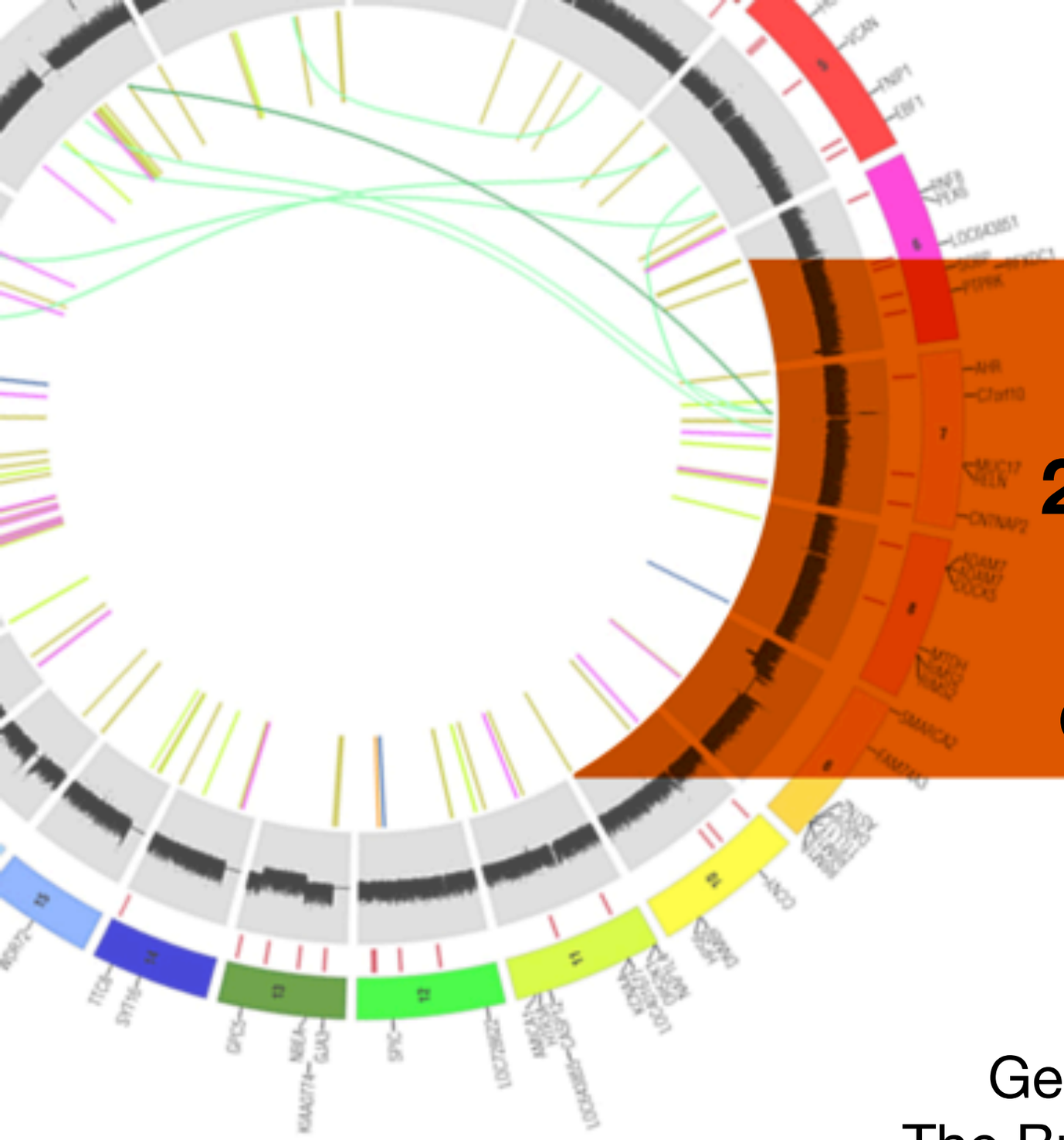


The Cancer Genome Atlas

Firehose Workshop
2nd TCGA Symposium
November 27, 2012

Crystal City, Virginia, U.S.A.

Michael S. Noble
Genome Data Analysis Center
The Broad Institute of MIT & Harvard

BROAD INSTITUTE

FIREHOSE
Broad GDAC

Acknowledgements

PI: Lynda Chin, Gaddy Getz

Broad Institute
Michael S. Noble
Douglas Voet
Daniel DiCara
Gordon Saksena
Hailei Zhang
David Heiman
Juok Cho
William Mallard
Michael Lawrence
Petar Stojanov
Lihua Zou
Chip Stewart
Scott Frazer
Pei Lin
Kristian Cibulskis
Rui Jing
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
Scott Carter
Juliann Chmielecki
Andrew Cherniack
Rameen Beroukhim

Peter Park
Nils Gehlenborg
Semin Lee
Richard Park





OUTLINE

I. Refresher: TCGA Data Flow

II. Refresher: Why Firehose?

III. What Firehose produces

IV. How To Get It

V. Summary & Future

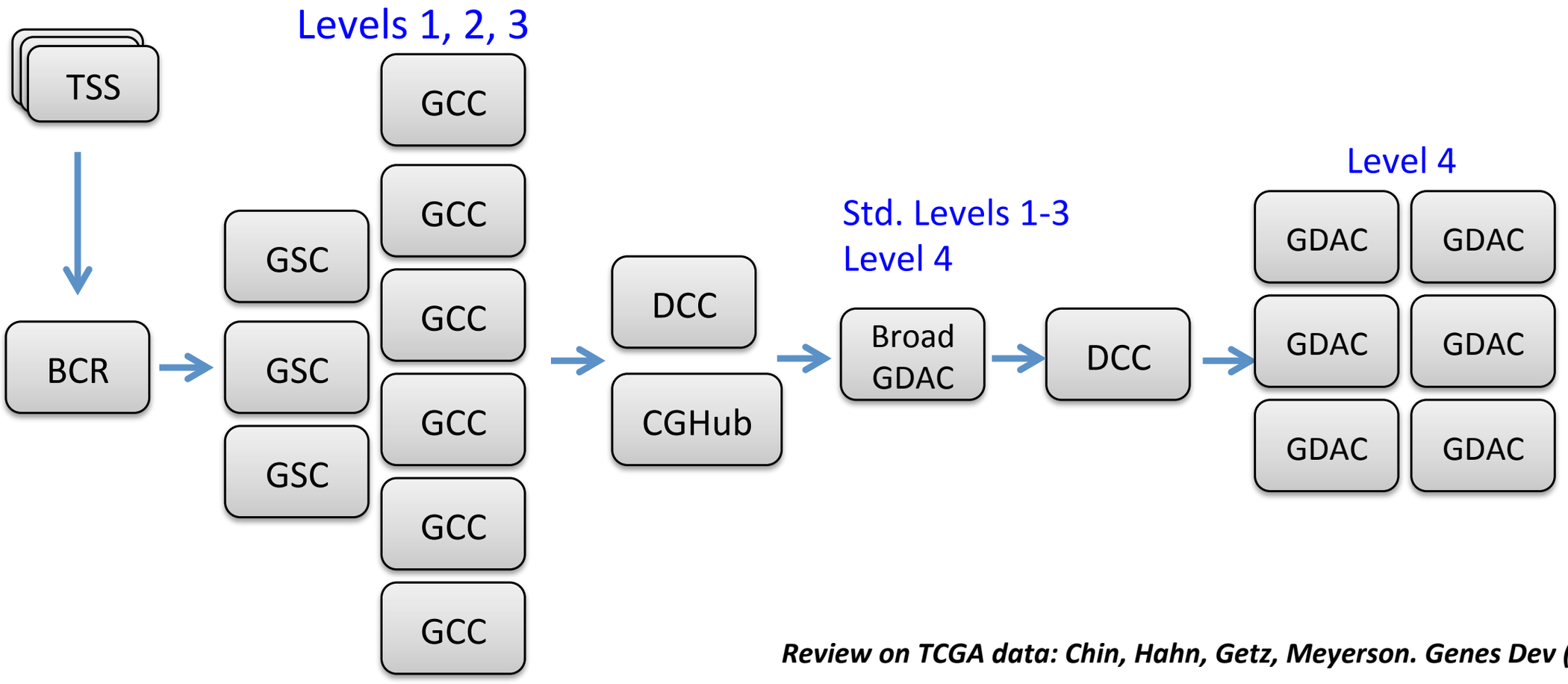
I. TCGA November 2012: The flood Continues

- 7K patient cases, heading to 11K total
- 26 tumor cohorts (plus clinical)
- 6 marker papers published, more underway
- Swirling amongst 20 centers nationwide (and ICGC)

Understanding TCGA : data flow & levels

~150 tissue sites

20 research centers



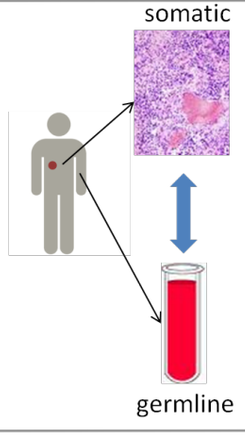
Review on TCGA data: Chin, Hahn, Getz, Meyerson. Genes Dev (2011)

Purpose I: Characterization:

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

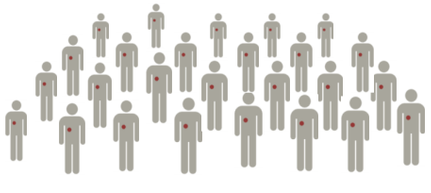
Level 2 – Normalized data (e.g. aligned reads – BAM files, intensity matched files)

