



# Introduction To Firehose: The Broad GDAC Pipeline


Michael Noble & Nils Gehlenborg  
The Broad Institute of MIT & Harvard

National Cancer Institute  
Translational Science Meeting

July 28, 2011

# Outline



- I. What is  ?
- II. Use Cases
- III. Science Content at a Glance: Reports
- IV. Perspective

# I. What Is Firehose?



- At this point you have a broad sense of TCGA goals, data stream, and portals
- But how do they come together to answer common biological questions?
- For example:
  - 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

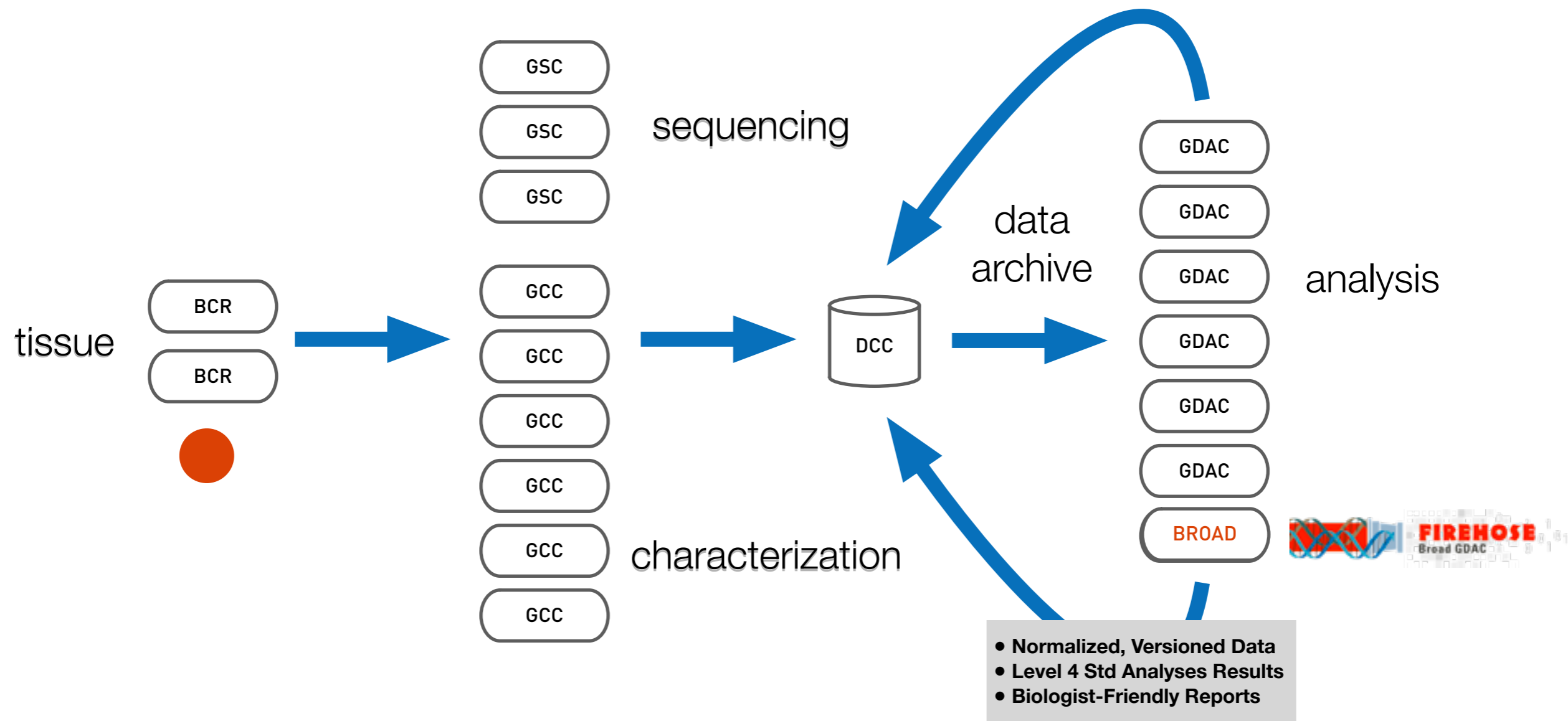
# I. What Is Firehose?

- These common results then become building blocks for higher-level analysis
- So that downstream users do not have to repeat each time
- Or ad-hoc reinvent methods for
- 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 and traceability 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.**

# Where Does Firehose Fit in TCGA?

By tracing the life cycle of a sample ...



