insight 6:

Cross-tumor studies increasingly valuable

some genes mutated across tumor types

TP53

some not

BRAF melanoma

gene	description	N	n	npat	nsite	nsil	aml	dl	crc	gbr	mel	mm	pr	ov	br	hn	nb	lusc	S_counts	S_dust	S_funct	S_overall
TP53	tumor protein p53	1196851	437	420	231	6	6	8	3	49	1	3	0	283	29	18	0	37	10	10	10	10
PTEN	phosphatase and tensin homolog (m	1115222	61	60	50	0	0	0	0	48	1	0	1	2	3	3	0	3	10	4	10	10
EGFR	epidermal growth factor receptor (c	3827203	56	52	39	7	0	1	0	43	0	1	0	6	3	0	1	1	10	2	5	10
PIK3CA	phosphoinositide-3-kinase, catalyti	3034809	45	44	17	4	0	0	1	4	1	0	0	2	29	2	1	5	10	10	10	10
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral	680341	28	28	13	0	6	1	5	1	0	10	0	2	0	0	0	3	10	4	3	10
NRAS	neuroblastoma RAS viral (v-ras) or	578341	24	24	7	0	1	0	2	2	8	8	0	2	0	0	1	0	10	10	0	10
BRAF	v-raf murine sarcoma viral oncoge	2197251	18	17	8	2	0	0	0	1	12	1	0	2	1	0	0	1	10	10	4	10
FBXW7	F-box and WD repeat domain contai	2484107	21	18	18	0	0	4	1	3	4	0	0	3	1	2	0	3	9	4	3	9
IDH1	isocitrate dehydrogenase 1 (NADP+	1233336	11	11	3	0	2	1	0	8	0	0	0	0	0	0	0	0	8	10	2	8
CDKN2A	cyclin-dependent kinase inhibitor 2	500455	14	14	13	0	1	0	0	4	0	0	0	0	1	1	0	7	10	0	1	8
SI	sucrase-isomaltase (alpha-glucosid	5402709	35	35	35	4	1	2	0	2	0	3	0	9	3	1	3	11	10	0	0	7
RB1	retinoblastoma 1 (including osteosa	2545919	24	24	24	1	0	0	0	11	0	1	0	9	0	1	0	2	9	1	0	7
MYD88	myeloid differentiation primary res	754554	7	7	3	0	0	7	0	0	0	0	0	0	0	0	0	0	8	5	0	7

gene	description	N	n	npat	nsite	nsil	aml	dl	crc	gbr	mel	mm	pr	ov	br	hn	nb	lusc	S_counts	S_dust	S_funct	S_overall
TP53	tumor protein p53	1196851	437	420	231	6	6	8	3	49	1	3	0	283	29	18	0	37	10	10	10	10
PTEN	phosphatase and tensin homolog (m	1115222	61	60	50	0	0	0	0	48	1	0	1	2	3	3	0	3	10	4	10	10
EGFR	epidermal growth factor receptor (c	3827203	56	52	39	7	0	1	0	43	0	1	0	6	3	0	1	1	10	2	5	10
PIK3CA	phosphoinositide-3-kinase, catalyti	3034809	45	44	17	4	0	0	1	4	1	0	0	2	29	2	1	5	10	10	10	10
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral	680341	28	28	13	0	6	1	5	1	0	10	0	2	0	0	0	3	10	4	3	10
NRAS	neuroblastoma RAS viral (v-ras) or	578341	24	24	7	0	1	0	2	2	8	8	0	2	0	0	1	0	10	10	0	10
BRAF	v-raf murine sarcoma viral oncoge	2197251	18	17	8	2	0	0	0	1	12	1	0	2	1	0	0	1	10	10	4	10
FBXW7	F-box and WD repeat domain contai	2484107	21	18	18	0	0	4	1	3	4	0	0	3	1	2	0	3	9	4	3	9
IDH1	isocitrate dehydrogenase 1 (NADP+	1233336	11	11	3	0	2	1	0	8	0	0	0	0	0	0	0	0	8	10	2	8
CDKN2A	cyclin-dependent kinase inhibitor 2	500455	14	14	13	0	1	0	0	4	0	0	0	0	1	1	0	7	10	0	1	8
SI	sucrase-isomaltase (alpha-glucosid	5402709	35	35	35	4	1	2	0	2	0	3	0	9	3	1	3	11	10	0	0	7
RB1	retinoblastoma 1 (including osteosa	2545919	24	24	24	1	0	0	0	11	0	1	0	9	0	1	0	2	9	1	0	7
MYD88	myeloid differentiation primary res	754554	7	7	3	0	0	7	0	0	0	0	0	0	0	0	0	0	8	5	0	7