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

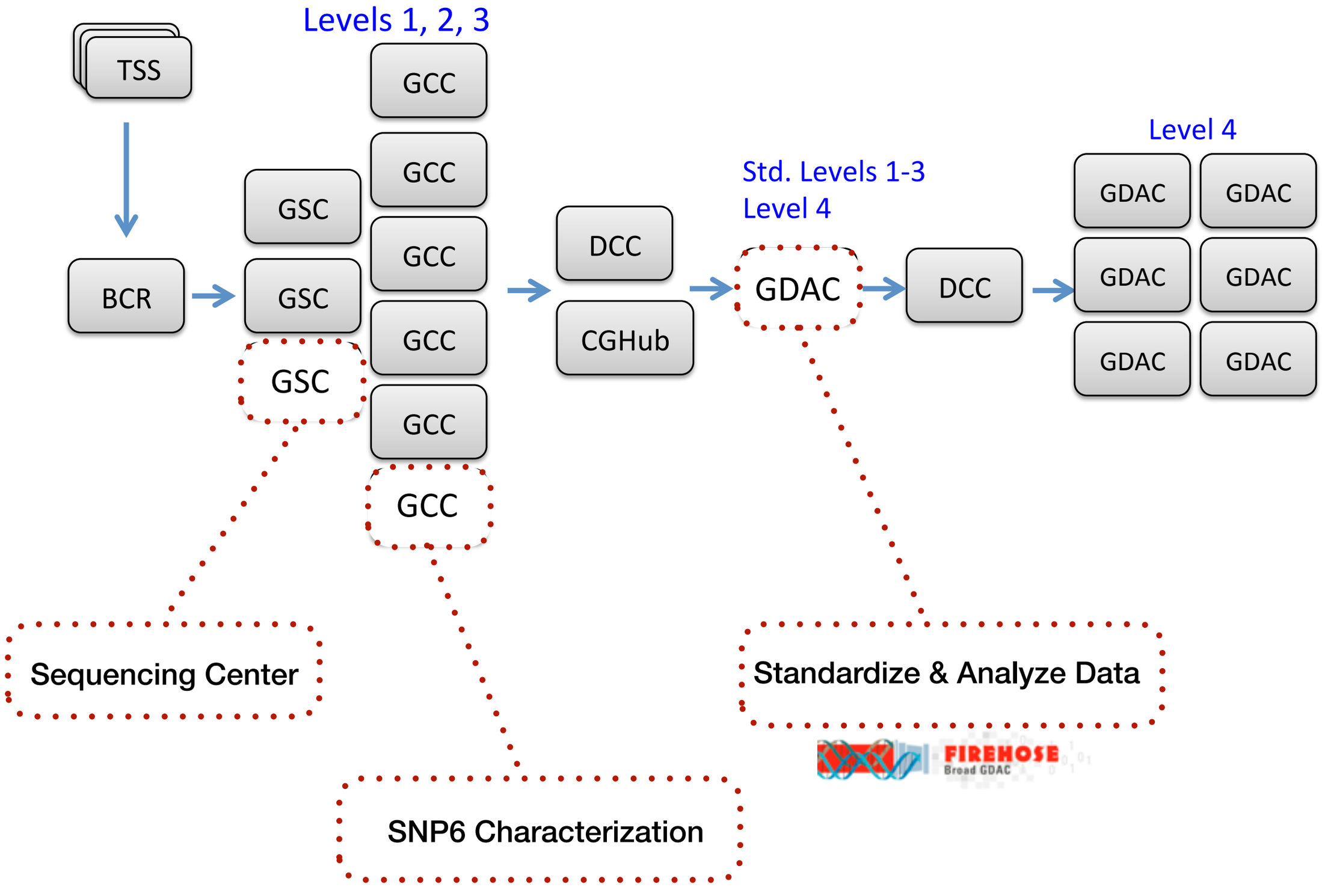


Purpose 2: Interpretation:

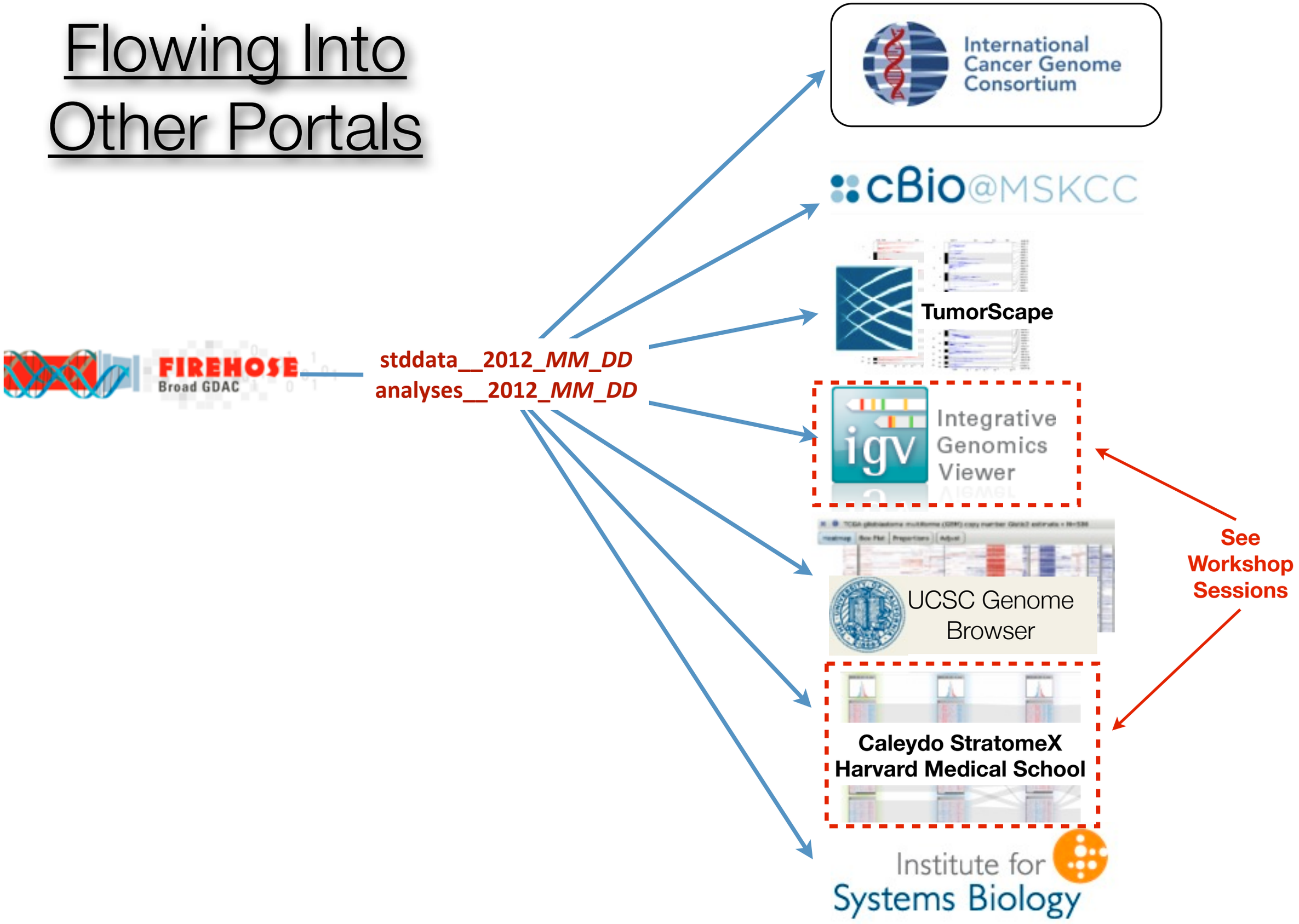
Level 4 – Analysis across a cohort (e.g. sub-types discovery, correlate data types, significantly mutated genes/regions/pathways, correlation to clinical parameters)



Broad Contribution: 3 of 20 research centers



Flowing Into Other Portals



But How Is Data Stream Used 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?
  How might those features map to 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 downstream users do not have to repeat each time

  (e.g. automation a boon when 20 new samples added to 250)

- Nor perform ad-hoc reinvention of methods

- Nor download all low-level data from which they were generated

- Nor 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

Firehose aims to address such concerns

II. WHY FIREHOSE?

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 ~35 terabytes of TCGA data and reliably executes more than 2300 pipelines per month.

Because The Bad Old Days ...

Of solitary, manual experimentation on few dozen samples ...

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

Then do it again Nov 13, 17, ...
Then forget, search ... lose track, search ...
Then repeat ALL for 20 more tumors
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						0					0
KICH						0					0
KIRC						72					403
KIRP						16					0
LAML						0					199
LGG						27					0
LIHC						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						454	412	316
PAAD	57	0	48						34	0	0
PANCAN8	4086	3882	3907						3169	2282	2152
PRAD	180	127	171						170	0	83
READ	169	168	162						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 TCGA Symposium (~11K samples)

New data types since Nov 2011 (12.5K samples)