# Firehose Goals



- 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.**



# The Bad Old Days: Manual Experiments

```
% create a folder
```

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

```
% run_your_computational_analysis
```

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

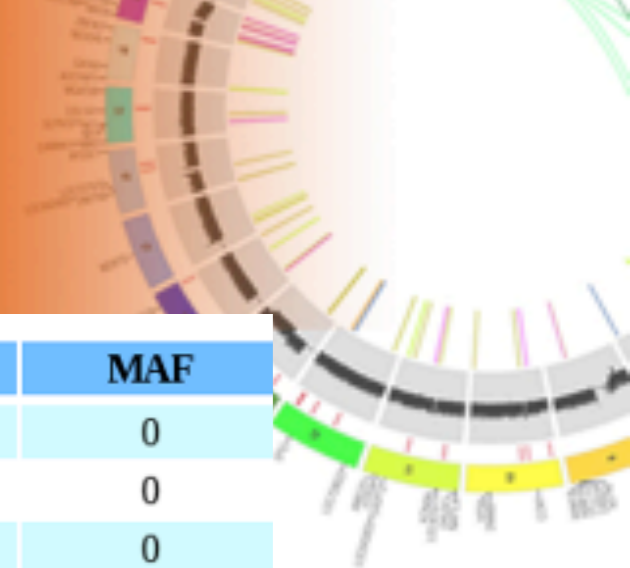
Then forget ... and search, search, search

Then repeat ALL for 19 more tumors

GBM, LUNG, AML, ...

Then multiply by 5, 10 ... comp bios at your site

# Doesn't Scale to TCGA



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
	+222	+738	+58	+486	+0	-7	+0	+291

May  
2011  
Data

} Diffs  
Since  
April

- 21 tumor sets, (up to 5 data types)
- Mutation calls for 8 tumor types

- 3085 patient cases



# 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, normalized datasets

*Generated at Broad, precursor to automated pipeline*

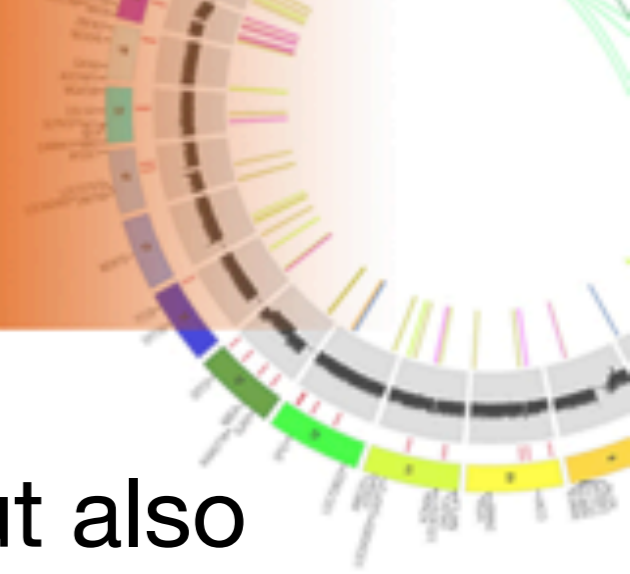
**These broadly map to 3 use cases, loosely corresponding to computational preference.**

# 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

“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
- Normalized, versioned data VERY 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

# III: Science Content at A Glance



Now Nils Gehlenborg will provide an overview of results & reports generated in Firehose.

- Our hope is that they enable readers [\(like clinical trialists\)](#)
- With just a few glances at common representational figures
- Not deep head-scratching
- To quickly take pulse of pipeline for a given tumor type



