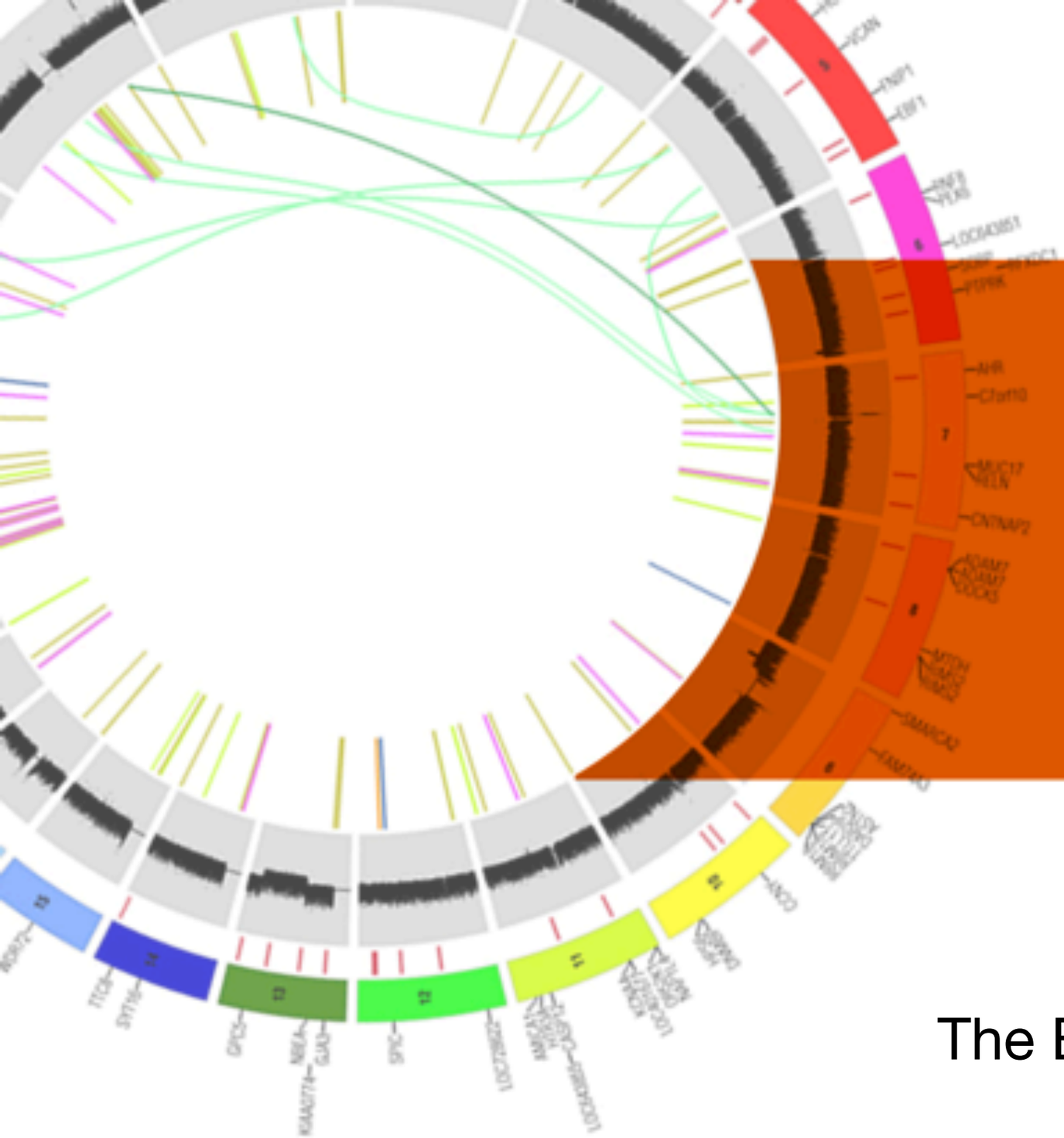


The Cancer Genome Atlas

Firehose Overview

M. Noble
The Broad Institute of MIT & Harvard

January 31, 2012

BROAD INSTITUTE

FIREHOSE Broad GDAC

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

THE CANCER GENOME ATLAS

Outline

I. Why (yet another pipeline)?

II. What (is Firehose, anyway)?

III. How (will it help)?

IV. Insights (gained so far)

I : Why?