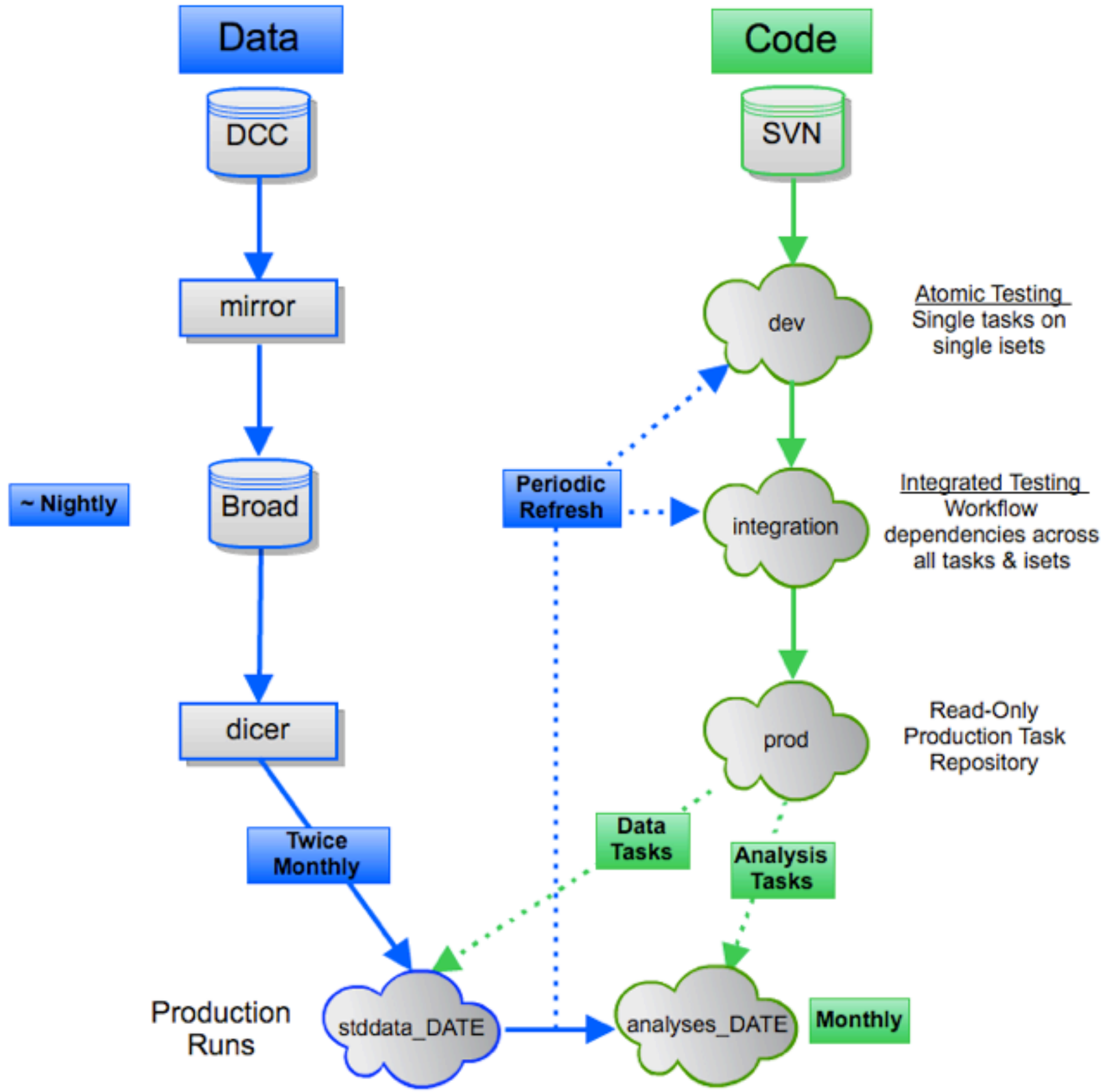
1 year: ~24K new samples

Acute Need for Automation, Systematic Rigor, and Transparency

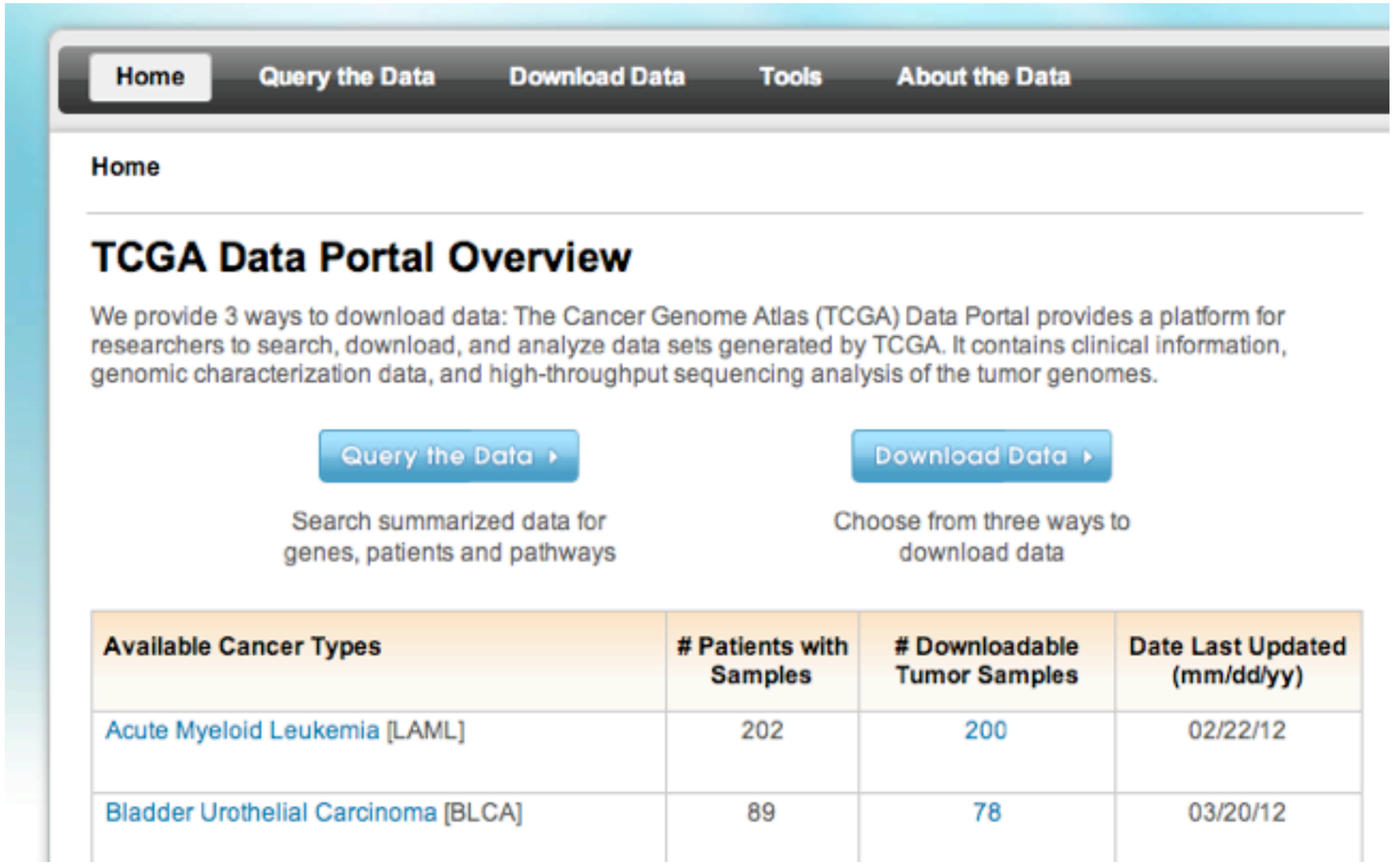
GDAC Firehose == Virtual Data Factory

Subject to Same Engineering Constraints of Timeliness, Transparency & Rigor as Physical Factories

Not academic one-off



But is this necessary ...



... given TCGA/DCC data portal already exists?

DCC portal: great resource, but more “raw” ...

No data aggregate / versioning

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

... and does not encompass analytics

What if I just want to view OV Gistic (CN) peaks?

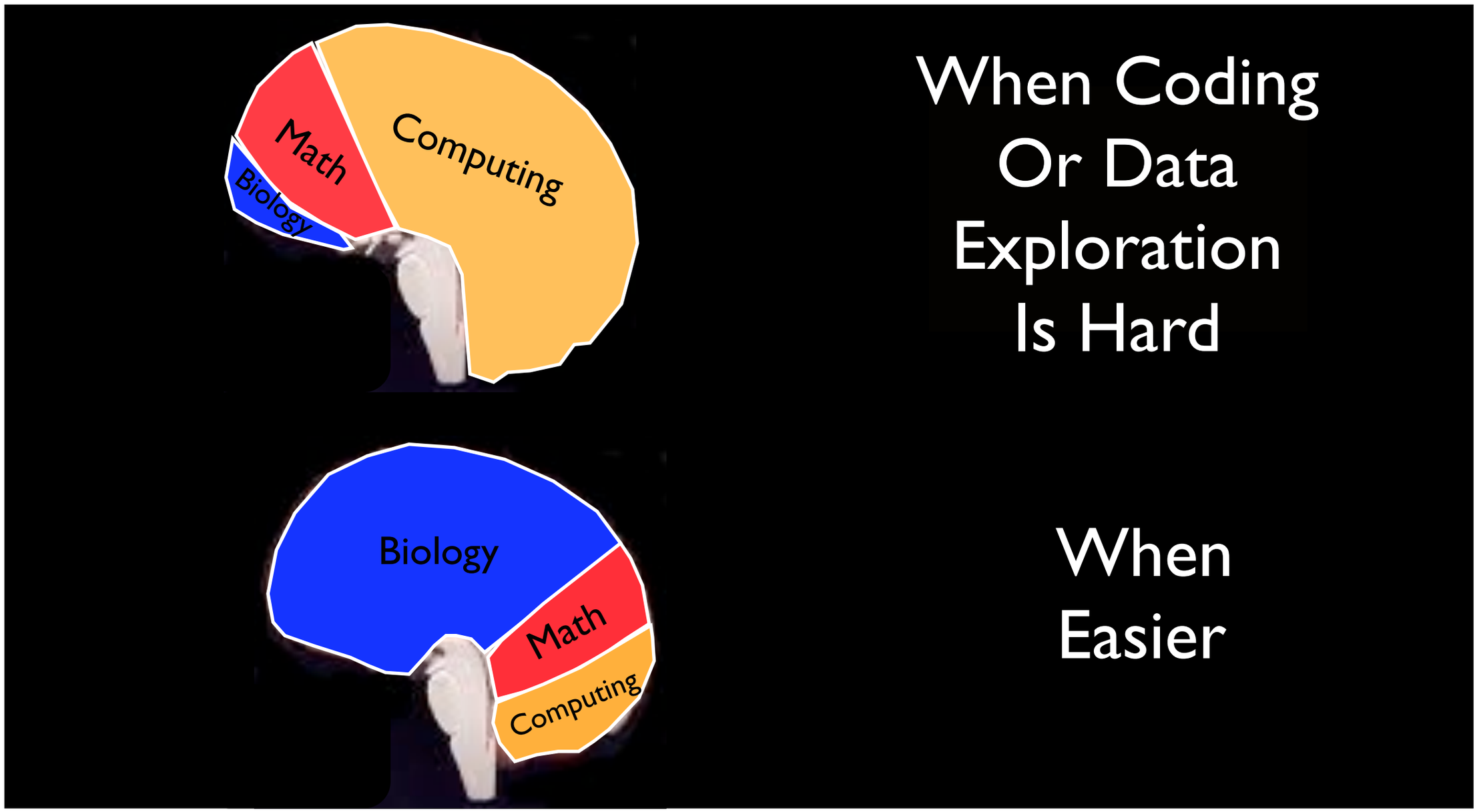
Or peek at an expression or methylation cluster?

Must I become an expert first?

Complexity, volume, and limited time preclude this level of involvement for many

Especially those without dedicated bioinformatics staff (e.g. MD or wet lab PH.D.)

It Must Be Faster/Easier/Simpler, Because ...



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

III. So Firehose Data Factory Produces

1. Version-stamped, standardized datasets (2X / month)

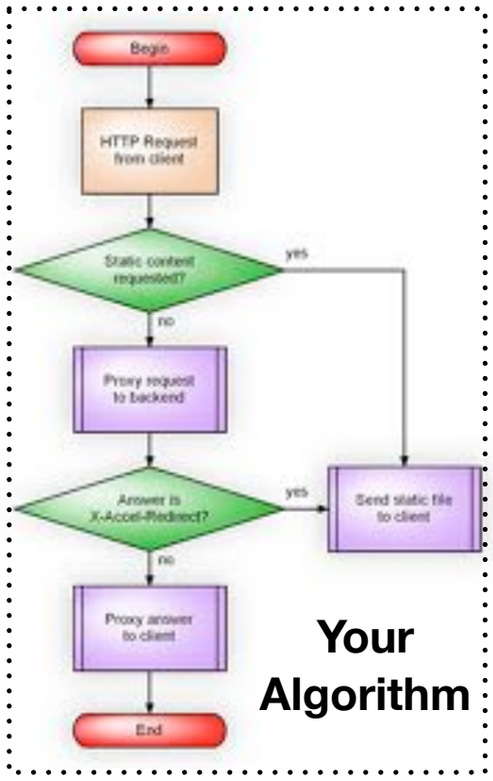
   Precursor to automated analyses, durable (DCC) & citable

2. Regular package of standard analyses results (1X / month)

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

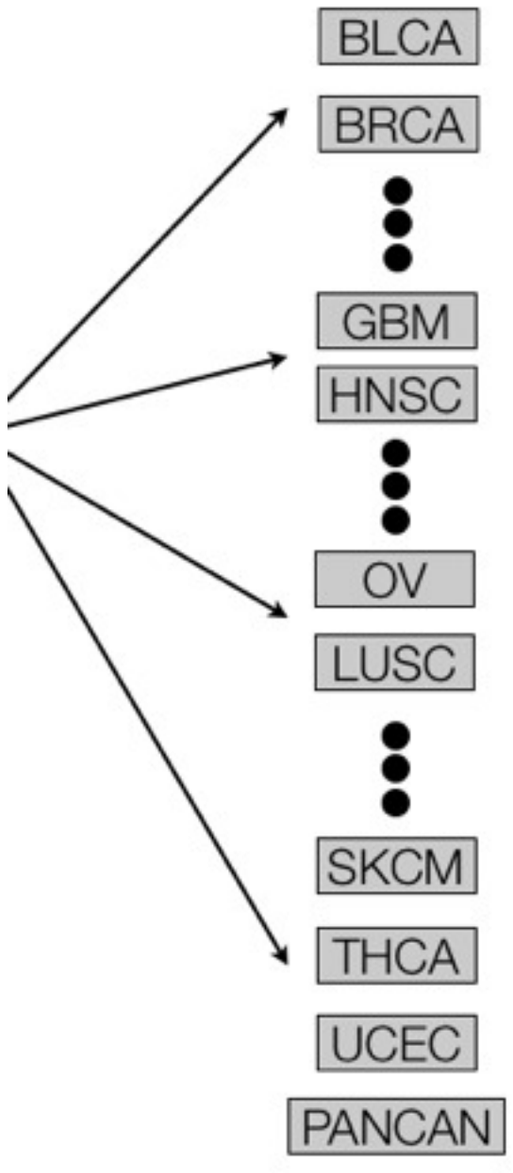
3. Companioned with biologist-friendly reports (475 / month)

For Tremendous Benefits of Scale & Richness



Put a new code in ...

