Firehose As A Data Normalization Service for TCGA

Michael S. Noble
Broad Institute of MIT & Harvard
May 19, 2011

The Babel Problem

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

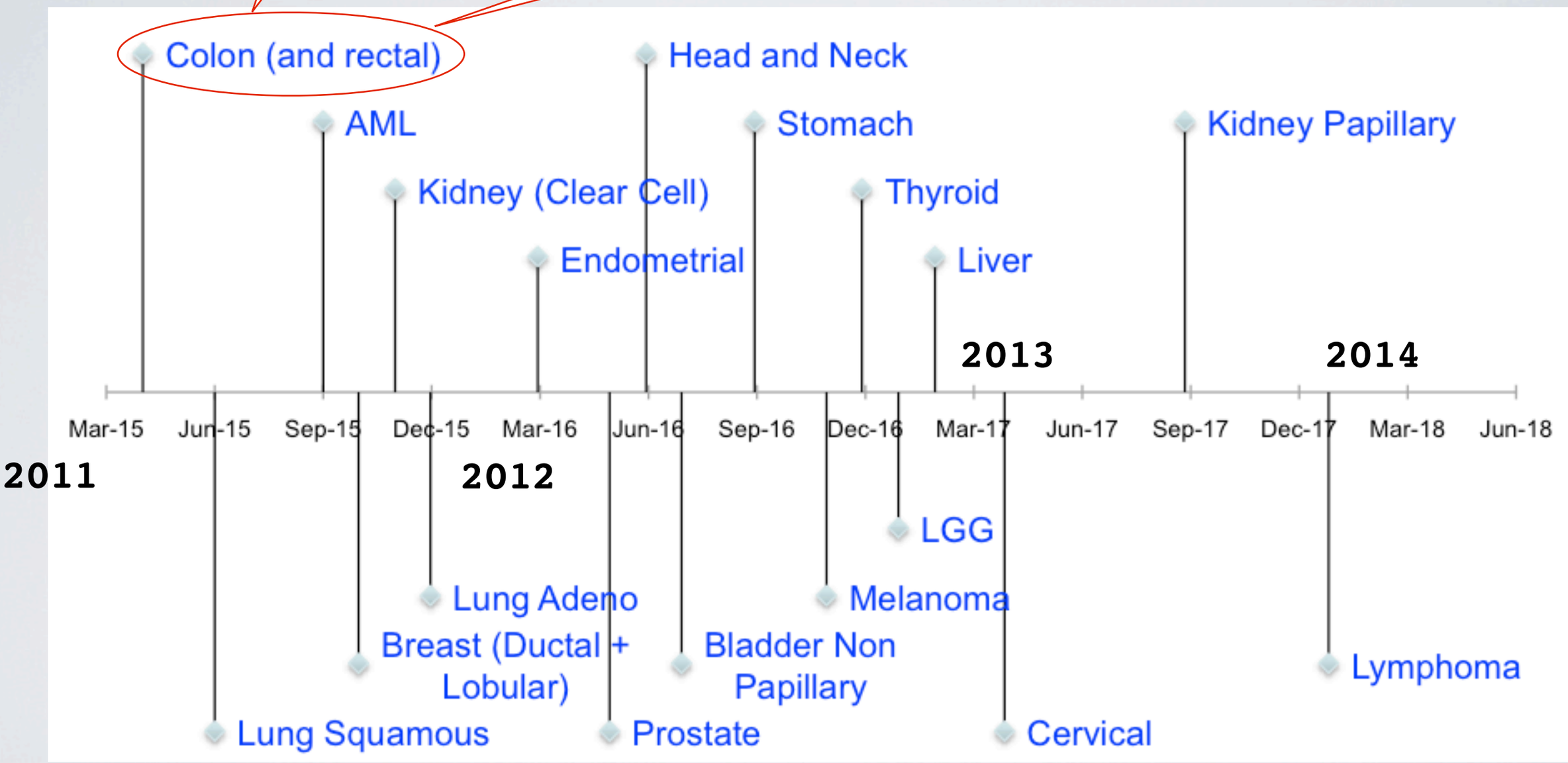
PROOF: ASK YOURSELF ...

