Firehose As a Data Normalization Service for TCGA

Michael S. Noble
Broad Institute of MIT & Harvard
May 12, 2011

Outline

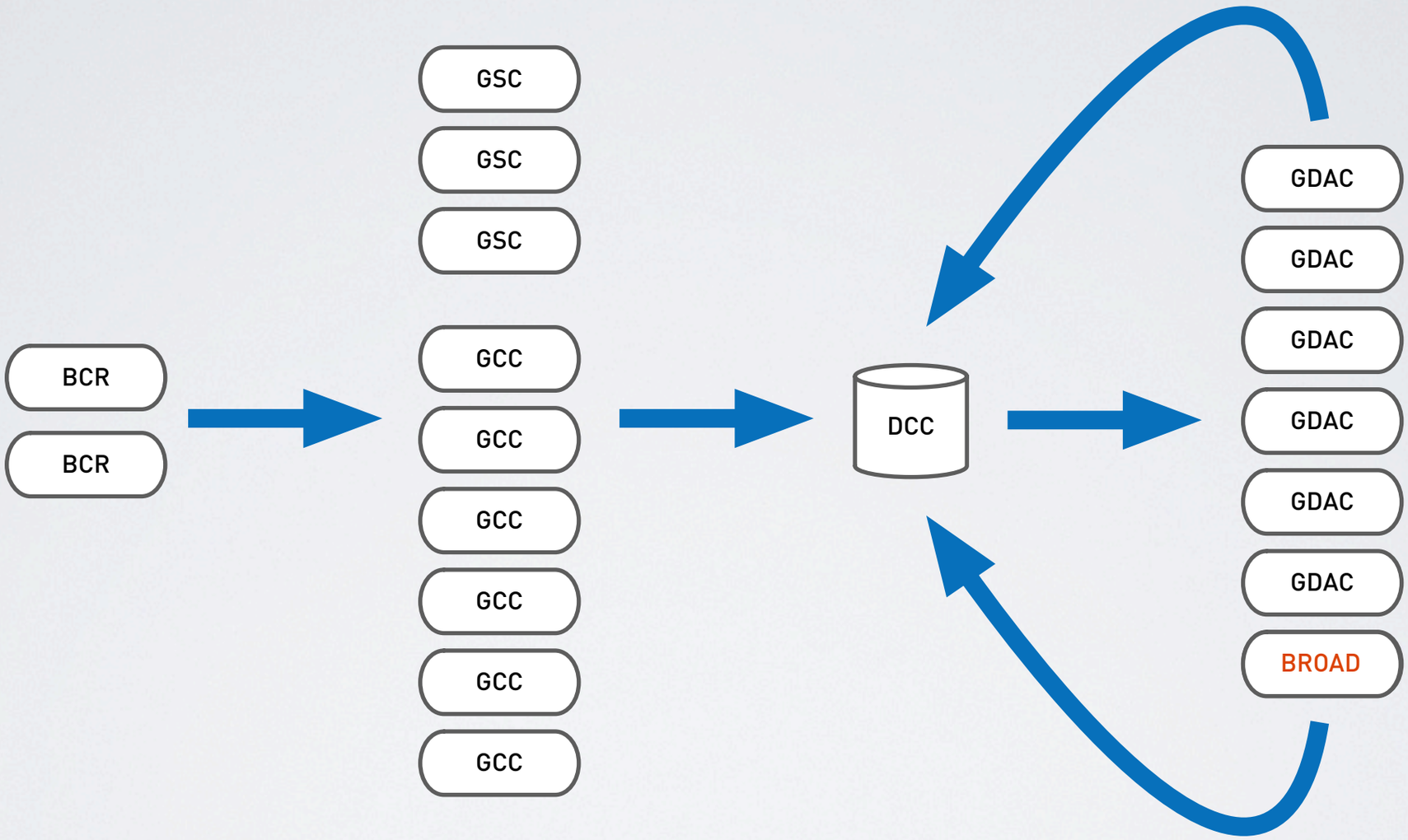
I. Data Flow

II. The Babel Problem

III. A Better Way

I : Data Flow

The Life Cycle of TCGA Data



19 Centers

NOVEMBER 2010

Tumor Type	Biospecimen #	Any level I data	clinical data	CNAs	Methylation	mRNA	miRNA	Maf File
BRCA	280	186	0	176	186	0	0	0
COAD	167	155	0	137	154	0	0	0
GBM	481	448	454	444	261	444	415	0
KIRC	213	41	19	39	40	41	0	0
KIRP	48	41	0	39	36	41	0	0
LAML	202	188	0	0	188	0	0	0
LUAD	129	33	0	21	32	33	0	0
LUSC	133	116	0	116	115	116	0	0
OV	586	571	520	570	425	568	566	384
READ	51	69	0	50	69	69	0	0
STAD	82	35	0	35	0	0	0	0
UCEC	70	24	0	24	24	0	0	0
Total	**2442**	**1907**	**993**	**1651**	**1530**	**1312**	**981**	**384**

- 12 tumor types
- 1907 patient cases
- 2442 BCR samples
- 22 Firehose analyses

- MAFs only for OV
- No TIER1 CDEs list
- Manual package/upload to DCC
- No SDRFs for results