FIREHOSE
Broad GDAC



... access thousands of samples + clinical
... across 9 datatypes and 25 tumor sets
... with press of button
... hardening your code
... in arguably richest cancer-data laboratory in world ... TCGA!
... cross-coupled with other important algorithms

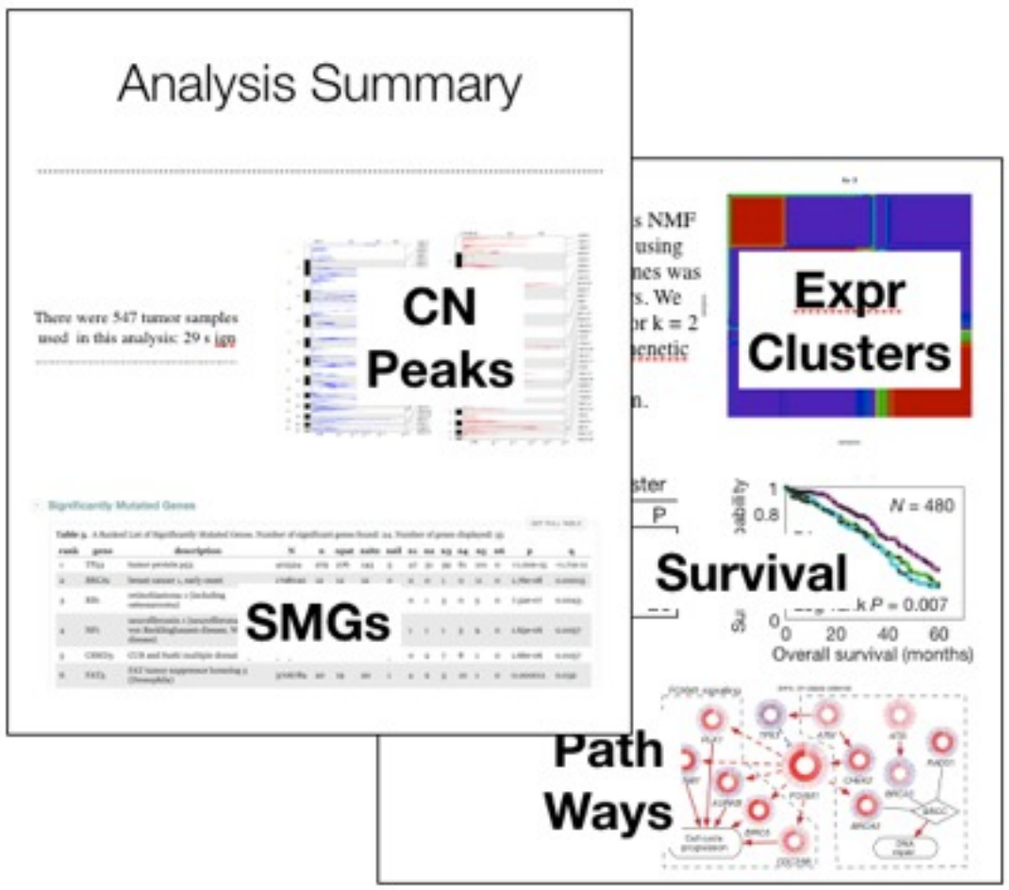
Unprecedented Scale: KiloPipeline per Month

stddata__2012_10_04	**547 datasets submitted to DCC**
stddata__2012_10_24	**940 datasets ...**
analyses__2012_10_24	**837 analyses ...**

2324 pipelines across 26 disease cohorts

With up to 33
biologist-friendly
analysis reports
per disease
(475 reports in all)