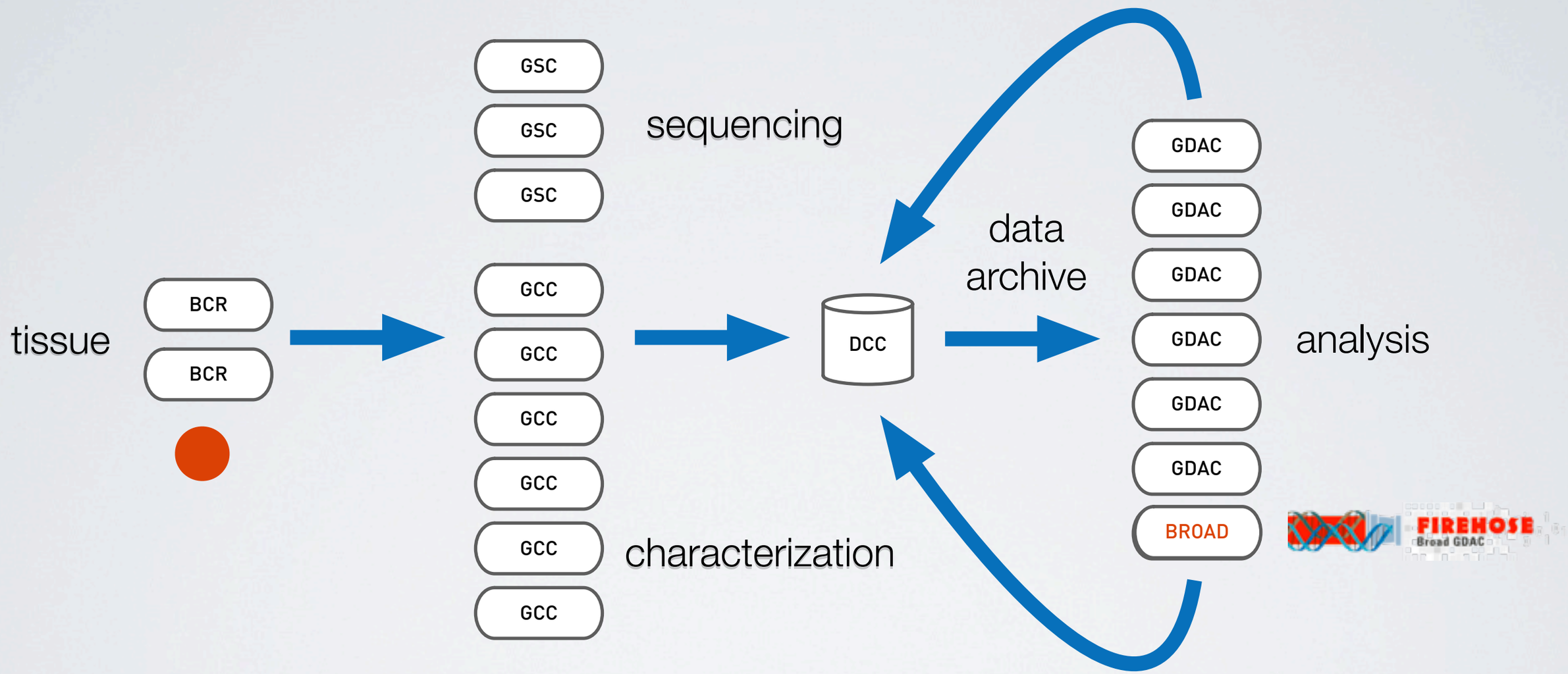
TCGA

ACRONYM: THE CANCER GENOME ATLAS

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

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?

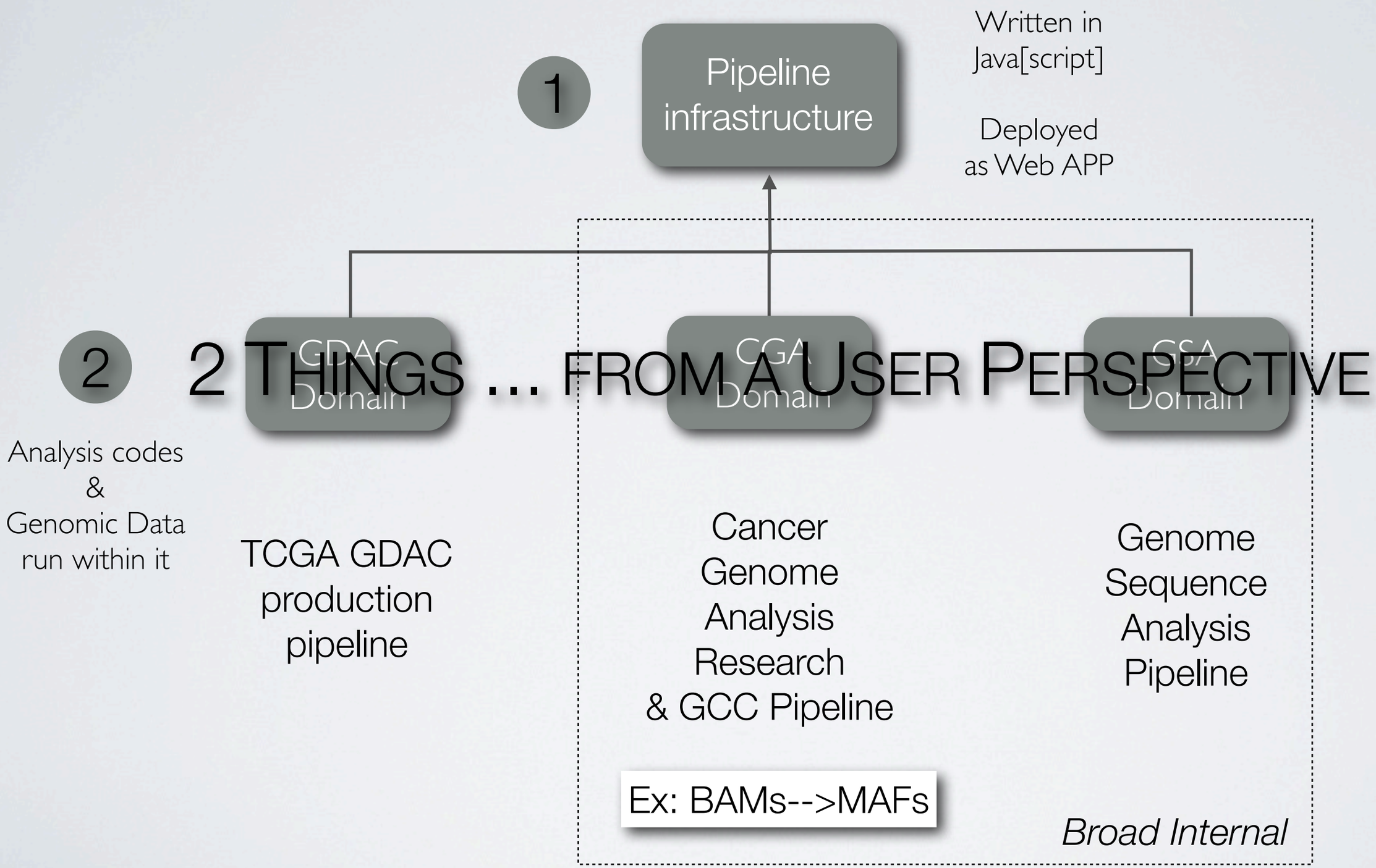
- There is no one-size-fits-all, cookie-cutter method to answer such questions
- But some analyses are common to many questions and can be automated:
  - Mutation calling, classifying, summarizing and significance-testing
  - Copy number alteration detection and significance-testing
  - Expression- and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

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



Providing

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

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

Summary of TCGA Tumor Data
Ingested into Broad GDAC Pipeline
2012_01_24 stddata Run

Tumor	BCR	Clinical	CN	Methylation	mRNA	mRNAseq	miR	miRseq	MAF	Protein
BLCA	59	36	35	0	0	0	0	54	0	
BRCA	855	825	781	313	529	450	0	781	507	
CESC	75	6	36	0	0	0	0	8	0	
COADREAD	591	584	565	236	224	78	0	255	224	
DLBC	10	0	0	0	0	0	0	0	0	
GBM	596	544	537	287	542	0	491	0	276	
HNSC	263	163	165	0	0	5	0	89	0	
KIRC	502	502	489	219	72	469	0	463	327	
KIRP	107	63	43	36	16	14	0	16	0	
LAML	202	200	0	192	0	179	0	187	199	
LGG	119	85	80	0	27	0	0	30	0	
LIHC	59	42	53	0	0	17	0	28	0	
LNNH	2	0	0	0	0	0	0	0	0	
LUAD	331	235	205	127	32	0	0	95	147	
LUSC	283	225	211	133	154	220	0	202	178	
OV	592	580	547	551	568	0	564	46	316	
PAAD	14	0	14	0	0	0	0	0	0	
PRAD	153	0	82	0	0	0	0	63	0	
SKCM	219	0	0	0	0	0	0	0	0	
STAD	149	148	134	66	0	58	0	125	0	
THCA	230	42	85	0	0	0	0	45	0	
UCEC	448	391	363	117	54	266	0	359	239	
Totals	5859	4671	4425	2277	2218	1756	1055	2846	2413	