5 Months Later: April 2011

TumorType	Biospecimen	Any_Level_1	Clinical	CNA	Methylation	mRNA	miR	MAF
BLCA	26	12	9	9	0	0	0	0
BRCA	647	390	353	375	186	434	0	0
CESC	23	8	5	8	0	0	0	0
COAD	245	151	207	182	167	155	0	88
COADREAD	338	203	285	253	236	224	0	139
GBM	508	476	465	466	288	506	415	199
HNSC	59	59	0	57	0	0	0	0
KIRC	460	347	192	345	219	72	0	0
KIRP	75	16	17	16	36	41	0	0
LAML	202	0	0	0	188	0	178	135
LGG	30	0	19	0	0	0	0	0
LIHC	38	0	0	0	0	0	0	0
LUAD	158	21	47	56	128	33	0	0
LUSC	184	161	72	142	133	134	0	0
OV	592	570	528	519	425	570	566	383
PRAD	65	0	0	0	0	0	0	0
READ	93	52	78	71	69	69	0	51
STAD	111	35	0	81	82	0	0	0
THCA	39	25	0	24	0	0	0	0
UCEC	298	24	127	133	70	0	0	0
Totals	3853	2347	2119	2484	1991	2014	1159	856
	+1411	**+440**	**+1126**	**+883**	**+461**	**+702**	**+178**	**+472**

- 8 new tumor sets (21 total)
- +1411 BCR samples (3853 total)
- 24 analyses
- MAFs for 6 tumor types
- TIER1 CDEs list for 9 tumors

∴ FLOOD OF DATA & ALGORITHMS

- Thousands of samples: 19 tumor types + clinical
- Scores of modules comprising 20+ *standard analyses*
- From 19 **decentralized** TCGA centers nationwide
- **TODAY ... AND EVOLVING DAILY**
- Standards/Coordination **NIGHTMARE**

II : The Babel Problem

We're Not Speaking the Same Language

We have 19 Centers in TCGA

And >19 Opinions on
A Central Question:

How Much Data Do We Have?

There should be ONLY ONE ANSWER

And It should not be a Matter of Opinion

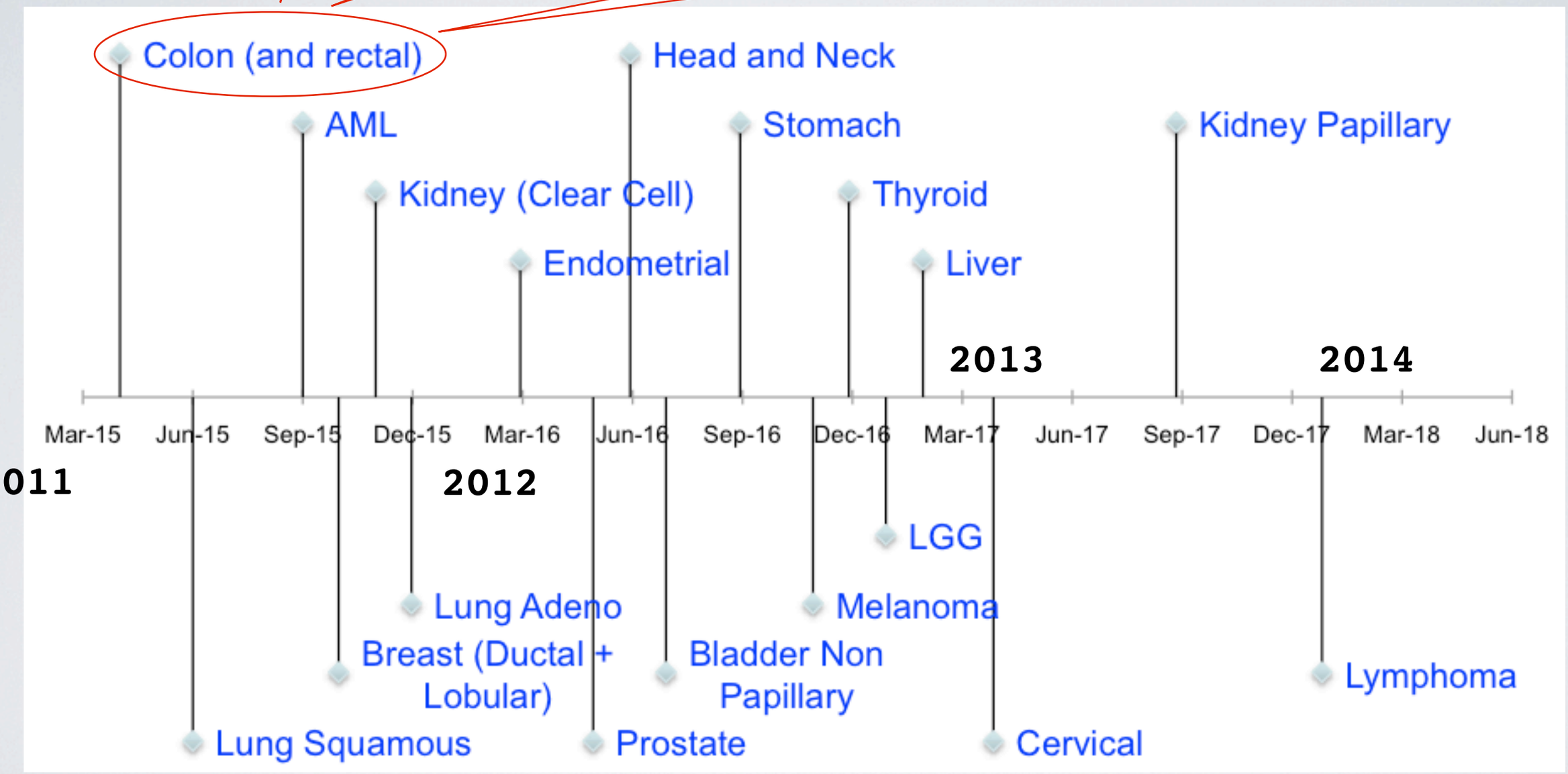
Proof: Ask Yourself ...