Single Month: Oct 2012



Again: Not Mutually Exclusive With ***"Easy"***

```
%  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 diffs in 35+ terabytes

Version stamp: rigor & clarity ⟶ ease

Easy Corroboration: first-pass, low hanging fruit

- 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 2012_07_25 results, too?"

Durability of DCC archive fosters citable referencing:

"Our analyses were performed against TCGA dataset version 2012_07_25, and validated against ...

But Mind The 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

Automatic Machines Are Dumb & Imperfect
Expert Judgement Still Required

`firehose2nature` tool is organic, not in-silico

Actively Used by Multiple TCGA AWGs

2012_08_25 awg_pancan8 Analyses Run

Tables of Ingested Data: HTML PNG TSV Redactions: Report

AnalysisReport	# Pipelines	% Successful
BRCA	35	100%
COADREAD	35	100%
GBM	34	100%
KIRC	35	100%
LUSC	35	100%
OV		
UCEC		
PANCAN		

Analysis Overview for Thyroid Adenocarcinoma: 2012_10_24

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

Unique Tumor Sample Counts

Tumor	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
THCA	435	218	330	94	353	0	254	0	349	224	323

...ose_get version 0.3.8

...eference point, enabling a wide range of cancer biologists, ...asily incorporate TCGA into the backdrop of ongoing ...ta and algorithms are of the highest possible quality, these

Analysis Overview for Lung Adenocarcinoma: 2012_11_15

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

Unique Tumor Sample Counts

Tumor	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
LUAD	439	294	358	0	432	32	355	0	366	237	229

Download run results with firehose_get version 0.3.8

Download command: ***firehose_get awg_luad 2012_11_15***

Task Dashboard

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.

UCEC

SKCM

LUSC

Future Potential: Automated Clinical Gold Mine?

Wealth of clinical data collected by TCGA

To date underrepresented in TCGA-based publications

Understandable byproduct of complex mix of scientific, technological & operational factors

But clear steps can be taken to minimize extent that

Sheer volume & complexity ... **alone** ...

Impede fuller exploitation of clinical in TCGA-based work

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

Precedent in 2011 Ovarian Manuscript

d

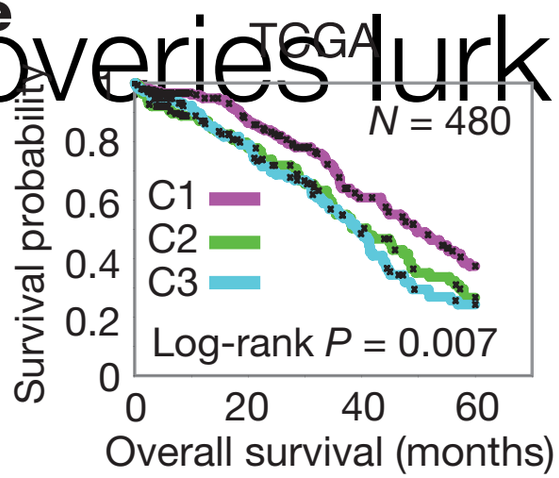
miRNA cluster	Gene cluster D	I	M	P
C1	55	48	15	89
C2	40	21	51	29
C3	39	37	43	20

CNMF clustering of OV miR expression yielded 3 subtypes

Given richeness of TCGA data stream, Discoveries lurk in our GDAC pipeline outputs

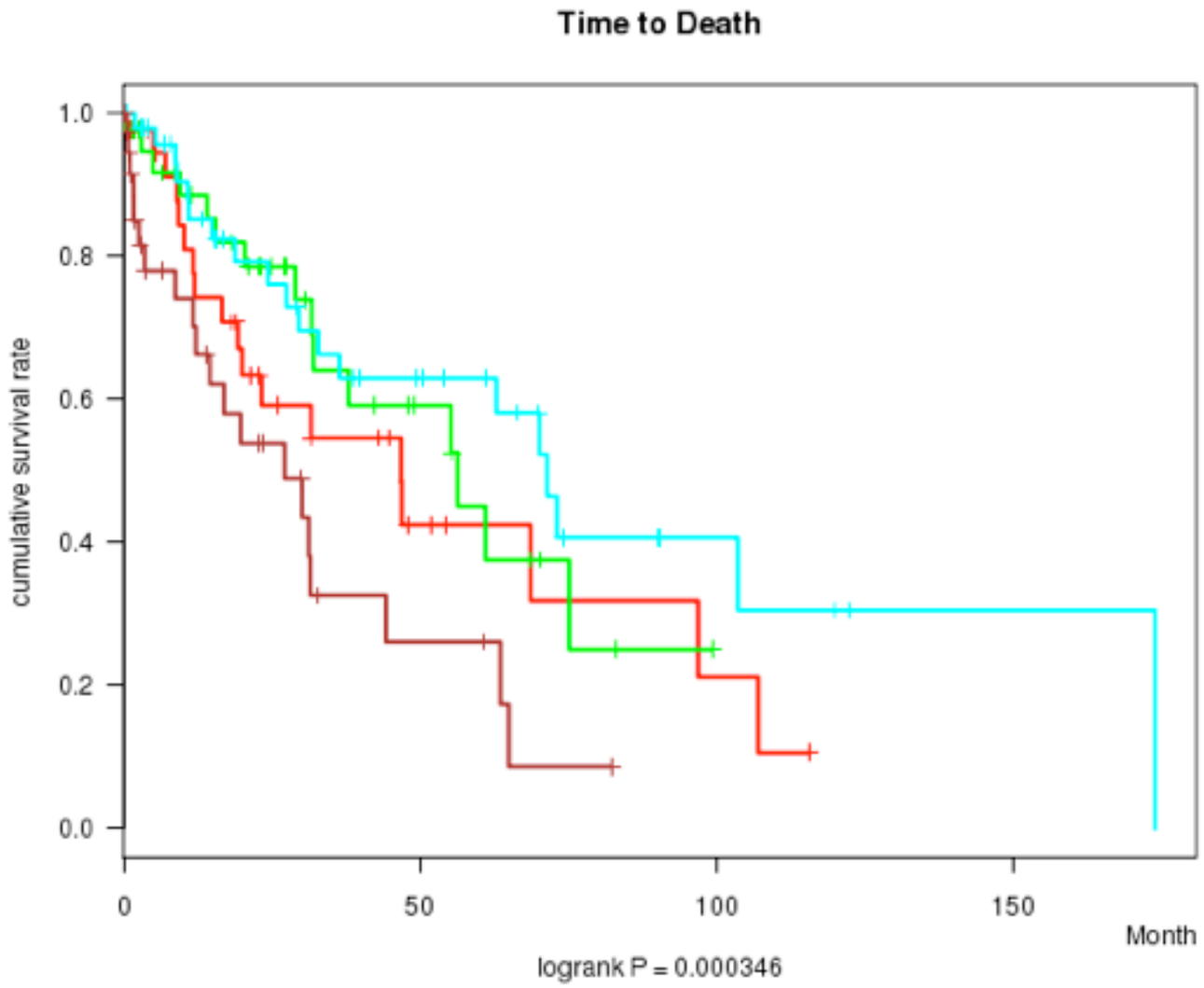
One of which correlated to significantly longer survivability

e



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

LUSC



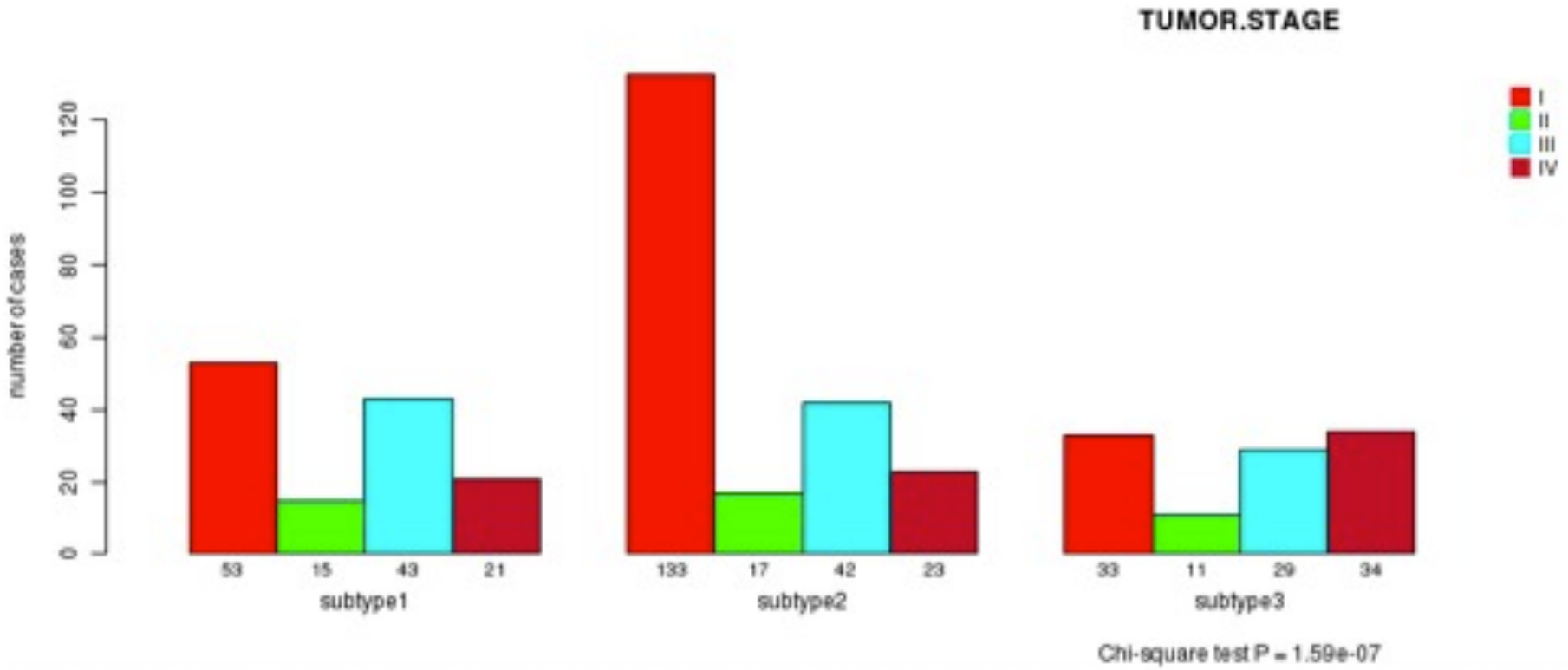
2012_09_13
Analyses

	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)

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

P value = *0.000346* (logrank test)

KIRC



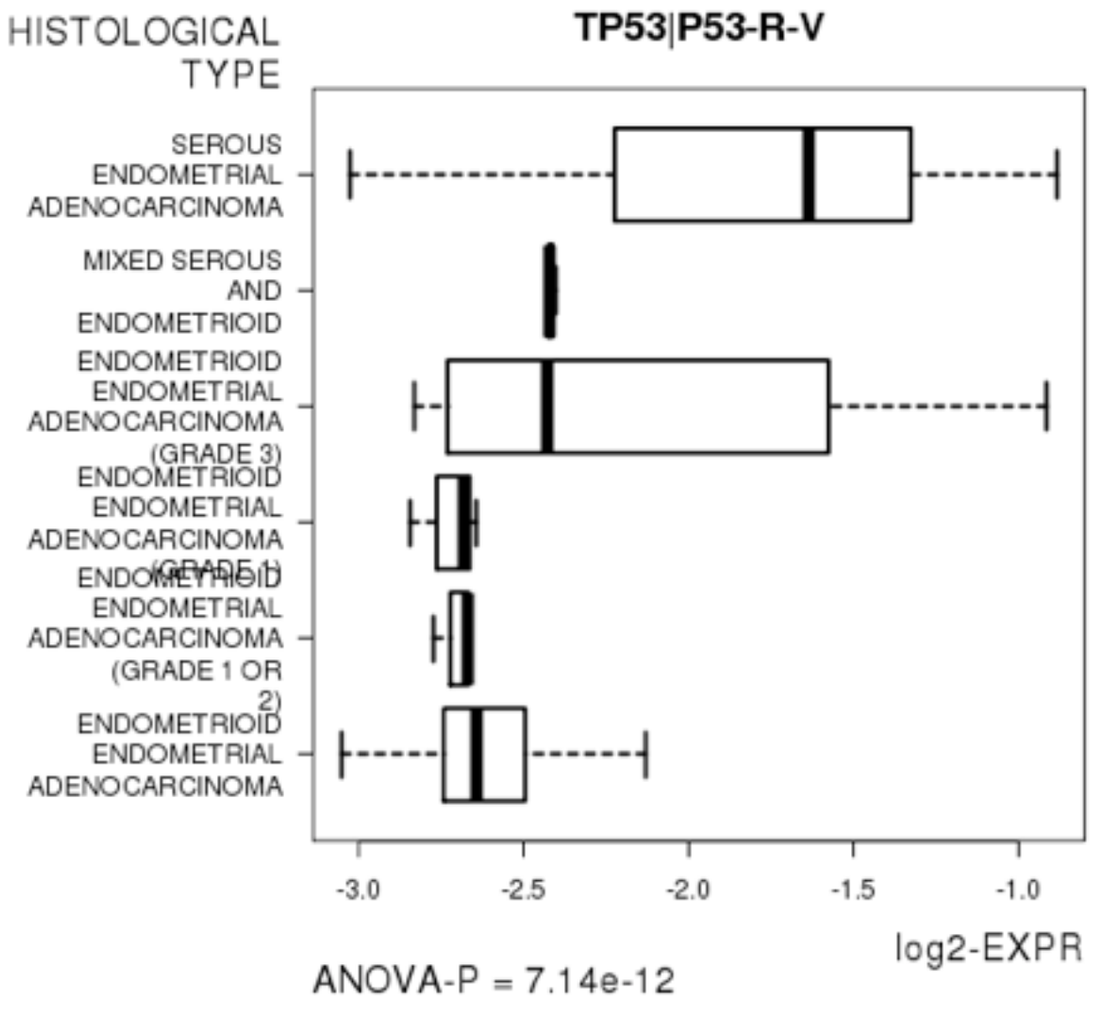
nPatients	I	II	III	IV
ALL	**219**	**43**	**114**	**78**
subtype1	53	15	43	21
subtype2	133	17	42	23
subtype3	33	11	29	34

'RPPA cHierClus subtypes'* versus *'TUMOR.STAGE'

P value = *1.59e-07* (Chi-square test)

UCEC

Correlation between RPPA expression and *'HISTOLOGICAL.TYPE'*



	ANOVA_P	Q
TP53|P53-R-V	7.144e-12	1.19e-09
CHEK2|CHK2_PT68-R-C	7.824e-09	1.29e-06
AKT1_AKT2_AKT3|AKT_PS473-R-V	6.908e-08	1.13e-05
PGR|PR-R-V	1.307e-07	2.13e-05
CDC2|CDK1-R-V	2.576e-07	4.17e-05
CDH1|E-CADHERIN-R-V	2.644e-07	4.26e-05
ESR1|ER-ALPHA-R-V	9.567e-07	0.000153
ESR1|ER-ALPHA_PS118-R-V	3.992e-06	0.000635
EEF2|EEF2-R-V	7.75e-06	0.00122
EIF4EBP1|4E-BP1_PS65-R-V	1.874e-05	0.00294

IV. How To Access

2012_10_24 stddata Run

DiseaseType	# Datasets	% Processed	Download
BLCA	38	100%	Open Protected
BRCA	48	100%	Open Protected
CESC	20	100%	Open Protected
COADREAD	38	100%	Open Protected
COAD	38	100%	Open Protected
DLBC	13	100%	Open Protected
GBM	51	100%	Open Protected
HNSC	40	100%	Open Protected
KICH			Open Protected
KIRC			Open Protected
KIRP			Open Protected
LAML			Open Protected
LGG			Open Protected
LIHC			Open Protected
LUAD			Open Protected
LUSC			Open Protected
OV	57	100%	Open Protected
PAAD	14	100%	Open Protected
PRAD	30	100%	Open Protected
READ	38	100%	Open Protected
SARC	13	100%	Open Protected
SKCM	24	100%	Open Protected
STAD	27	100%	Open Protected
THCA	40	100%	Open Protected
UCEC	48	100%	Open Protected
PANCAN8	87	95%	Open Protected

Data Dashboard

2012_10_24 analyses Run

AnalysisReport	# Pipelines	% Successful	Download
BLCA	36	100%	Open Protected
BRCA	44	100%	Open Protected
CESC	26	100%	Open Protected
COADREAD	44	100%	Open Protected
COAD	44	100%	Open Protected
GBM	46	100%	Open Protected
HNSC	18	100%	Open Protected
KIRC	44	100%	Open Protected
KIRP			Open Protected
LAML			Open Protected
LGG			Open Protected
LIHC			Open Protected
LUAD			Open Protected
LUSC			Open Protected
OV			Open Protected
PAAD	11	100%	Open Protected
PRAD	33	100%	Open Protected
READ	44	100%	Open Protected
SARC	7	100%	Open Protected
SKCM	25	100%	Open Protected
STAD	31	100%	Open Protected
THCA	37	100%	Open Protected
UCEC	44	100%	Open Protected
DLBC	7	88%	Open Protected
KICH	6	75%	Open Protected
PANCAN8	9	56%	Open Protected

Analysis Dashboard

http://gdac.broadinstitute.org

Open Public Resource Interactive Desktop Use

Nexus Resource for Evolving Community

- Thousands of views, 140K+ hits / month
- Hundreds of GB downloads / month
- Across dozens of centers & portals
- Research / Academic / Commercial
- National & International
- Beyond genomics : e.g. CPTAC / proteomics

Continuing to Gain Traction

(more useful than pretty, but facelift coming in 2013)

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

Browse ▾ Log In Search

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>

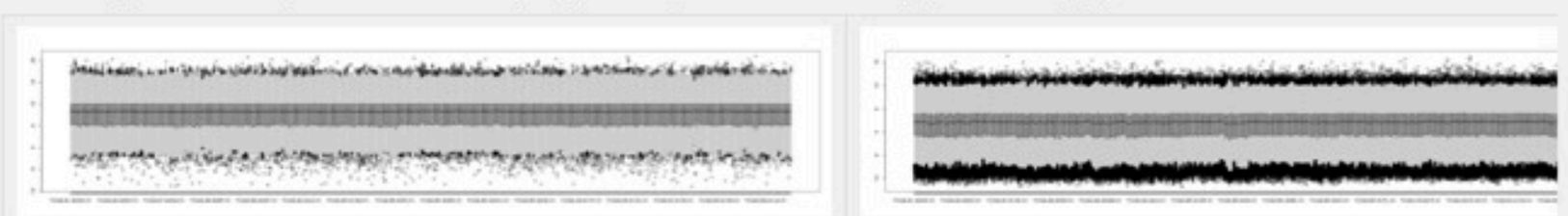
>
> Dear David,
>
> Apologies for the delay in responding. Yes, you are right: our outputs do
> not
> contain normals. This is partly a legacy held over from the TCGA pilot
> studies, which is where many of the analyses in our GDAC originally stem
> from. Our FAQ online at gdac.broadinstitute.org discusses 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).



stddata dashboard

The Broad GDAC standardized data packages represent a frozen snapshot of all TCGA analysis data at a given time:

- ***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***

2012_08_04 stddata Run

Tables of Ingested Data: HTML PNG TSV Redactions: Report

ReleaseNotes	# Datasets	% Processed	Download	
BLCA	20	100%	Open	Protected
BRCA	27	100%	Open	Protected
CESC	11	100%	Open	Protected
COADREAD	21	100%	Open	Protected
DLBC	5	100%	Open	Protected
GBM	27	100%	Open	Protected
HNSC	20	100%	Open	Protected
KIRC	27	100%	Open	Protected
KIRP	23	100%	Open	Protected
LAML	11	100%	Open	Protected
LGG	17	100%	Open	Protected
LIHC	17	100%	Open	Protected