# Firehose Reports: Example 1



**Cell**  
PRESS

Cancer Cell  
**Article**

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

**Glioblastoma Multiforme: Clustering of mRNA expression: consensus NMF**

Overview

- Introduction
- Summary

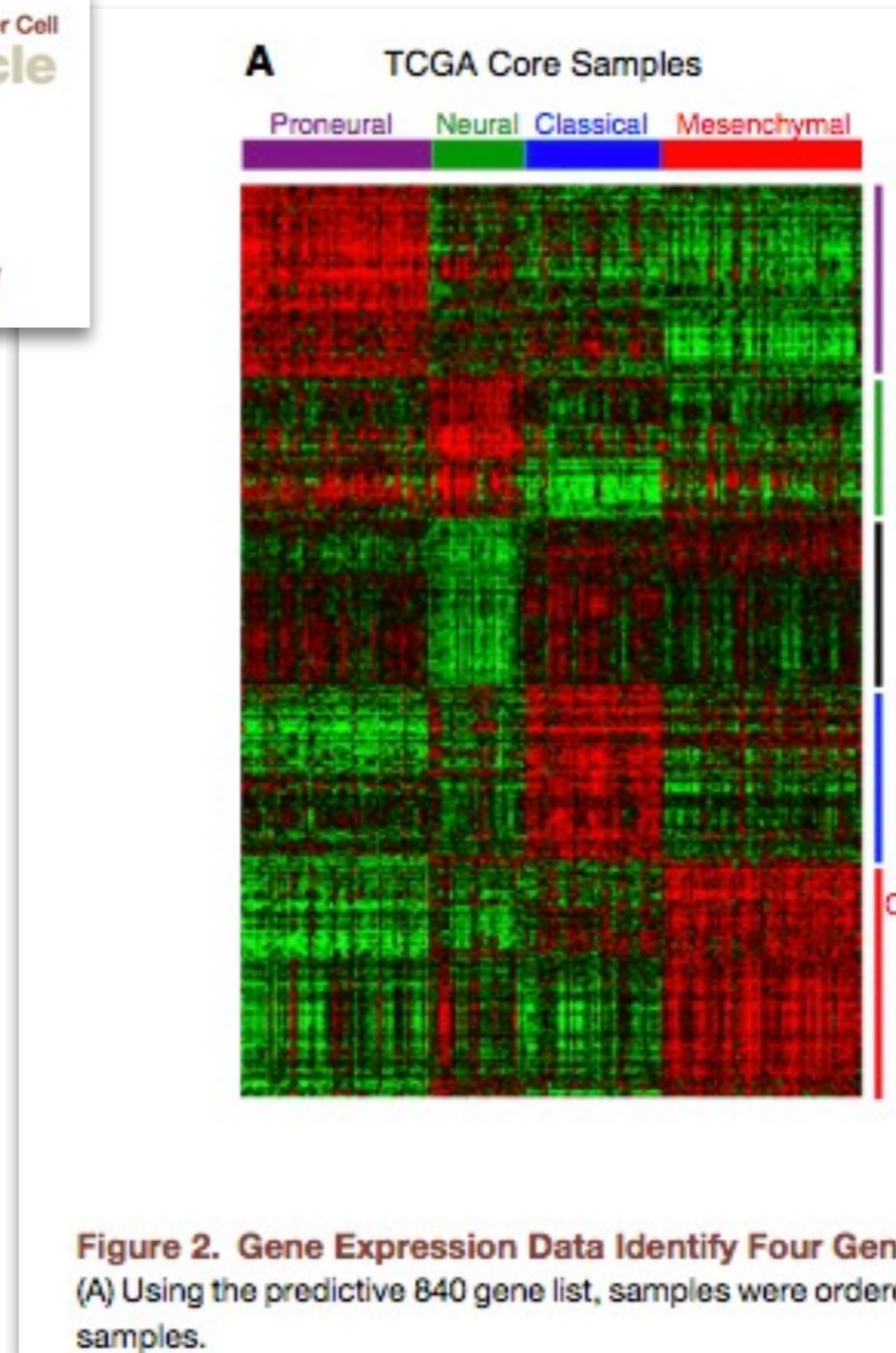
The most robust consensus NMF clustering of 492 samples using the 1500 most variable genes was identified for  $k = 4$  clusters. We compared the clustering for  $k = 2$  to  $k = 8$  and used the exploratory correlation coefficient to determine the best solution.

Results

- Gene expression patterns of molecular subtypes
- Consensus and correlation matrix

Figure 2. The consensus matrix after clustering across 4 clusters with linked nodes between clusters.

Figure 3. The correlation matrix also shows 4 clusters.