- How many samples does my DWG have?
- Where are they?
- Can I download myself from DCC?
- What about mutation?
- Or RNA-Seq?
- Or clinical parameters?

Datasets seem “cobbled together by hand”
Who has what samples? How many?
Where’s mutation?

Observation
“We can’t do it this way for 19 more tumor types”

Firehose missed workshop by ~1 day ...
Despite weekends & nights by several groups



TCGA Phase II Disease WG Timeline

III. A Better Way

The value of a Single, Standard, Normalized data source cannot be overstated.

And we're tantalizingly close to being able to provide it!

Vol 455|23 October 2008|doi:10.1038/nature07385

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-dimensional analysis of these molecular characteristics in human cancer and to provide the data rapidly to the research community. Here we report the interim integrative analysis of DNA copy number, gene expression and DNA methylation aberrations in 206 glioblastomas—the most common type of primary adult brain cancer—and nucleotide sequence aberrations in 91 of the 206 glioblastomas. This analysis provides new insights into the roles of *ERBB2*, *NF1* and *TP53*, uncovers frequent mutations of the phosphatidylinositol-3-OH kinase regulatory subunit gene *PIK3R1*, and provides a network view of the pathways altered in the development of glioblastoma. Furthermore, integration of mutation, DNA methylation and clinical treatment data reveals a link between *MGMT* promoter methylation and a hypermutator phenotype consequent to mismatch repair deficiency in treated glioblastomas, an observation with potential clinical implications. Together, these findings establish the feasibility and power of TCGA, demonstrating that it can rapidly expand knowledge of the molecular basis of cancer.

Cancer is a disease of genome alterations: DNA sequence changes, copy number aberrations, chromosomal rearrangements and modification in DNA methylation together drive the development and progression of human malignancies. With the complete sequencing of the human genome and continuing improvement of high-throughput genomic technologies, it is now feasible to contemplate comprehensive surveys of human cancer genomes. The Cancer Genome Atlas aims to catalogue and discover major cancer-causing genome alterations in large cohorts of human tumours through integrated multi-dimensional analyses.

The first cancer studied by TCGA is glioblastoma (World Health Organization grade IV), the most common primary brain tumour in adults[1]. Primary glioblastoma, which comprises more than 90% of biopsied or resected cases, arises *de novo* without antecedent history of low-grade disease, whereas secondary glioblastoma progresses from previously diagnosed low-grade gliomas[1]. Patients with newly diagnosed glioblastoma have a median survival of approximately 1 year with generally poor responses to all therapeutic modalities[2]. Two decades of molecular studies have identified important genetic events in human glioblastomas, including the following: (1) dysregulation of growth factor signalling via amplification and mutational activation of receptor tyrosine kinase (RTK) genes; (2) activation of the phosphatidylinositol-3-OH kinase (PI(3)K) pathway; and (3) inactivation of the p53 and retinoblastoma tumour suppressor pathways[1]. Recent genome-wide profiling studies have also shown remarkable genomic heterogeneity among glioblastoma and the existence of molecular subclasses within glioblastoma that may, when fully defined, allow stratification of treatment[3–6]. Albeit fragmentary, such baseline knowledge of glioblastoma genetics sets the stage to explore whether novel insights can be gained from a more systematic examination of the glioblastoma genome.

Results

Data release. As a public resource, all TCGA data are deposited at the Data Coordinating Center (DCC) for public access (http://cancergenome.nih.gov/). TCGA data are classified by data type (for example, clinical, mutations, gene expression) and data level to allow structured access to this resource with appropriate patient privacy protection. An overview of the data organization is provided in the Supplementary Methods, and a detailed description is available in the TCGA Data Primer (http://tcga-data.nci.nih.gov/docs/TCGA_Data_Primer.pdf).

Biospecimen collection

Retrospective biospecimen repositories were screened for newly diagnosed glioblastoma based on surgical pathology reports and clinical records (Supplementary Fig. 1). Samples were further selected for having matched normal tissues as well as associated demographic, clinical and pathological data (Supplementary Table 1). Corresponding frozen tissues were reviewed at the Biospecimen Core Resource (BCR) to ensure a minimum of 80% tumour nuclei and a maximum of 50% necrosis (Supplementary Fig. 1). DNA and RNA extracted from qualified biospecimens were subjected to additional quality control measurements (Supplementary Methods) before distribution to TCGA centres for analyses (Supplementary Fig. 2).