RPPA Status @ Broad GDAC

- All available TCGA RPPA data mirrored At Broad

  ✓ brca
  ✓ coad
  ✓ gbm
  ✓ kirc
  ✓ ov
  ✓ read
  ✓ ucec

  7 tumor types so far

- Anticipated in Production Data Pipeline in Feb 2012

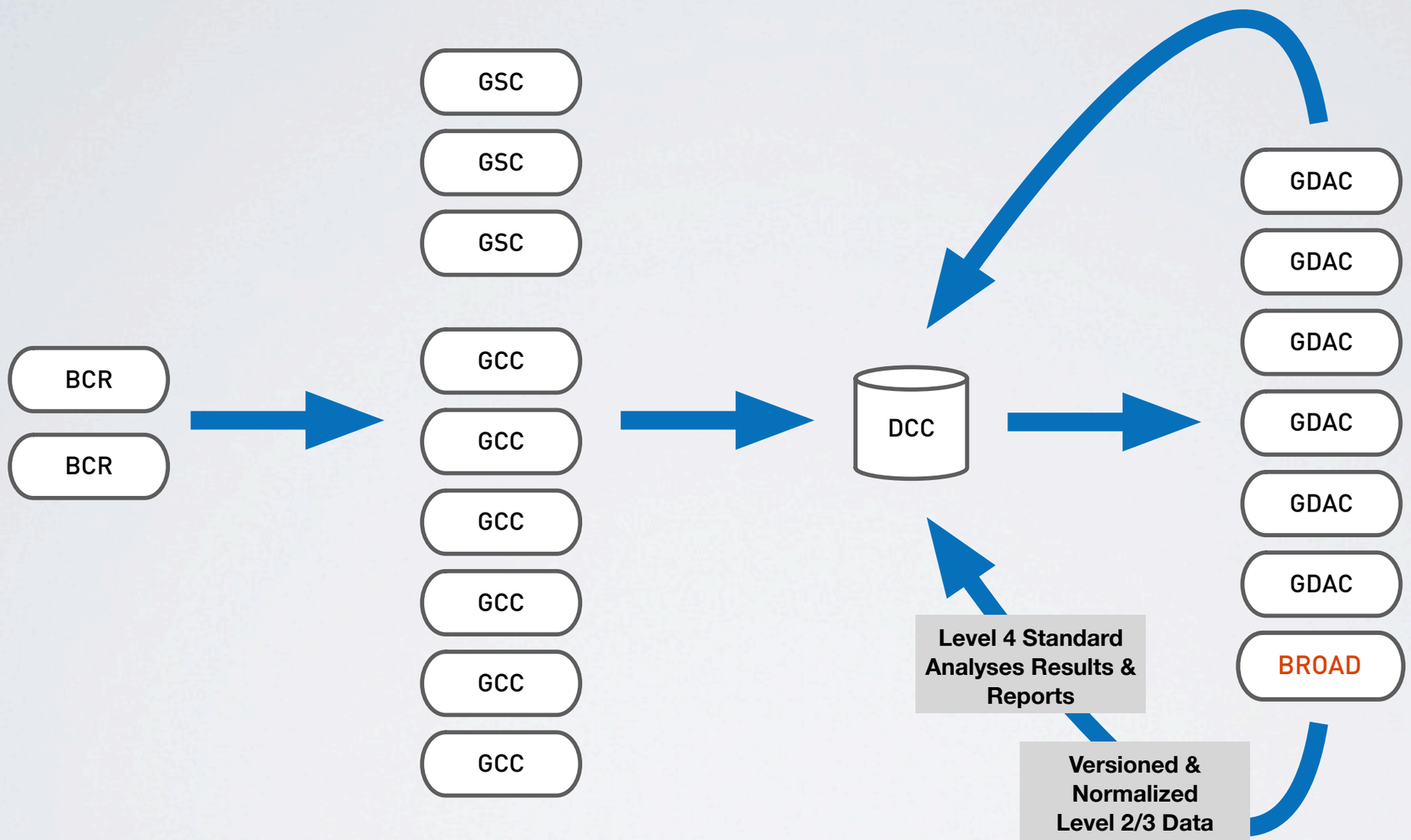
- Analysis pipelines under development:

  collaboration with MD Anderson GDAC (Chin et al)

- Contact: Spring Liu (yingchun@broadinstitute.org)

III : How?

Where Firehose Lives In TCGA



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

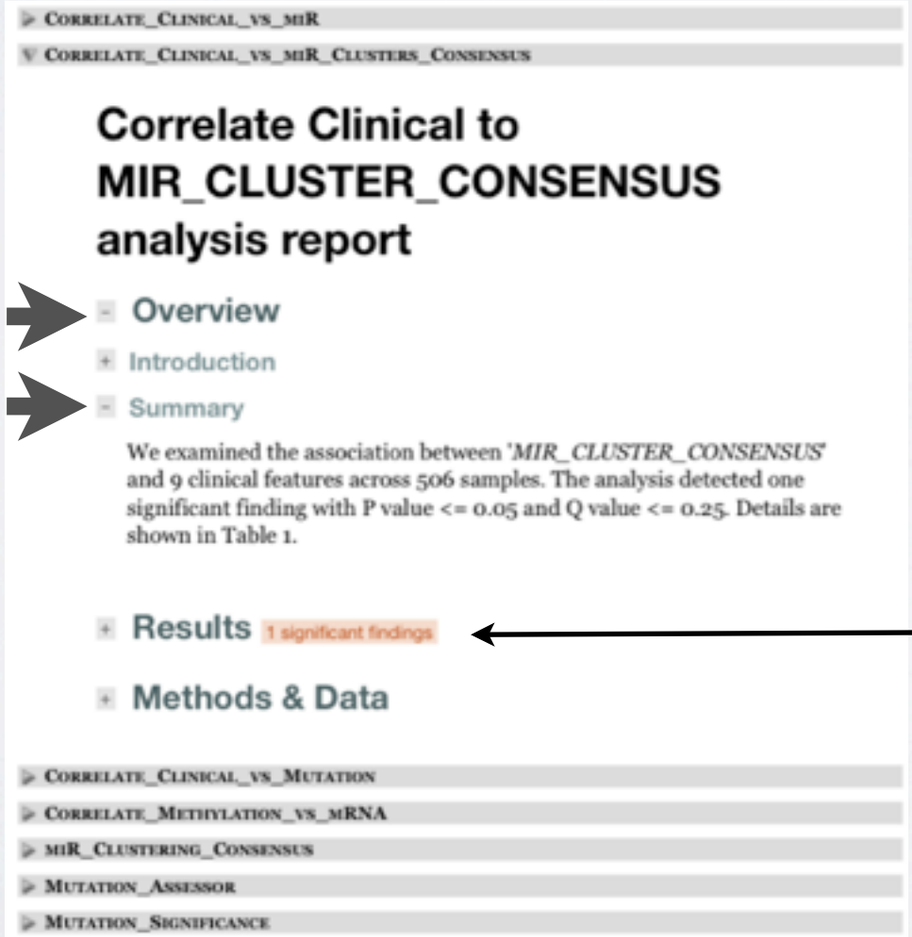
   Generated at Broad, precursor to automated pipeline