# Firehose Reports: Example 2



ARTICLE

doi:10.1038/nature10166

## Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network\*

**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.





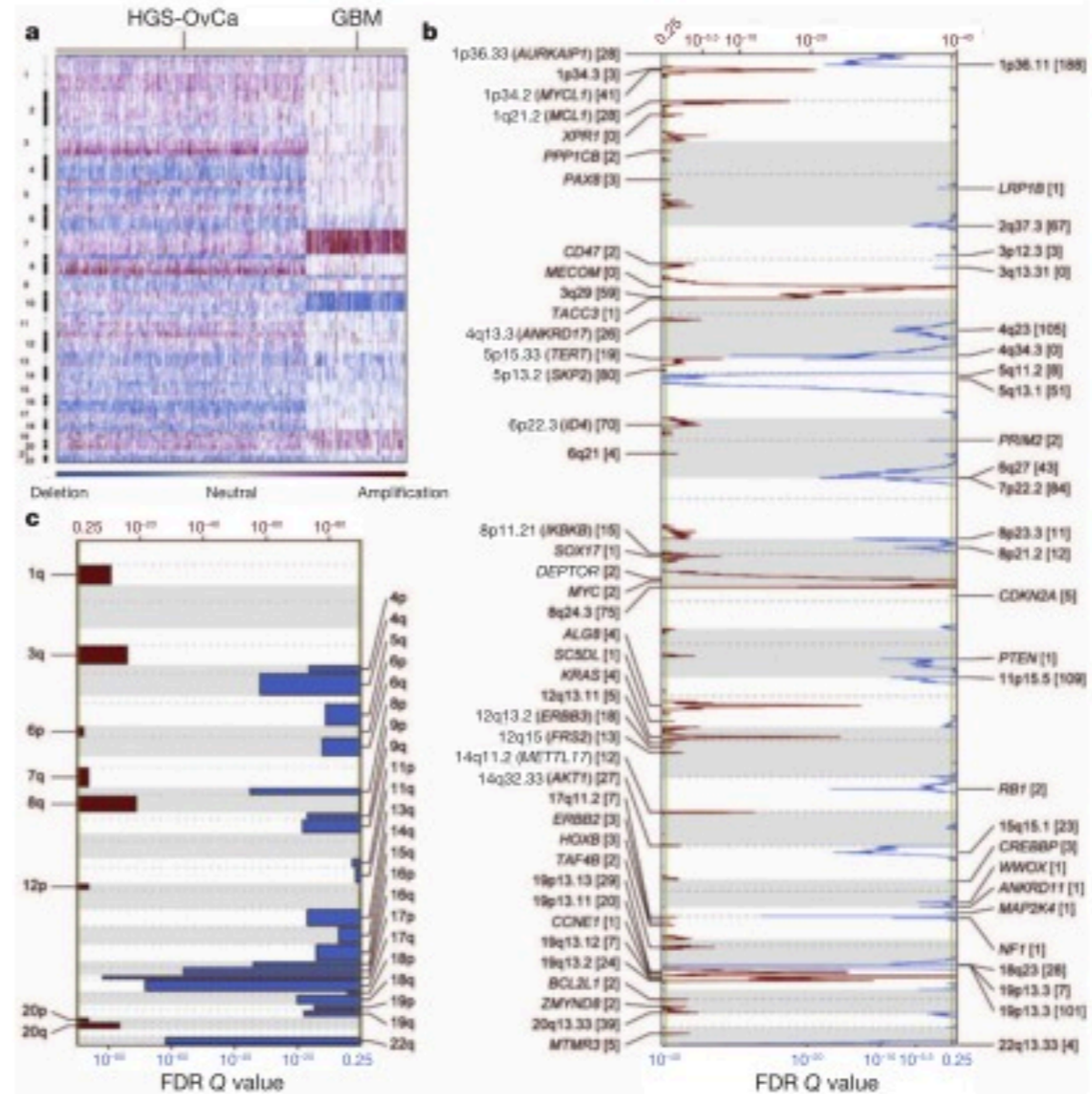
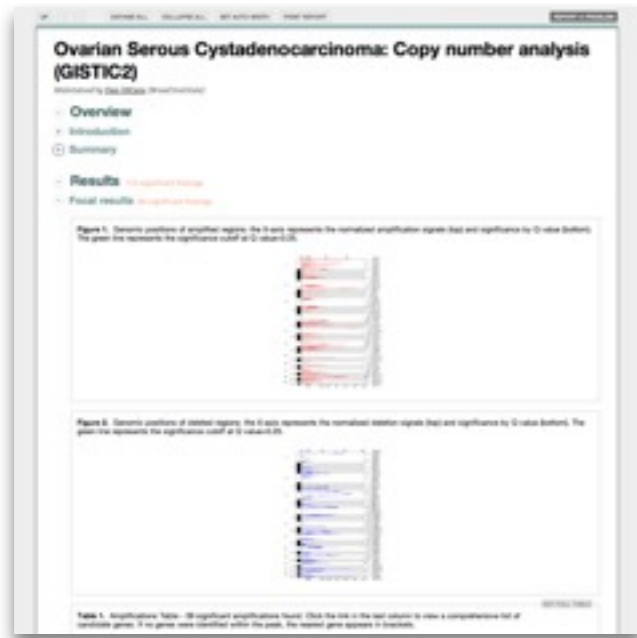
# Firehose Reports: Example 3