LUAD	26	100%	Open	Protected
LUSC	34	100%	Open	Protected
OV	32	100%	Open	Protected
PAAD	6	100%	Open	Protected
PRAD	16	100%	Open	Protected
SKCM	14	100%	Open	Protected
STAD	18	100%	Open	Protected
THCA	18	100%	Open	Protected
UCEC	22	100%	Open	Protected
PANCANCER	48	87%	Open	Protected

Data/Provenance Rigor

Towards solving
BABEL Problem

Launch Point For
Analysis-Ready
TCGA Data

→ ICGC, too!

UP < > EXPAND ALL COLLAPSE ALL SET AUTO WIDTH PRINT

stddata__2012_11_14 Samples Summary Report

Overview

Introduction

For TCGA data, redaction is the removal of cases from the data prior to publication or release. Redacted cases are generally rare, but cases must be redacted when the TSS/BCR subject link is incorrect ("unknown patient identity"), or in the case of genotype mismatch, completely wrong cancer, or completely wrong organ/tissue. Redaction occurs regardless of a case's analyte characterization or DCC data deposition status.

Rescission is the removal of samples from the list of redactions. This happens if the reason for redaction is eventually cleared up. For clarity, rescinded redactions do not appear in this report.

Summary

There were 60 redactions.

Results

Redactions

Table 1.

Barcode	UUID	Date	Type	Notes
TCGA-BR-4190	282e979d-4ad9-4d42-8ffa-7487a94fa1f3	11/08/2012	STAD	Site found that there was duplicate tissue in their biobank with another ID and different clinical data than that sent to TCGA. Case is being redacted but may be salvaged if the site can reconsile the correct clinical data to the tissue.
TCGA-BR-4194	2c650fe1-48b0-4f88-bc11-04096be48571	11/08/2012	STAD	Site found that there was duplicate tissue in their biobank with another ID and different clinical data than that sent to TCGA. Case is being redacted but may be salvaged if the site can reconsile the correct clinical data to the tissue.
TCGA-BR-4195	7917234c-63be-4320-b7af-535381f99d99	11/08/2012	STAD	Site found that there was duplicate tissue in their biobank with another ID and different clinical data than that sent to TCGA. Case is being redacted but may be salvaged if the site

Rigor, Transparency, Ease

Comprehensive report on ingested samples

From online dashboard

Nov 8 STAD redactions

+ Redactions

+ Blacklisted Samples

- Filtered Samples

GET FULL TABLE

Table 3. Click on any filtered samples count to display a table detailing the filtered samples for the associated tumor type.

Tumor Type	Filtered Samples Count
BLCA	40
BRCA	693
CESC	9
COAD	1080
DLBC	18
GBM	930
HNSC	602
KIRC	923
KIRP	209
LIHC	222
LUAD	930
LUSC	726
OV	658
PRAD	548
READ	157
SARC	30
SKCM	35
STAD	274
THCA	242
UCEC	122

Clear disposition of every ingested sample, every run

Downloadable as TSV

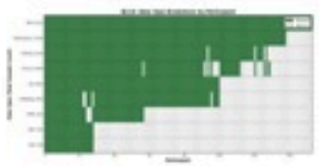
Or view heatmap figure

- Sample Heatmaps

BLCA

GET HIGH-RES IMAGE

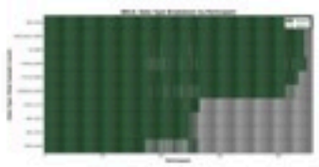
Figure 1. This figure depicts the distribution of available data on a per participant basis.



BRCA

GET HIGH-RES IMAGE

Figure 2. This figure depicts the distribution of available data on a per participant basis.



CESC

analysis dashboard

2012_07_25 analyses Run

Tables of Ingested Data: HTML PNG TSV

AnalysisReport	# Pipelines	% Successful	Download	
BLCA	18	100%	Open	Protected
BRCA	29	100%	Open	Protected
CESC	12	100%	Open	Protected
COADREAD	29	100%	Open	Protected
GBM	28	100%	Open	Protected
HNSC	15	100%	Open	Protected
KIRC	29	100%	Open	Protected
LAML	13	100%	Open	Protected
LGG	22	100%	Open	Protected
LIHC	10	100%	Open	Protected
OV	35	100%	Open	Protected
PRAD	14	100%	Open	Protected
SKCM	12	100%	Open	Protected
THCA	15	100%	Open	Protected
UCEC	29	100%	Open	Protected
KIRP	22	96%	Open	Protected
LUAD	23	96%	Open	Protected
LUSC	20	95%	Open	Protected
STAD	16	94%	Open	Protected
PAAD	4	80%	Open	Protected
PANCANCER	8	41%	Open	Protected

View: Analysis reports Release notes FAQ Download: firehose_get

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									214	276
HNSC	312									0	0
KICH	65									0	0
KIRC	502									454	403
KIRP	135									0	0
LAML	202									0	199
LGG	181									0	0
LIHC	99									0	0
LNNH	2									0	0
LUAD	439									0	229
LUSC	360									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	0	0	0	1	0
13	Correlate_Methylation_vs_mRN	0	0	0	1	0
14	Methylation_Clustering_CNMF	0	0	0	1	0
15	Methylation_Preprocess	0	0	0	1	0
16	miRseq_Clustering_CNMF	0	0	0	1	0
17	miRseq_Clustering_Consensus	0	0	0	1	0
18	miRseq_Preprocess	0	0	0	1	0
19	miR_Clustering_CNMF	0	0	0	1	0
20	miR_Clustering_Consensus	0	0	0	1	0
21	miR_FindDirectTargets	0	0	0	1	0
22	miR_Preprocess	0	0	0	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

Linked to Biologist-Friendly Reports

2012_07_25 analyses Run

Tables of Ingested Data: HTML PNG TSV Redactions: Report

AnalysisReport	# Pipelines	% Successful	Download	
BLCA	18	100%	Open	Protected
BRCA	29	100%	Open	Protected
CESC	12	100%	Open	Protected
COADREAD	29	100%	Open	Protected
GBM	28	100%	Open	Protected
HNSC	15	100%	Open	Protected
KIRC	29	100%	Open	Protected
LAML	13	100%	Open	Protected
LGG	22	100%	Open	Protected
LIHC	16	100%	Open	Protected
OV	35	100%	Open	Protected
PRAD	14	100%	Open	Protected
SKCM	12	100%	Open	Protected
THCA	15	100%	Open	Protected
UCEC	29	100%	Open	Protected
KIRP	22	96%	Open	Protected
LUAD	23	96%	Open	Protected
LUSC	20	95%	Open	Protected
STAD	16	94%	Open	Protected
PAAD	4	80%	Open	Protected
PANCANCER	8	41%	Open	Protected

View: Analysis reports Release notes FAQ Download: firehose_get

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.

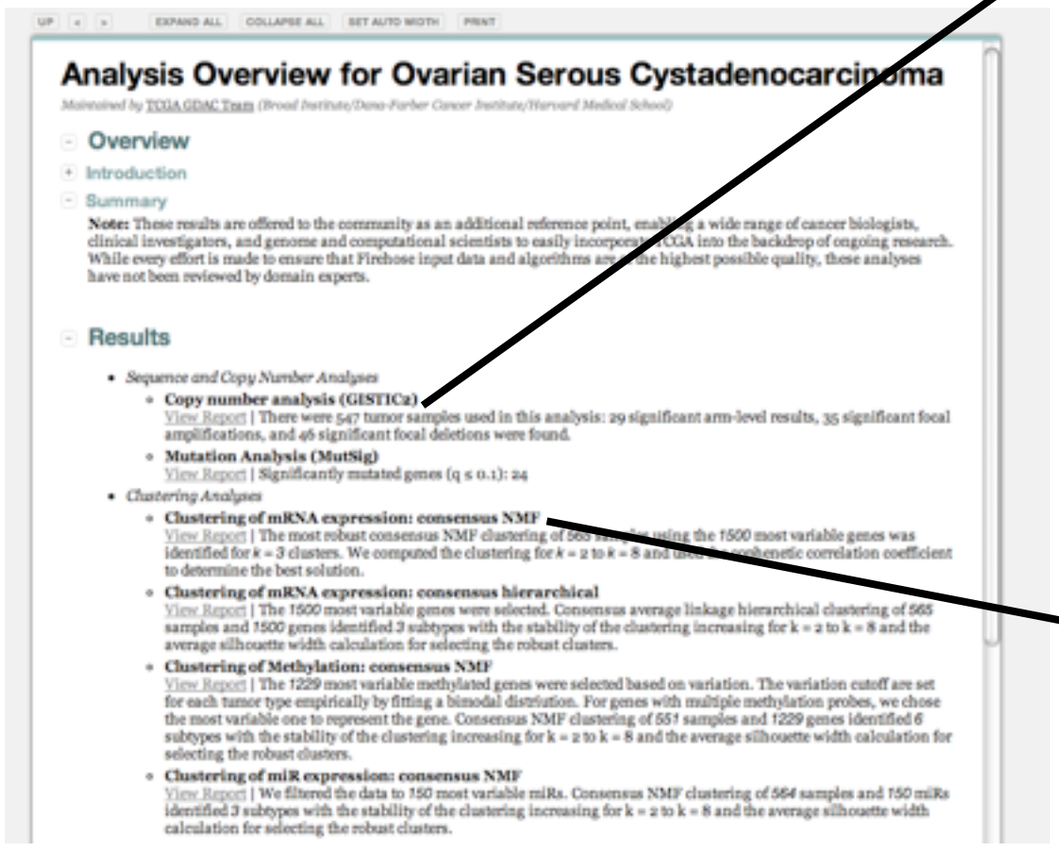
Results

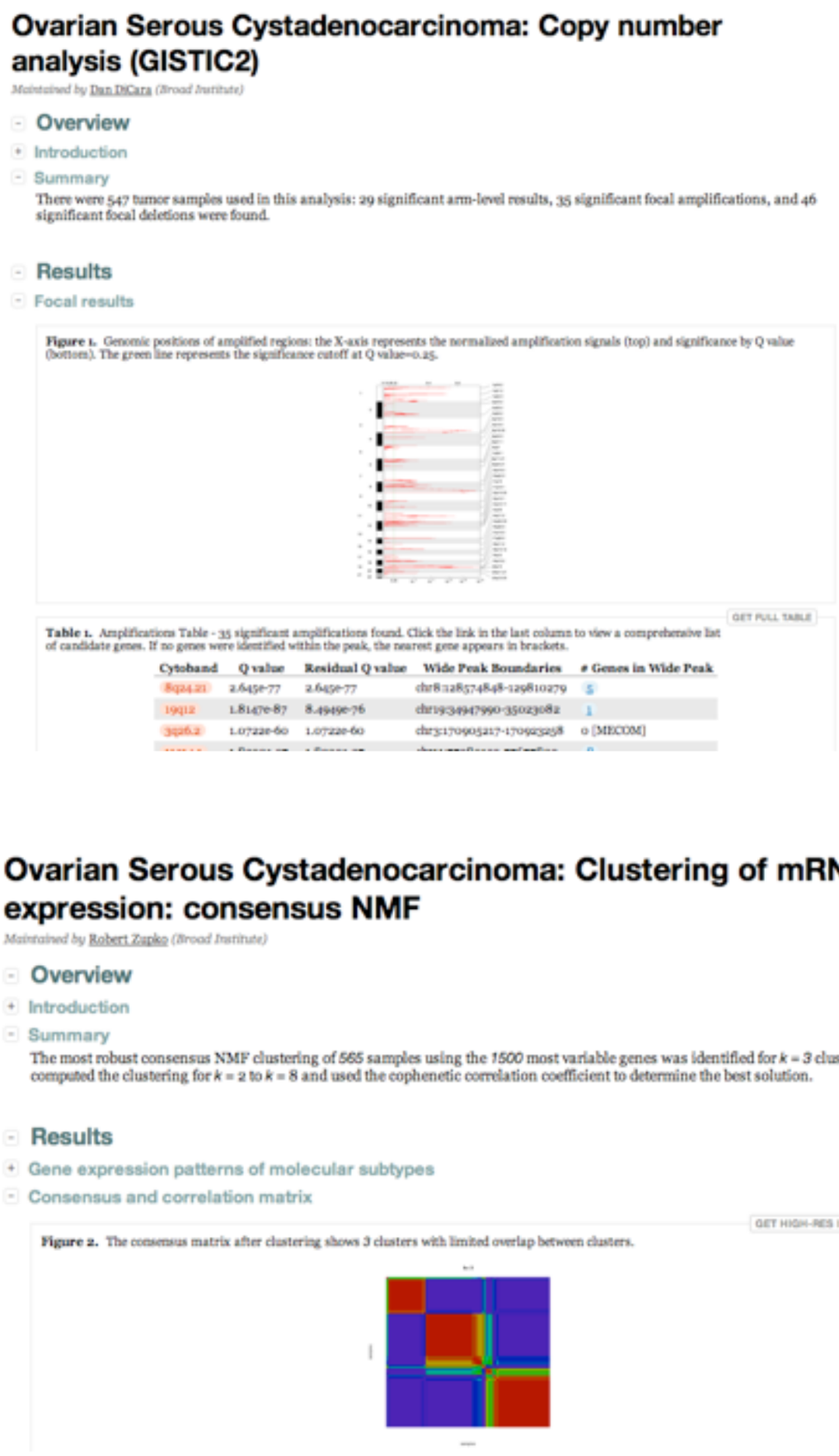
- *Sequence and Copy Number Analyses*
  - **Copy number analysis (GISTIC2)**
    View Report | There were 547 tumor samples used in this analysis: 29 significant arm-level results, 35 significant focal amplifications, and 46 significant focal deletions were found.
  - **Mutation Analysis (MutSig)**
    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





Completely Open: no passwords
Linked to downloadable data

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+.

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 figures to enlarge. Click again to scale down.

Red markers indicate statistically significant results in this section.

Red boxes indicate statistically significant results.

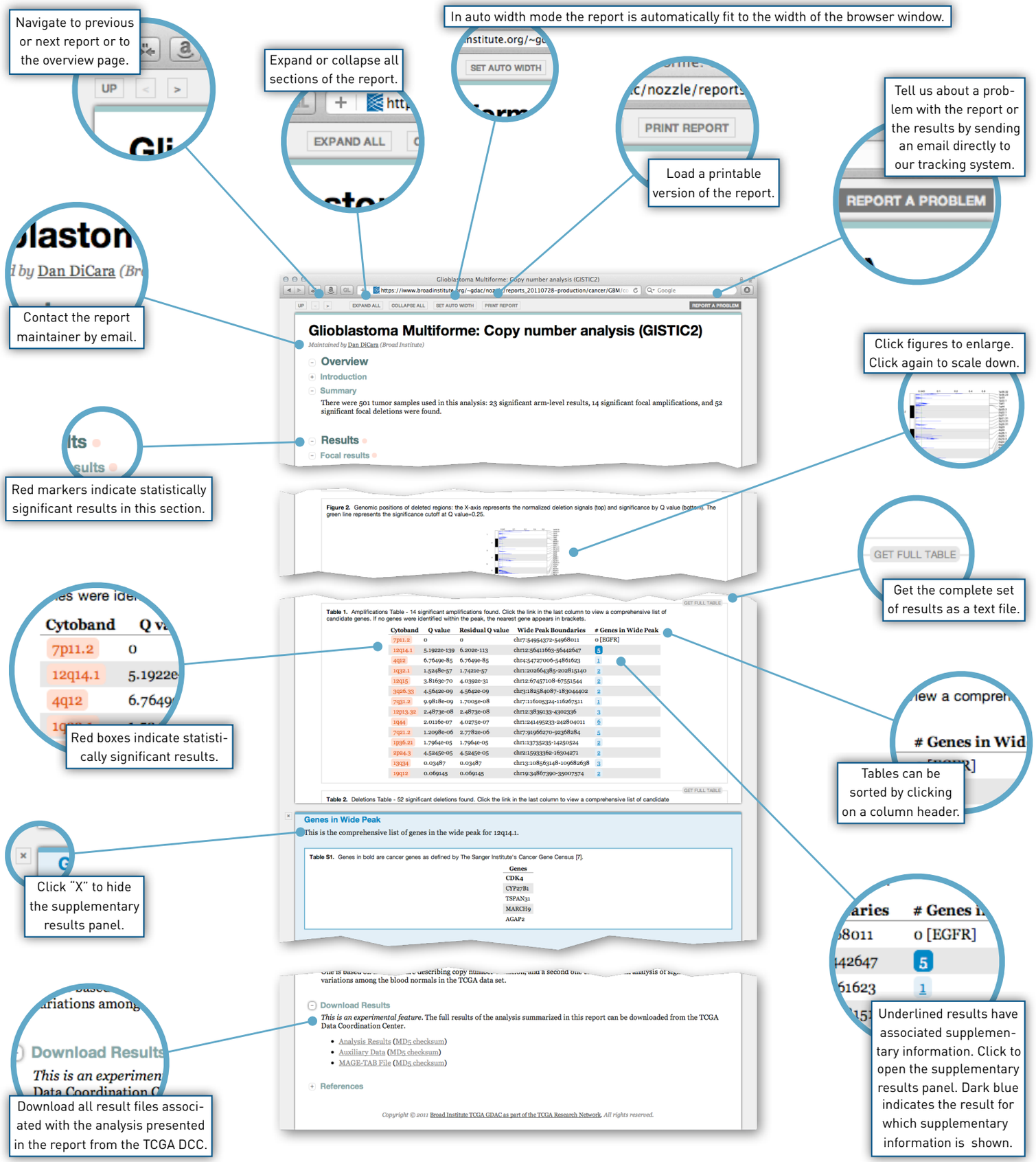
Get the complete set of results as a text file.

Tables can be sorted by clicking on a column header.

Click "X" to hide the supplementary results panel.

Download all result files associated with the analysis presented in the report from the TCGA DCC.

Underlined results have associated supplementary information. Click to open the supplementary results panel. Dark blue indicates the result for which supplementary information is shown.



Again, aimed at solid design & engineering

Nozzle package downloadable as open source

Used in multiple external projects

Programmatic, Too