These broadly map to 3 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 <u>EASILY COMPREHENDED SUMMARY REPORTS</u>
- LIKE DRAFT RESULTS SECTION ... SANS PUBLICATION DELAY

Nozzle : Analyst & Biologist-Friendly Reports

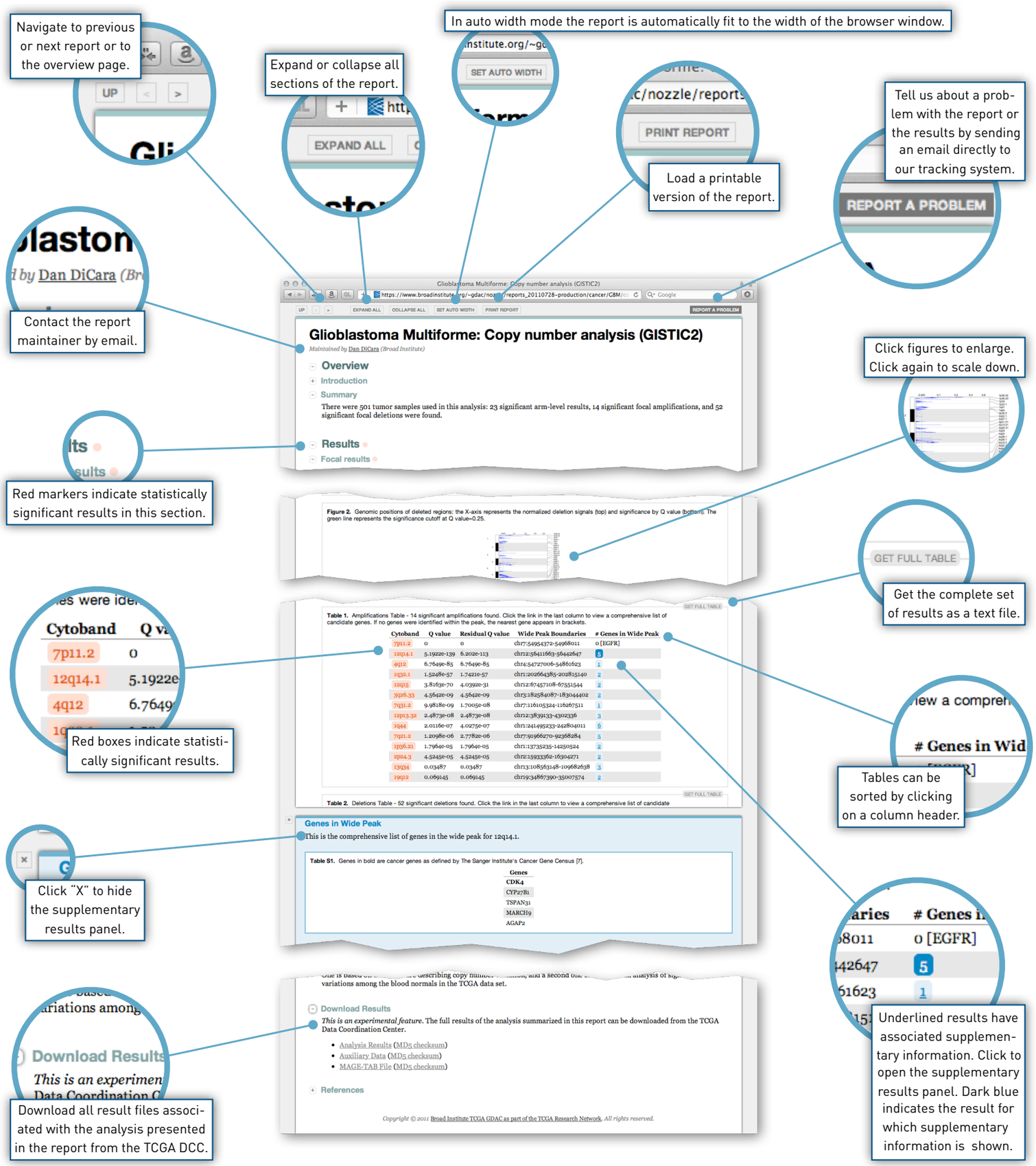


- Standard visual format for ALL pipelines
- Intelligent Scoping:
  - drill from overview to details
  - Significant results “bubble up”
- **don’t miss needle in haystack**

Firehose Reports | At-a-Glance

BROAD INSTITUTE

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



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 Petar 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

GET FULL TABLE

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
10	OR5D16	olfactory receptor, family 5, subfamily D, member 16	295338	4	2	0	1	1	0	0.00015	0.29
11	ZNF470	zinc finger protein 470	314800	5	0	2	2	1	0	0.00017	0.30