ARTICLE

doi:10.1038/nature10166

## Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network\*



**Figure 1 | Genome copy number abnormalities.** **a**, Copy number profiles of 89 HGS-OvCa, compared with profiles of 197 glioblastoma multiforme (GBM) tumors. Copy number increases (red) and decreases (blue) are significant amplified and deleted regions, well-localized regions with fewer genes, and regions with known cancer genes or genes identified as copy number loss of function genes. The number of genes in

# Firehose Reports: Rationale



**Make *effective* interpretation of analysis results as *efficient* as possible.**

- quick overview of an analysis
- easy to navigate to “relevant” findings
- in-depth information available

# Firehose Reports: Structure



## Report content is organized like a paper

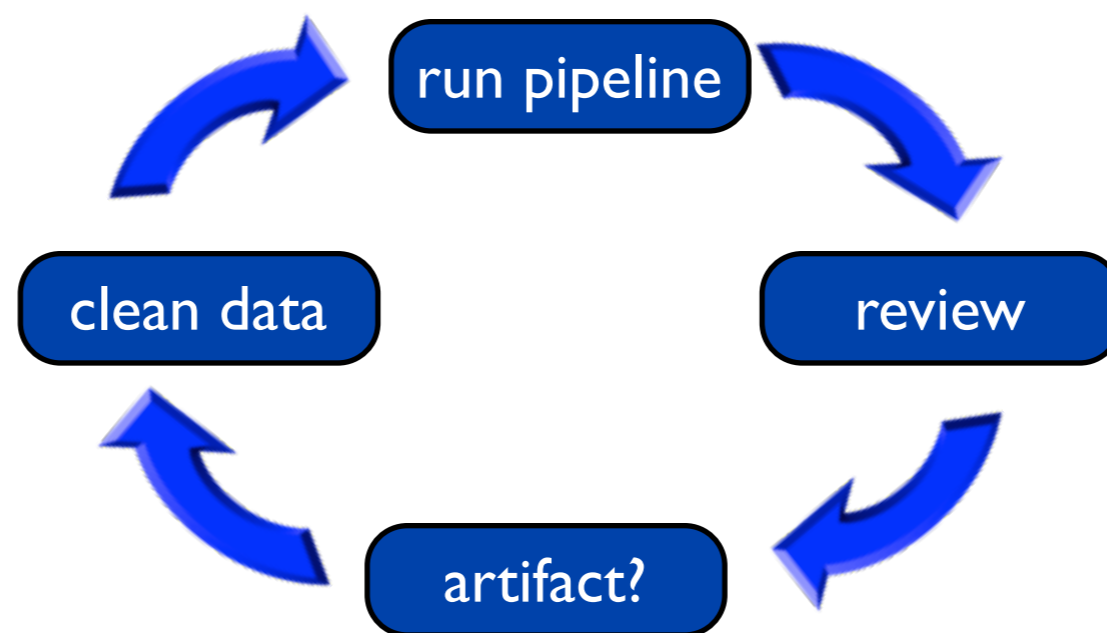
- Overview (“Abstract”)
- Results
- Methods & Data



# IV. Dumb Computers Iterating To Excellence

- Work in Progress: not easy, partly because Algorithms & data evolving rapidly  
QC still developing, in areas such as:  
**batch effects, analytic verification**
- Humans needed in **cycle** to interpret biology & stats

Example  
olfactory receptor  
gene culled from list  
of significant GBM  
mutations,  
by accounting for  
expression levels



Will  
you  
Join?

# Following Up



Reports Signup: [http://bit.ly/nci\\_tsm\\_tcga](http://bit.ly/nci_tsm_tcga)

Firehose Website: <http://gdac.broadinstitute.org>

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