After exclusion based on insufficient tumour content ($n = 234$) and suboptimal nucleic acid quality or quantity ($n = 147$), 206 of the 587 biospecimens screened (35%) were qualified for copy number, expression and DNA methylation analyses. Of these, 143 cases had matched normal peripheral blood or normal tissue DNAs and were therefore appropriate for re-sequencing. This cohort also included 21 post-treatment glioblastoma cases used for exploratory comparisons

*Lists of participants and their affiliations appear at the end of the paper.

FIREHOSE
Broad GDAC

Operational 6 months

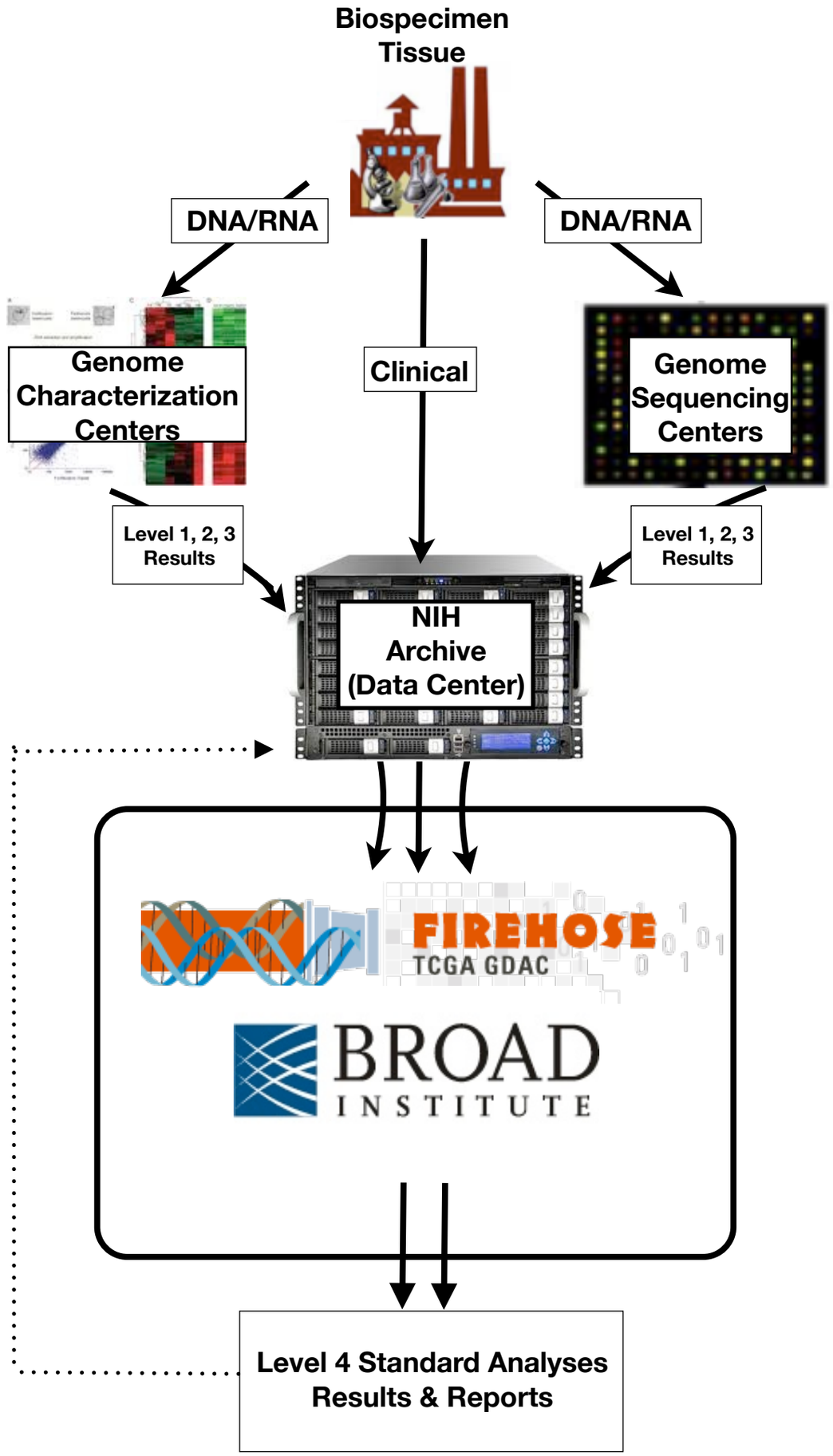
Reproduce ~90% of 2-3 years TCGA pilot results in 2-3 days

So What?