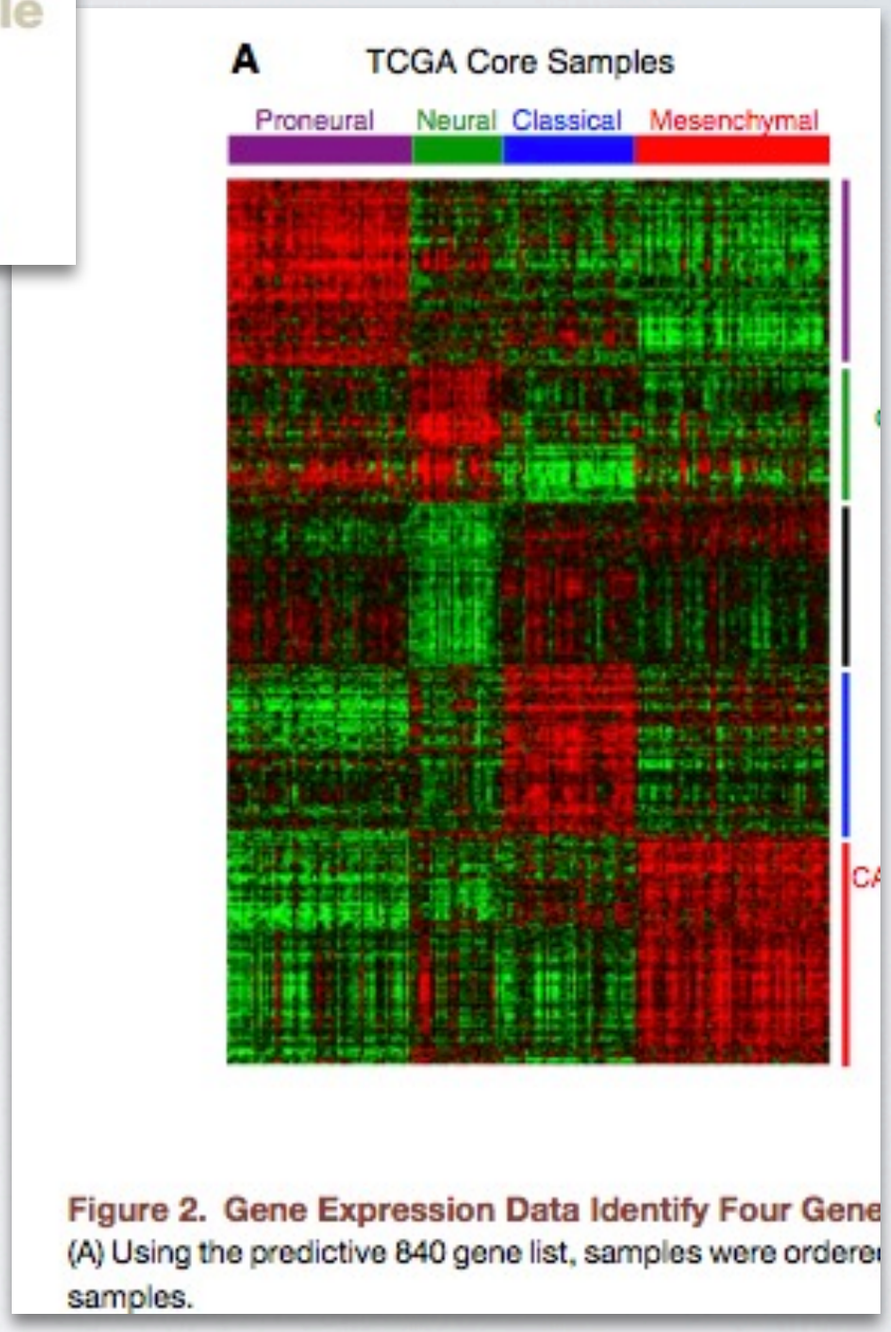
Table 2 | Significantly mutated genes in HGS-OvCa