MutSig: M. Lawrence, G. Getz, et al

Firehose makes these cross-tumor analyses comparatively easy & automatic

Insight 7 : Clinical Correlations Hard

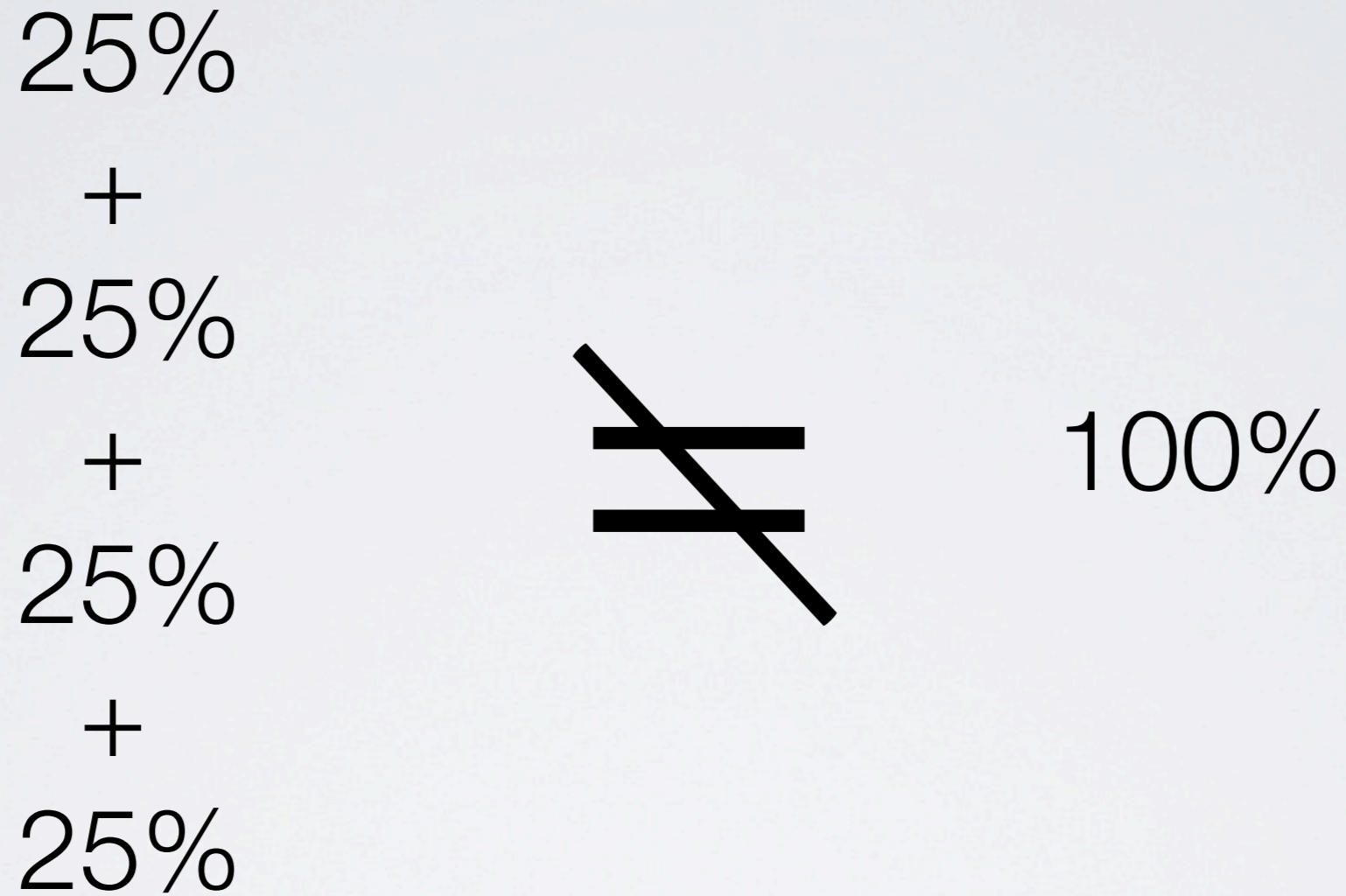
Flux in parameter definitions stymie automation
Manual transcribing implies uncertainty