<u>Babel problem shows FH value is limited if ...</u>

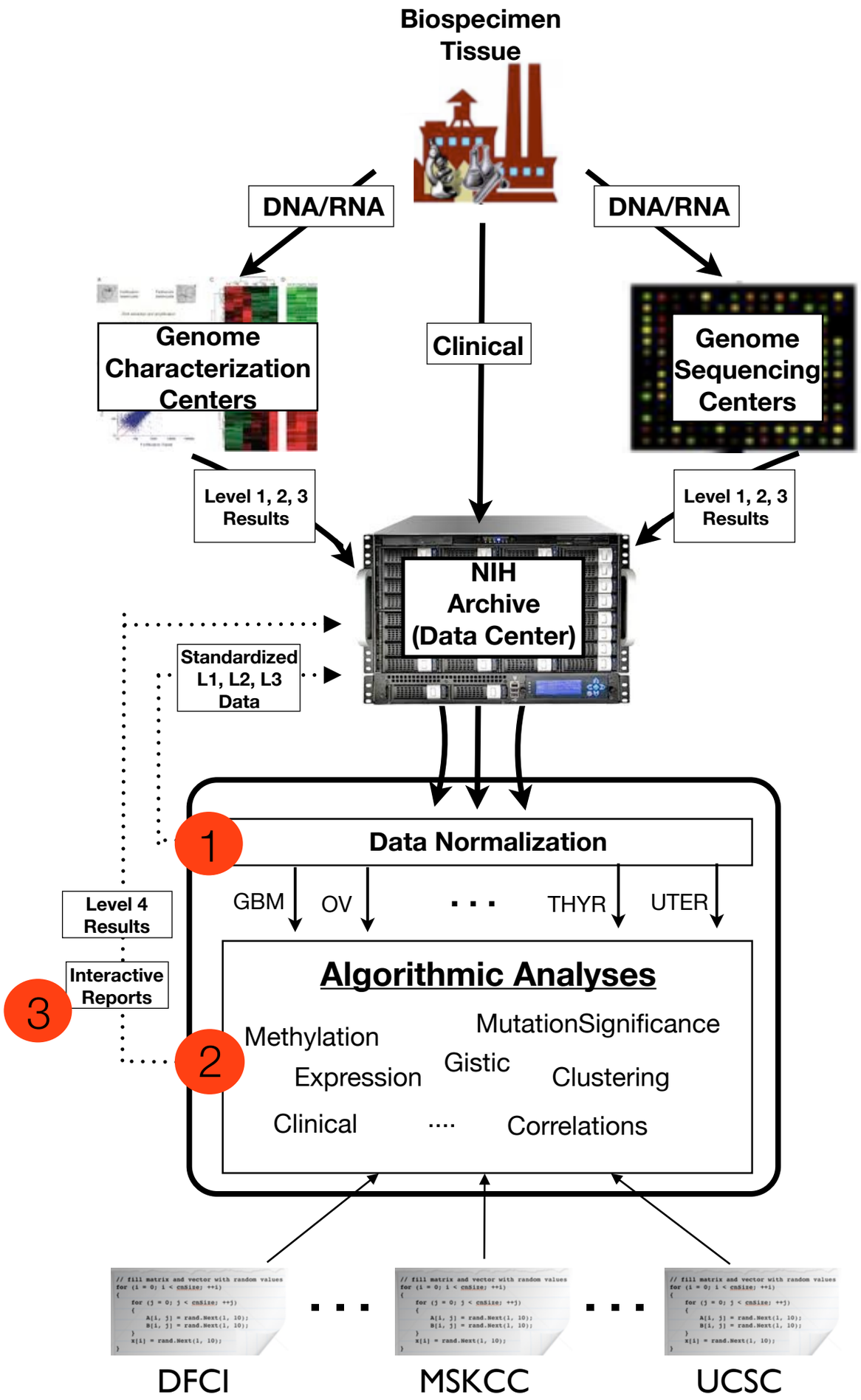
1. We do not put ***normalized*** data ...
   Into analyst & ultimately biologist hands ...
2. Via ***timely standard analyses*** ...
3. All in ***comprehensible*** form

We're VERY CLOSE, but not there yet.



Most in TCGA Share This General View

Aggregation Point For Standard Analyses



Remaining Blissfully Unaware Of Operational Details

And that “standard analyses” is meaningless without standardized data

1

Normalized Data

- Daily auto-mirror from DCC to Broad local disk
- **Partition:** to one sample per file
- **Cleanup:** remove variations problematic for automation

- Daily ingestion into FireHose DEV & PROD workspaces
- Controlled ingestion into production analyses: press GO
- **Selection:** filtered (by DNU list) samples merged ...

We use these normed data for TCGA analyses.

And claim that entire TCGA ~~should~~ must, too.

Automated Parsing of DCC Data

Gordon Saksena, Gad Getz

THE CANCER GENOME ATLAS

BROAD INSTITUTE

Abstract

Firehose provides its algorithms with data that are up-to-date and follow a regularized format. To do this, it mirrors the DCC site nightly, scans for new SDRF files, and transforms each file referenced by the SDRF file into a highly regular format, containing one sample per file. The transformation process eliminates two types of variation: that which is explicitly allowed by the spec (single vs multisample files, filenames, hybridization ids) and that on which the spec is silent (line termination styles, spaces in IDs, files with no data, uneven number of fields per line, duplicated samples, and other novel variants). Next, a collection of samples is identified, using criteria such as tumor type and exclusion lists from Disease Working Groups (DWGs) or the Biospecimen Core Resource (BCR); we would like to also use clustering group membership. Finally, the chosen collection of per-sample files is merged together into a single file, providing Firehose-hosted algorithms with the latest submitted data in a consistent format.