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



Gene Expression Clustering

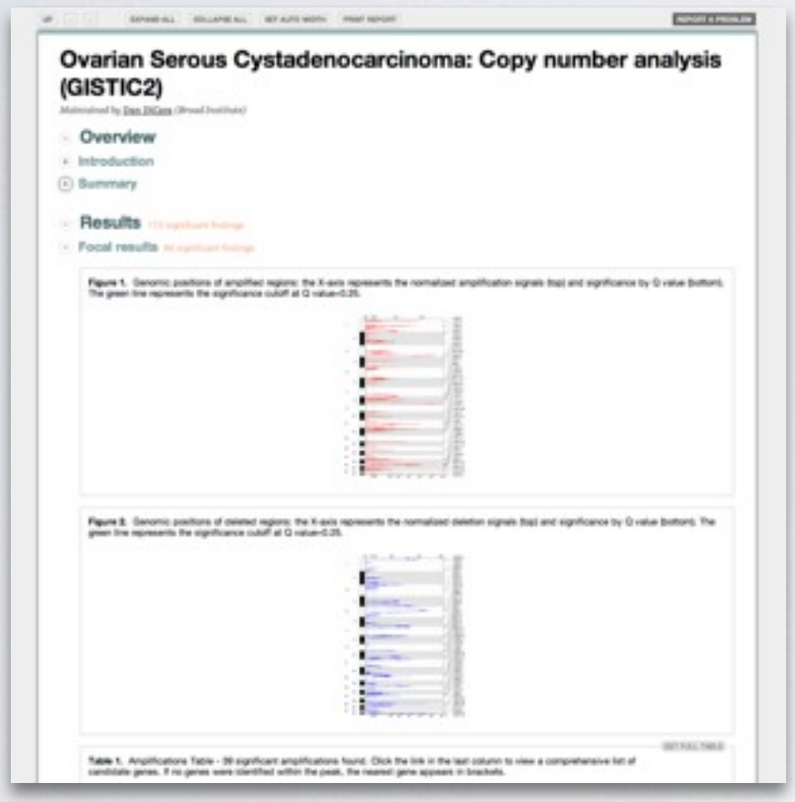
Firehose Reports: Example 3

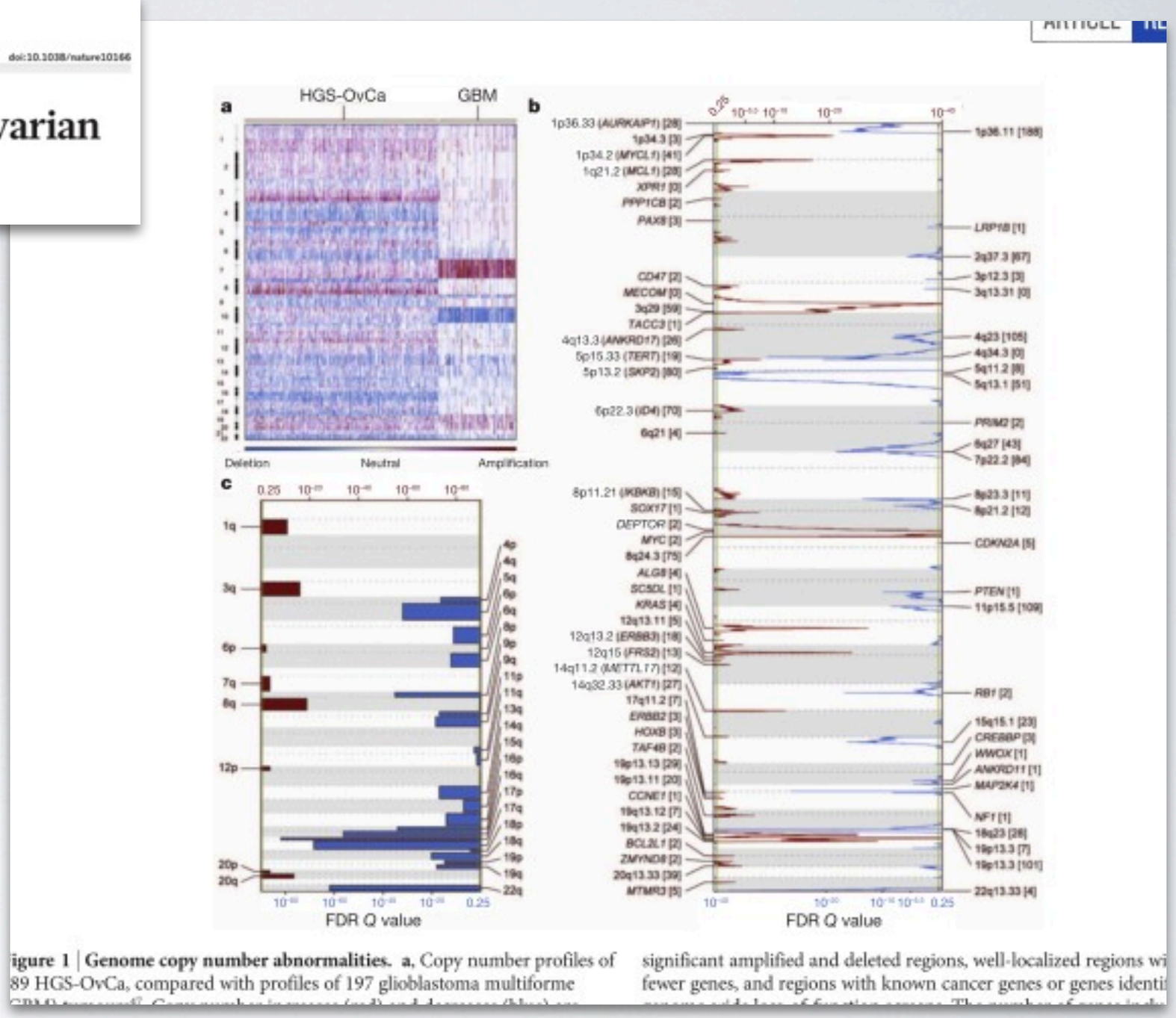

ARTICLE

doi:10.1038/nature10166

Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*





igure 1 | Genome copy number abnormalities. a, Copy number profiles of 89 HGS-OvCa, compared with profiles of 197 glioblastoma multiforme

significant amplified and deleted regions, well-localized regions wi fewer genes, and regions with known cancer genes or genes identi

Copy Number Alterations

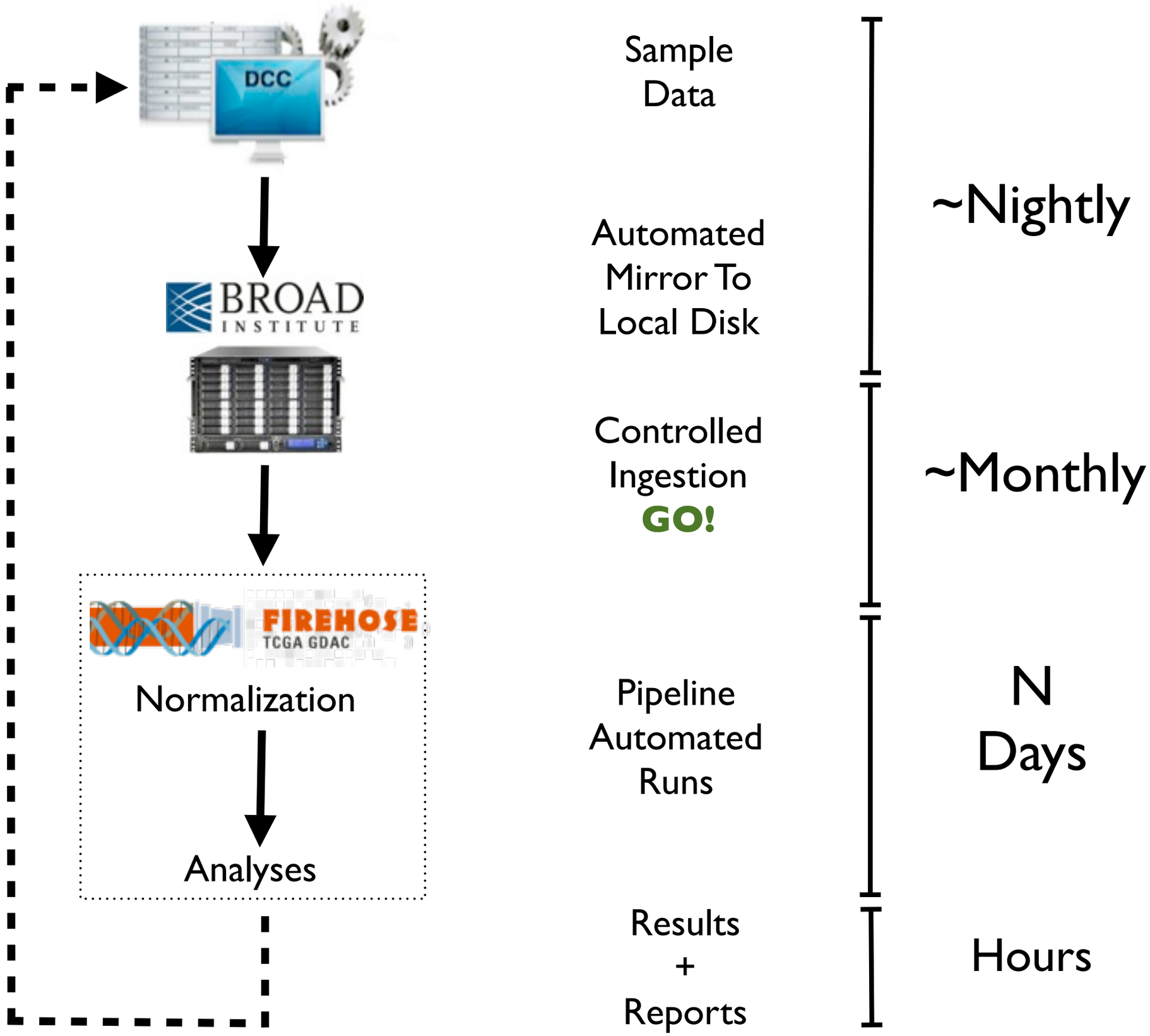
Fine Print

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.

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



But While Doing This We Constantly See

The Babel Problem

Rarely Is There Agreement On Central Question:

How Much Data Do We Have?