Program Office working hard to stabilize
But @ Broad we don't trust our own results:

Example: % BRCA samples with male gender too high

Removed from standard production runs (July 2011)
Will reappear in provisional runs (Nov/Dec 2011)

Insight 8 : Context Switching is Costly



Fewer full-timers > more part-timers

For More Information


Poster 76 : Firehose infrastructure (D.Voet)

Poster 58 : Firehose data standardization (G. Saksena)

Poster 58 : Integrative Genomics Viewer (J. Robinson)

Dashboard > Broad TCGA GDAC > Home Browse Michael Noble Search

> FAQ

 **FIREHOSE**
Broad GDAC

FAQ Edit Add Tools

Added by Michael Noble, last edited by Michael Noble on Nov 15, 2011 (view change)

Frequently Asked Questions

Q: When is the next run?

A: As of November 2011 the Broad Institute GDAC will aim to provide 3 runs per month:

- Standard Data Run: started on 1st of month
- Standard Data Run: started on 15th of month
- Analysis Run: started shortly after second data run completes

Q: What reference genome build are you using?

A: Presently we are using hg18, but recognize the need to transition to hg19 as soon as possible. Our understanding is that TCGA standards stipulate that OV, GBM, COAD/READ, and LAML data are hg18, and all else is hg19.

Q: How or where can I access the results of a run?

A: In one of two ways:

- Both analyses and standardized data are stored in the [Broad repository of the TCGA Data Coordination Center \(DCC\)](#). After signing in (TCGA credentials required), you should see something like

Index of /tcgfiles/ftp_auth/distro_ftpusers/tcga4yeo/other/gdacs/gdacbread

name	Last modified	Size
Parent Directory	08-Oct-2011 10:34	-
analysis/	04-Feb-2011 13:33	411
blca/	08-Oct-2011 11:03	-
brca/	08-Oct-2011 10:56	-
ccsk/	08-Oct-2011 11:03	-
coad/	08-Oct-2011 10:56	-
coadread/	08-Oct-2011 11:01	-
csll/	08-Oct-2011 10:56	-
gbm/	08-Oct-2011 10:56	-
hnscc/	08-Oct-2011 11:03	-
lihc/	08-Oct-2011 10:57	-
lihp/	08-Oct-2011 10:58	-
laml/	08-Oct-2011 10:58	-
lgs/	08-Oct-2011 11:03	-
liver/	08-Oct-2011 11:03	-
lsc/	08-Oct-2011 10:58	-
lscv/	08-Oct-2011 10:58	-
ov/	08-Oct-2011 10:58	-
panc/	08-Oct-2011 11:03	-
read/	08-Oct-2011 11:03	-
read/	08-Oct-2011 11:01	-
report/	12-Oct-2011 14:12	-
stml/	08-Oct-2011 11:01	-
thca/	08-Oct-2011 11:03	-
ucec/	08-Oct-2011 11:01	-

from which you may simply navigate to the tumor type and run date of interest.

- Standardized data packages can also be viewed directly within your [local IGV installation](#), without signing in to the DCC, by following [the instructions given here](#).

WWW
Email

<http://gdac.broadinstitute.org>
gdac@broadinstitute.org

Broad GDAC Analysis Summary 2011_05_25 Run

Tables of Ingested Data: [HTML](#) [PNG](#) [TSV](#)

Tumor Type	# Completed	Percentage
OV	24	<u>100%</u>
GBM	24	<u>100%</u>
READ	17	<u>71%</u>
LUSC	17	<u>71%</u>
LUAD	17	<u>71%</u>
COAD	17	<u>71%</u>
COADREAD	17	<u>71%</u>
BRCA	12	<u>50%</u>
KIRC	10	<u>42%</u>
KIRP	7	<u>29%</u>
UCEC	4	<u>17%</u>
LGG	4	<u>17%</u>
CESC	4	<u>17%</u>
BLCA	4	<u>17%</u>
STAD	3	<u>13%</u>
LIHC	3	<u>13%</u>
HNSC	3	<u>13%</u>
THCA	2	<u>8%</u>
PRAD	2	<u>8%</u>
LAML	2	<u>8%</u>

What Analyses?
Look at our dashboard ...