Clinical Data Normalization

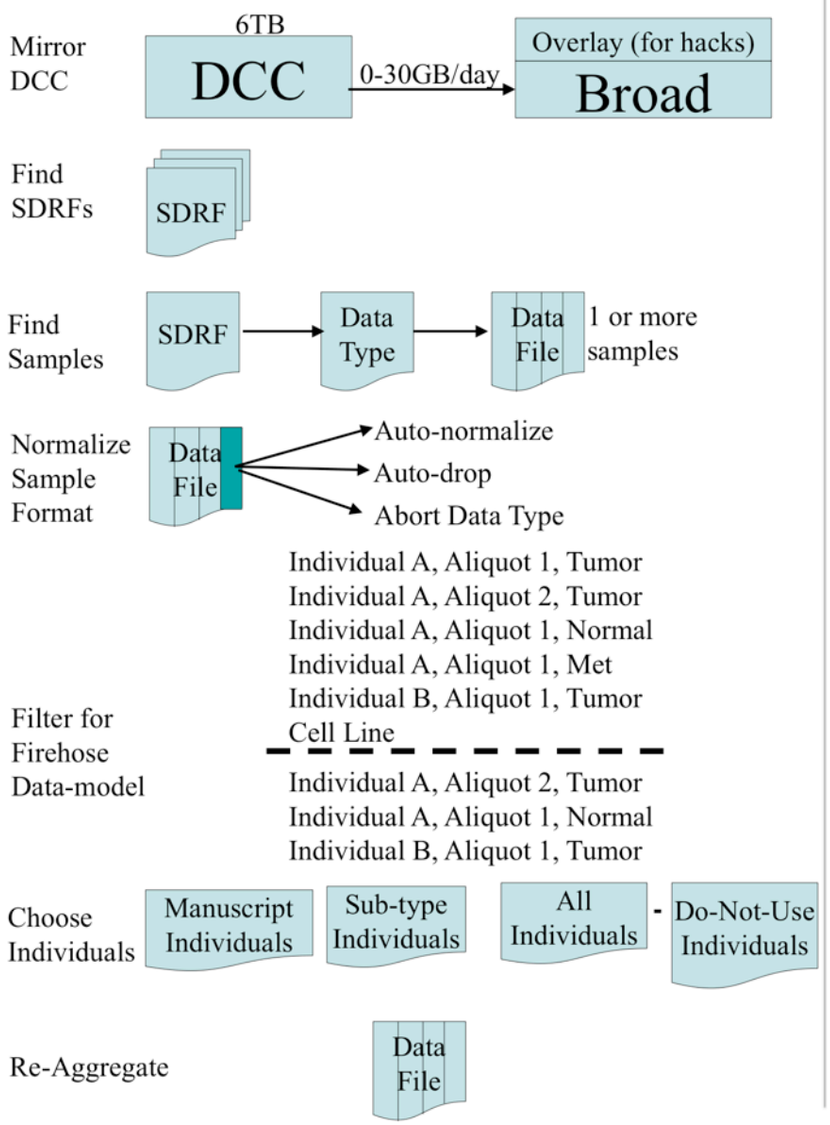
XML -> 2 column parameter-value pair

patient.tumortissuesite	ovary
patient.drugs.drug-2.drugname	cisplatin

Aggregate columns, with superset of parameters

Map certain parameters to short name - Volatile

Data Flow



Data Variations

Tolerable Variation	Approach
1 or many samples per datafile	1 sample per datafile
Hybridization ID different than TCGA ID	Replace Hybridization ID with TCGA ID
LF or CR/LF line terminators	LF line terminators
4 dialects of seg files, one per center	1 dialect of seg files: 6 columns, chr='1' – '26'
Empty fields	Map empty fields to NA
Arbitrary filename	Standardized filename
Include for Analysis= no	Auto-drop data

Awkward Variation	Approach
Missing SDRF file (BCRs, GSCs)	Auto-generate SDRF via frail heuristics
Nonsense data, eg column of all 'null'	Abort datatype, manually delete samples
Malformed TCGA ID, eg with trailing spaces	Abort datatype, manually edit SDRF file
Variable number of columns per row	Abort data type or SDRF
Header columns do not match other samples	Abort data type, manual reset if all resubmitted
Data file not found in expected directory	Abort data type

Poster from Nov 2010 F2F
Available online @ gdac.broadinstitute.org

2 Timely Analyses

- Switching to multiple runs per month
- Standard Analyses: on all tumors
- TOO Analyses: targets of opportunity
  - Such as manuscripts or DWG workshops
  - Example: 2 runs performed in April
    - Standard run
    - TOO: for May 2 LUNG DWG in NC
      (which largely served intended purpose)

2 Standard Analyses

- FH analyses not just archival / community
- Can be used in realtime
- As baselines for DWG work
- Addressing the fundamental problems:

"How much data do we have?
"We can't do it this way for 19 more tumor types"