New

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

JAN 2012 UPDATE: OUR STDDATA PKGS FED TO ICGC, TOO

New

ROLE 3: TARGETED AWG RUNS

- Analysis Targets Of Oppo

  e.g. for coordinatec

- Example: 2 runs perform
  - Standard analyses r
  - TOO for May 2 LUN

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	***ov has size***	***1.6T***
gbm has size	***1.8T***	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 ...

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

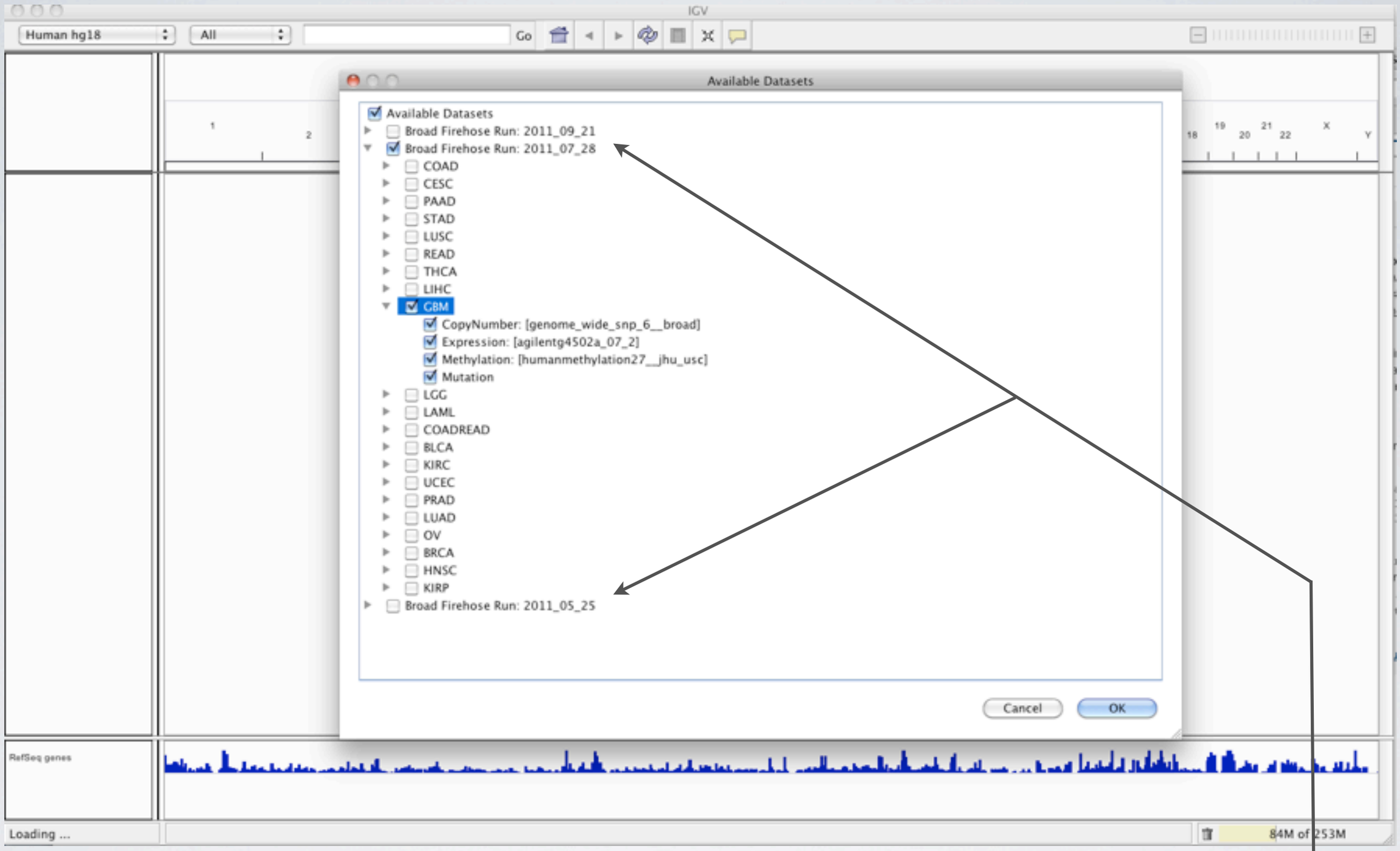
Index of /tcgafiles/ftp_auth/distro_ftpusers/tcga4yeo/other/gdacs/gdacbroad

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	-
hnsc/	08-Oct-2011 11:03	-
kirc/	08-Oct-2011 10:57	-
kirp/	08-Oct-2011 10:58	-
laml/	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	-

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



Directly from Broad, no TCGA credentials required

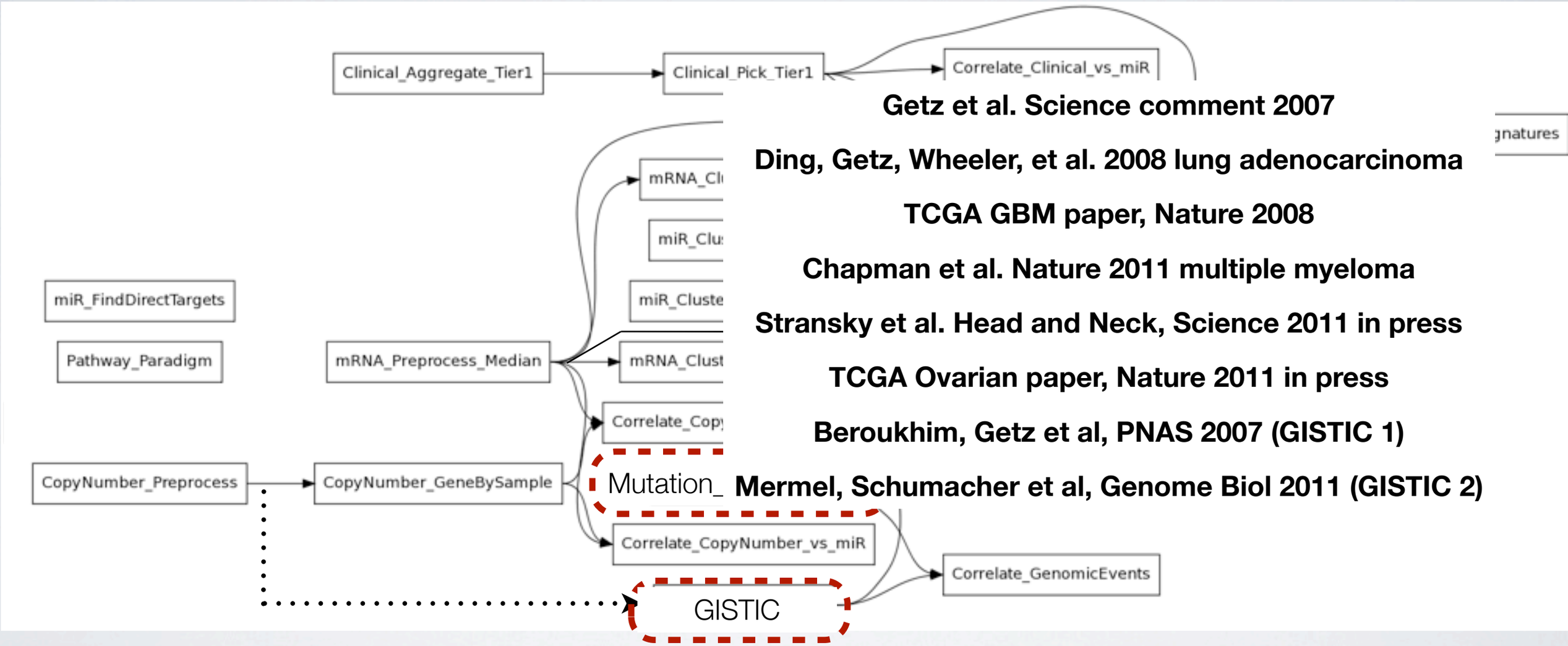
https://confluence.broadinstitute.org/display/GDAC/IGV+Data+Loading