```
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 data or analyses runs
- **Open access : no password**
- Select by run type & date
- Subselect by tumor type
- Or analyses type / name
- See what runs we did
- Or what tasks in each run

10K download from gdac.broadinstitute.org

% 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
```

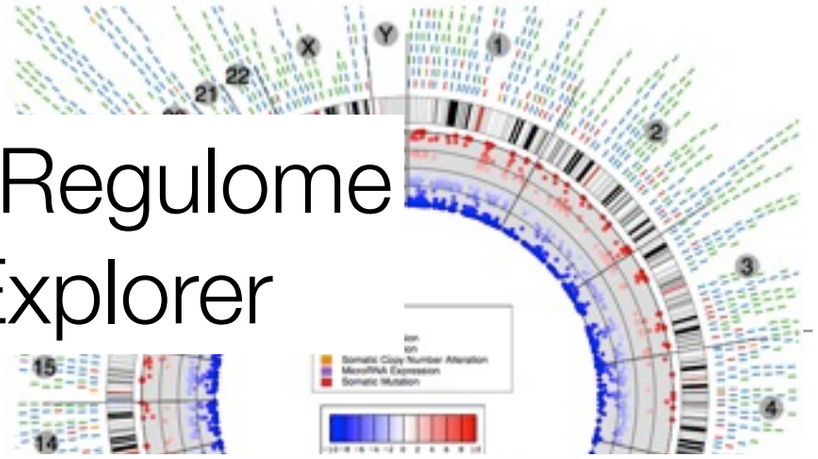
% firehose_get -tasks analyses 2012_07_25