- But Broad needs to SDRF normed data for DCC

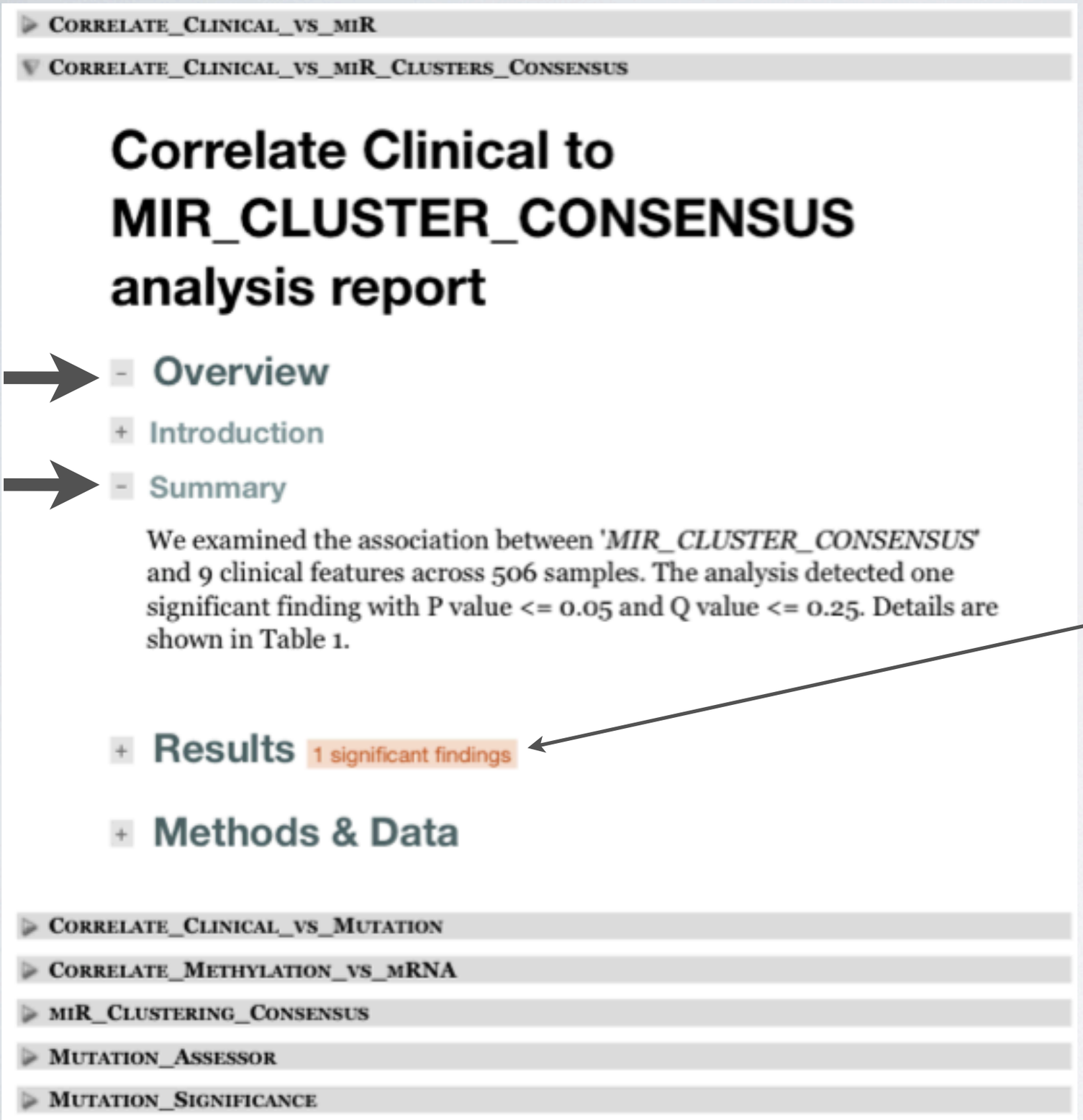
3

Comprehensible

Nozzle : Analyst & Biologist-Friendly Reports

1. All have same **structure**.
2. And same **layout**.
3. Quickly **guide** reader from summary to details.
4. With **advanced features** like foldable sections & zoomable figures.
5. Created with a **simple** set of instructions.
6. Companion to most standard analyses produced by Firehose

Nozzle : PAN-CANCER Dataset Example



- Standard visual format for ALL pipelines
- Interactive! Not just static display
- Intelligently Scoped:
  - drill from overview to details
  - Significant results “bubble up”
  - **don’t miss needle in haystack**
- Embedded tags: <INTRO>, <RESULTS>, ...
- Enable automatic processing:
  - auto-aggregation to summary report
  - focused mining in external tools (TAP)

Correlate Clinical to MIR_CLUSTER_CONSENSUS analysis report

+ **Overview**

- **Results**

- Overview of the results_

GET FULL TABLE

Table 1. Overview of the association results between 1 clustering variables and 9 clinical features. Shown in the table are P values (Q values). Thresholded by P value <= 0.05 and Q value <= 0.25, one significant finding detected.