WWW gdac.broadinstitute.org

Email gdac@broadinstitute.org

Broad GDAC Analysis Summary

2011_05_25 Run

Tables of Ingested Data: [HTML](#) [PNG](#) [TSV](#)

Tumor Type	# Completed	Percentage
OV	24	<u>100%</u>
GBM	24	<u>100%</u>
READ	17	<u>71%</u>
LUSC	17	<u>71%</u>
LUAD	17	<u>71%</u>
COAD	17	<u>71%</u>
COADREAD	17	<u>71%</u>
BRCA	12	<u>50%</u>
KIRC	10	<u>42%</u>
KIRP	7	<u>29%</u>
UCEC	4	<u>17%</u>
LGG	4	<u>17%</u>
CESC	4	<u>17%</u>
BLCA	4	<u>17%</u>
STAD	3	<u>13%</u>
LIHC	3	<u>13%</u>
HNSC	3	<u>13%</u>
THCA	2	<u>8%</u>
PRAD	2	<u>8%</u>
LAML	2	<u>8%</u>

TumorType	Biospecimen	Any_Level_1	Clinical	CNA	Methylation	mRNA	miR	MAF
BLCA	35	12	11	9	0	0	0	0
BRCA	704	524	358	507	186	434	0	0
CESC	40	8	5	8	0	0	0	0
COAD	245	202	208	186	167	155	0	102
COADREAD	338	276	287	257	236	224	0	158
GBM	547	511	465	498	288	499	415	199
HNSC	97	59	0	57	0	0	0	0
KIRC	460	453	241	448	219	72	0	0
KIRP	75	16	17	16	36	41	0	0
LAML	202	0	0	0	188	0	178	135
LGG	58	30	19	30	0	0	0	0
LIHC	45	38	0	37	0	0	0	0
LUAD	158	59	47	58	128	33	0	122
LUSC	184	184	72	142	133	134	0	150
OV	592	570	528	519	425	570	566	383
PRAD	65	65	0	64	0	0	0	0
READ	93	74	79	71	69	69	0	56
STAD	111	35	0	81	82	0	0	0
THCA	39	25	0	24	0	0	0	0
UCEC	325	220	127	215	70	0	0	0
Totals	4075	3085	2177	2970	1991	2007	1159	1147

	Pipeline	Not Ready	Failed	Succeed
1	Aggregate_Clusters	0	0	1
2	Clinical_Aggregate_Tier1	0	0	1
3	Clinical_Pick_Tier1	0	0	1
4	CopyNumber_GeneBySample	0	0	1
5	CopyNumber_Gistic2	0	0	1
6	CopyNumber_Preprocess	0	0	1
7	Correlate_Clinical_vs_miR	0	0	1
8	Correlate_Clinical_vs_Molecular_Signatures	0	0	1
9	Correlate_Clinical_vs_mRNA	0	0	1
10	Correlate_Clinical_vs_Mutation	0	0	1
11	Correlate_CopyNumber_vs_miR	0	0	1
12	Correlate_CopyNumber_vs_mRNA	0	0	1
13	Correlate_GenomicEvents	0	0	1
14	Correlate_Methylation_vs_mRNA	0	0	1
15	miR_Clustering_CNMF	0	0	1
16	miR_Clustering_Consensus	0	0	1
17	miR_FindDirectTargets	0	0	1
18	mRNA_Clustering_CNMF	0	0	1
19	mRNA_Clustering_Consensus	0	0	1
20	mRNA_Preprocess_Median	0	0	1
21	Mutation_Assessor	0	0	1
22	Mutation_Significance	0	0	1
23	Pathway_FindEnrichedGenes	0	0	1
24	Pathway_Paradigm	0	0	1
Total		0	0	24

[WWW](http://www.gdac.broadinstitute.org)

gdac.broadinstitute.org

[Email](mailto:gdac@broadinstitute.org)

gdac@broadinstitute.org

THANK YOU!