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

IV : Insights & Challenges

Insight 1:

This ... 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.
504 pipes and ~1000 GenePattern modules, per run
<u>Continuously evolving</u> through years of publication use.

A Tale of Two Coders

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

Like ENI

simple task

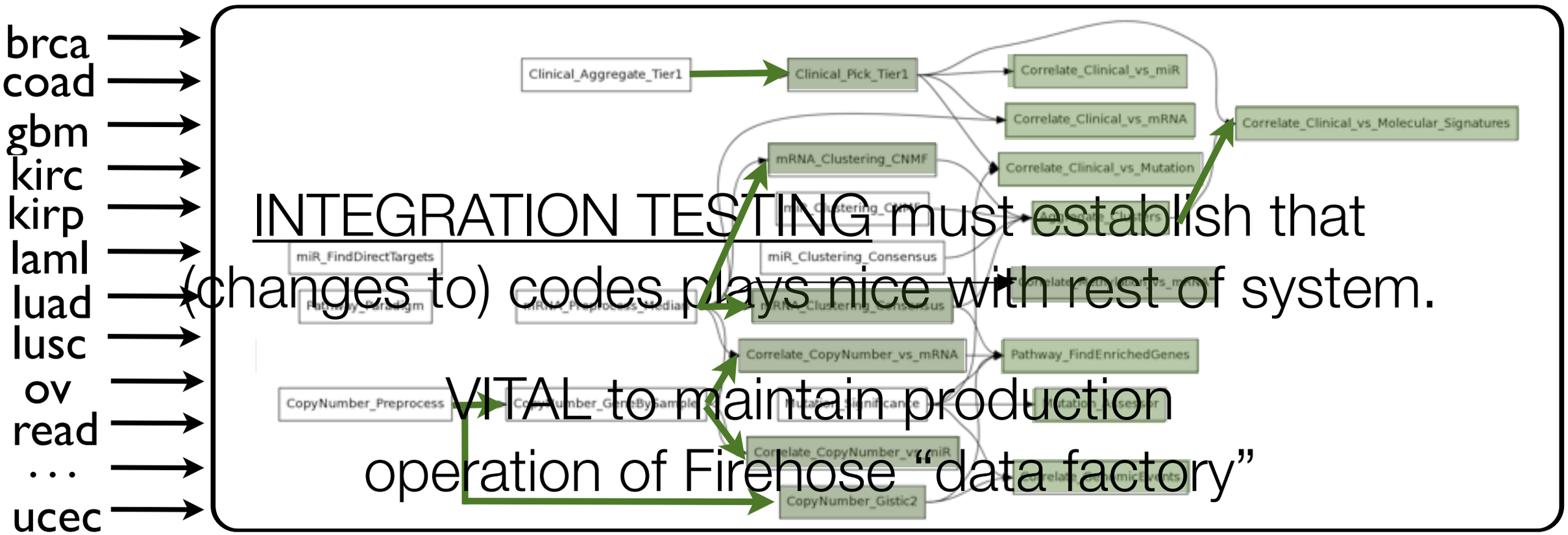
... in par

.

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.



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

VITAL to maintain production operation of Firehose "data factory"

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

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**

- 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

Insight 4:

Given that TCGA arguably largest/richest cancer data ever assembled

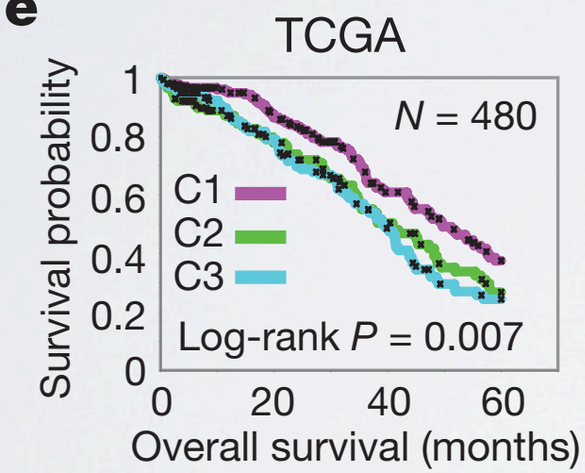
d

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 OV miR expression yielded 3 subtypes

Discoveries lurk in our GDAC pipeline outputs

e



One of which correlated to significantly longer survivability

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

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

For More Information

Dashboard › Broad TCGA GDAC › Home Browse Michael Noble Search
› FAQ

FIREHOSE Broad GDAC **FAQ** Edit Add Tools

2 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 /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	-
fpli/	08-Oct-2011 10:56	-
gbm/	08-Oct-2011 10:56	-
hnsc/	08-Oct-2011 11:03	-
kirc/	08-Oct-2011 10:57	-
kirp/	08-Oct-2011 10:58	-
laml/	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	-

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 http://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	100%
GBM	24	100%
READ	17	71%
LUSC	17	71%
LUAD	17	71%
COAD	17	71%
COADREAD	17	71%
BRCA	12	50%
KIRC	10	42%
KIRP	7	29%
UCEC	4	17%
LGG	4	17%
CESC	4	17%
BLCA	4	17%
STAD	3	13%
LIHC	3	13%
HNSC	3	13%
THCA	2	8%
PRAD	2	8%
LAML	2	8%

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 gdac.broadinstitute.org
Email gdac@broadinstitute.org