Clinical Features		MIR_CLUSTER_CONSENSUS
Time to Death	survival	0.0136 (0.123)
Time to Recurrence	survival	0.457 (1.00)
AGE	continuous	0.299 (1.00)
KARNOFSKY.PERFORMANCE.SCORE	continuous	0.8 (1.00)
NEOADJUVANT.THERAPY	binary	0.646 (1.00)
PRIMARY.SITE.OF.DISEASE	multiclass(3)	0.156 (1.00)
TUMOR.GRADE	binary	0.549 (1.00)
TUMOR.STAGE	multiclass(4)	0.174 (1.00)
BATCH.NUMBER	multiclass(12)	0.575 (1.00)

+ **Methods & Data**

Interactivity: Drill Down To Significant Findings

View Summary Tables

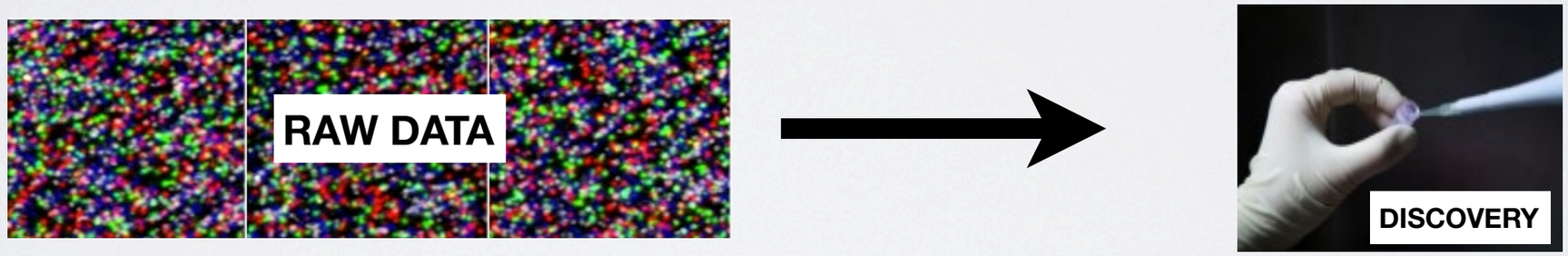
Or Fully Expanded

Take Aways

- Significant progress across TCGA

- But Holy Grail

  ✓ Data in hands of non-computational biologist
  ✓ Used as comprehensible baseline for AWG
  ✓ Facilitating the transformation of



Remains to be fully realized.

- With small tweaks, Firehose can HELP!

What I Need From You

- Clearer schedule of DWG activities

- Better sense of analyses / data (sub)groupings

  - ✓ We are starting to write “individual set service”
  - ✓ To allow easy subsetting/aggregating
  - ✓ Of individual/sample sets
  - ✓ Without needing Firehose login credentials
  - ✓ Will also appear in TAP: TCGA Analysis Portal

- Example: potential colorectal analyses (from Adam Bass)

  - ✓ All samples
  - ✓ All colon vs. All rectal
  - ✓ All non-hypermutated
  - ✓ Proximal vs. distal
  - ✓ ALL KRAS, BRAF, NRAS wild-type

gdac@broadinstitute.org

gdac.broadinstitute.org

These are Your FRIENDS
Use Them!

Broad GDAC Analysis Summary
2011_04_21 Run

Tables of Ingested Data: HTML PNG TSV

Tumor Type	# Completed	Percentage
OV	24	100%
GBM	24	100%
COAD	14	58%
READ	13	54%
FULL	13	54%
COADREAD	13	54%
LUSC	12	50%
LUAD	12	50%
BRCA	12	50%
KIRC	10	42%
KIRP	9	38%
UCEC	4	17%
CESC	4	17%
BLCA	4	17%
STAD	3	13%
HNSC	3	13%
THCA	2	8%
LAML	2	8%
LGG	1	4%
PRAD	0	0%
LIHC	0	0%

Summary of TCGA Tumor Data
Ingested into Broad GDAC Pipeline
04/21/2011 Run

TumorType	Biospecimen	Any_Level_1	Clinical	CNA	Methylation	mRNA	miR	MAF
BLCA	26	12	9	9	0	0	0	0
BRCA	647	390	353	375	186	434	0	0
CESC	23	8	5	8	0	0	0	0
COAD	245	151	207	182	167	155	0	88
COADREAD	338	203	285	253	236	224	0	139
GBM	508	476	465	466	288	506	415	199
HNSC	59	59	0	57	0	0	0	0
KIRC	460	347	192	345	219	72	0	0
KIRP	75	16	17	16	36	41	0	0
LAML	202	0	0	0	188	0	178	135
LGG	30	0	19	0	0	0	0	0
LIHC	38	0	0	0	0	0	0	0
LUAD	158	21	47	56	128	33	0	0
LUSC	184	161	72	142	133	134	0	0
OV	592	570	528	519	425	570	566	383
PRAD	65	0	0	0	0	0	0	0
READ	93	52	78	71	69	69	0	51
STAD	111	35	0	81	82	0	0	0
THCA	39	25	0	24	0	0	0	0
UCEC	298	24	127	133	70	0	0	0
Totals	3853	2347	2119	2484	1991	2014	1159	856

	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

The End!