```
...
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
...
```

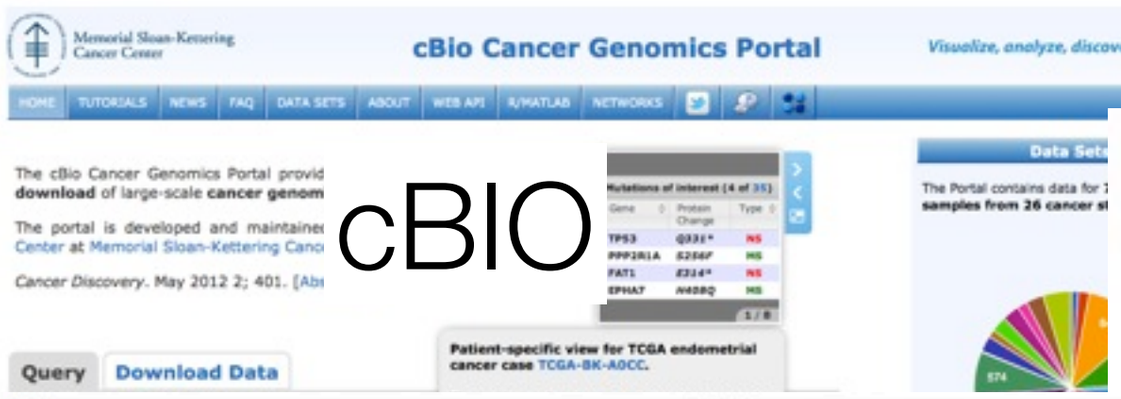
These analyses are what is described by the reports on our GDAC dashboards

And Higher-Level Portals

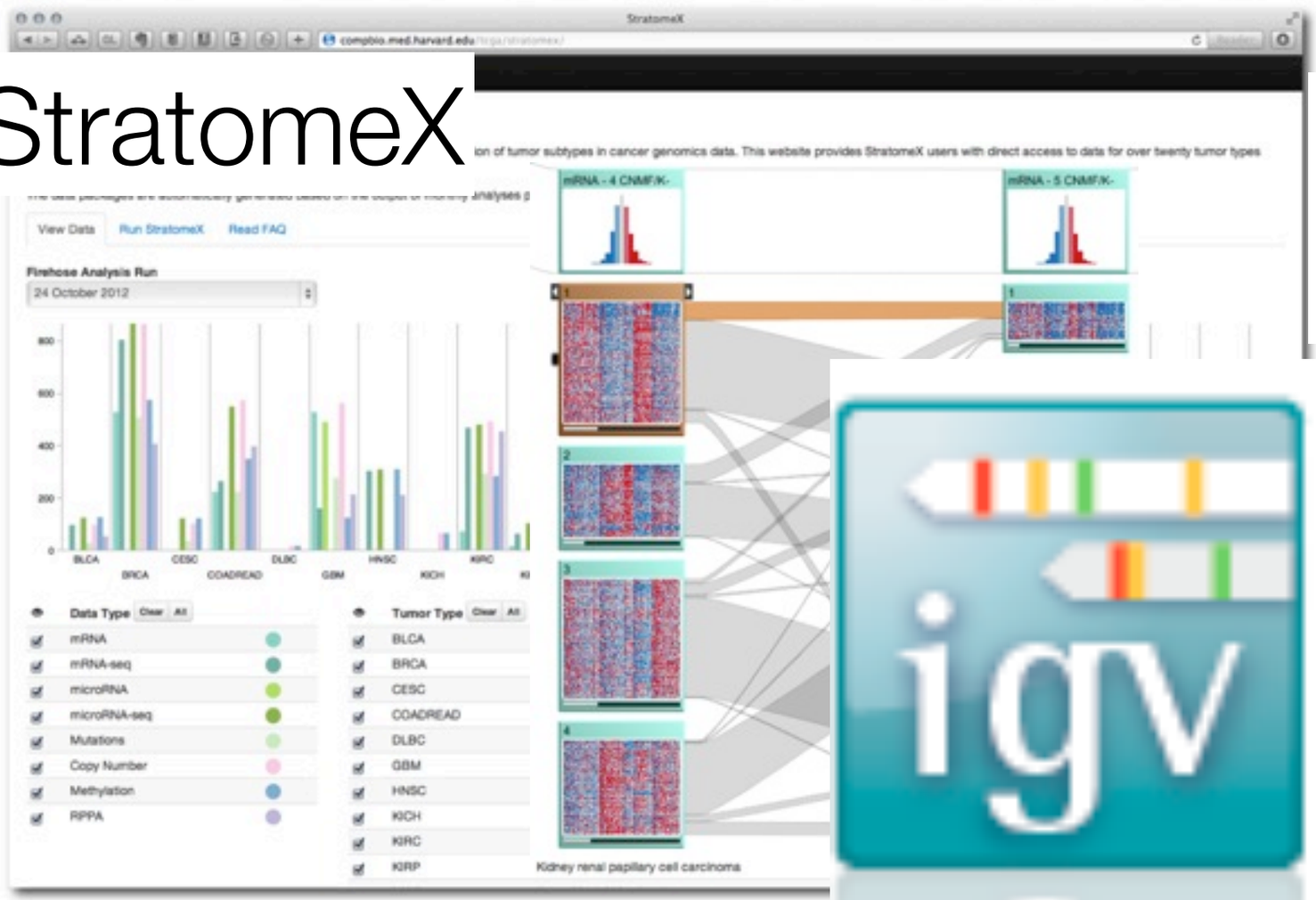
ISB Regulome Explorer



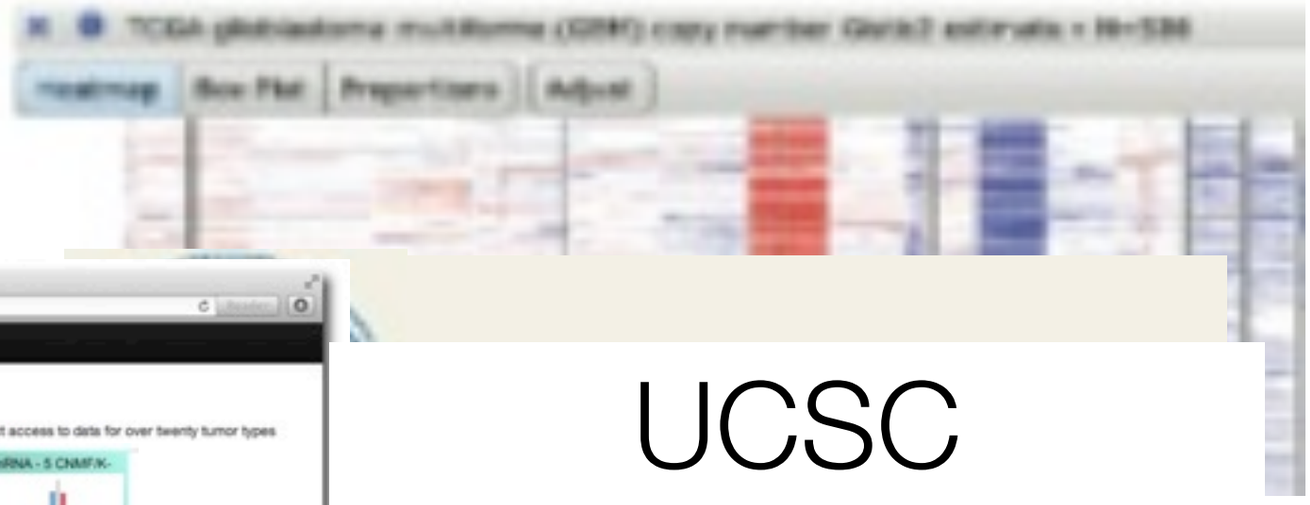
cBIO



StratomeX



UCSC Genome Browser



Integrative Genomics Viewer

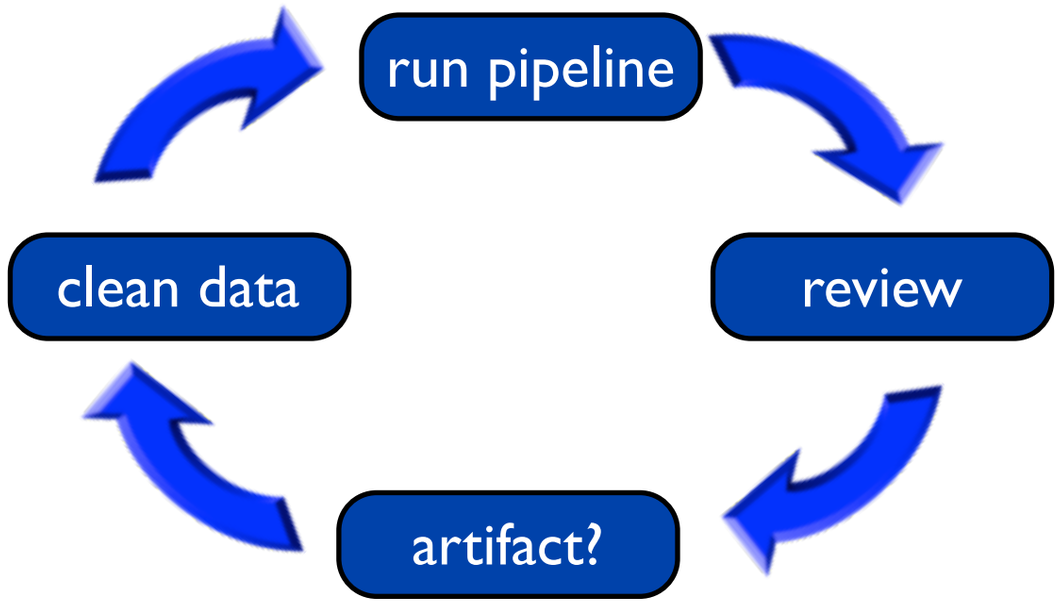
Each data run auto-loaded into (you do nothing)

Summary : Really just getting started

- TCGA::cancer ≈ Human Genome Project::genomics
- Amazing, but not just an end in itself
- Erecting signposts now: more maps & driving to come
- Decades-long impact as a catalyst
- Coming Soon to Firehose:
  - Even more codes (from 24 to 50 last 6 months)
  - Batch effects
  - Better website, with search
  - IGV & Stratomex integration from dashboards

- We've talked about scale through automation ... BUT
- Humans needed **<u>in cycle</u>** to interpret biology & stats

<u>Example</u>
olfactory receptor
gene culled from list
of significant GBM
mutations,
by accounting for
expression level
and ***replication time***



<u>And Add New Codes</u>

<u>We Want You To Collaborate!</u>

For More Information

Poster 2 : PanCancer CN alteration (A. Cherniack)
Poster 15 : Double Normals (C. Stewart)
Poster 66 : StratomeX visualizer (N. Gehlenborg)
Poster 72 : Meth27 & Meth450 in Firehose (D. Heiman)
Poster 82 : Optimization for Big Data (W. Mallard)
Poster 97 : Integrative Genomics Viewer (J. Robinson)

WWW http://gdac.broadinstitute.org
Email gdac@broadinstitute.org

This Talk Will Be Posted To

https://confluence.broadinstitute.org/display/GDAC/Presentations

Thank You!