- How many samples does my AWG have?
- Where are they?
- What about mutation?
- Or RNA-Seq?
- Or clinical parameters?

∴ *The practical value of a <u>cannonical</u> data source cannot be overstated.*

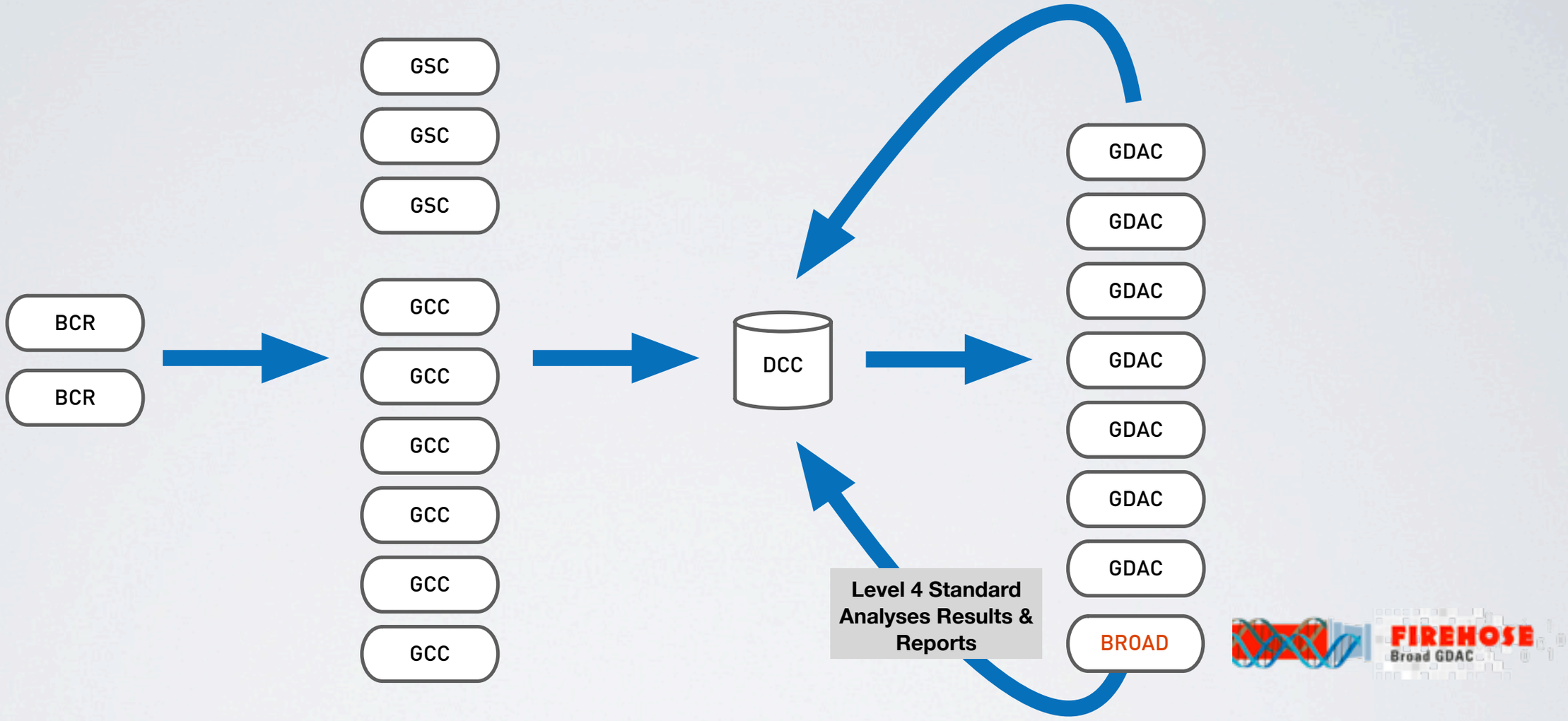
Datasets seem “cobbled together by hand”
Who has what samples? How many?
Where’s mutation?

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

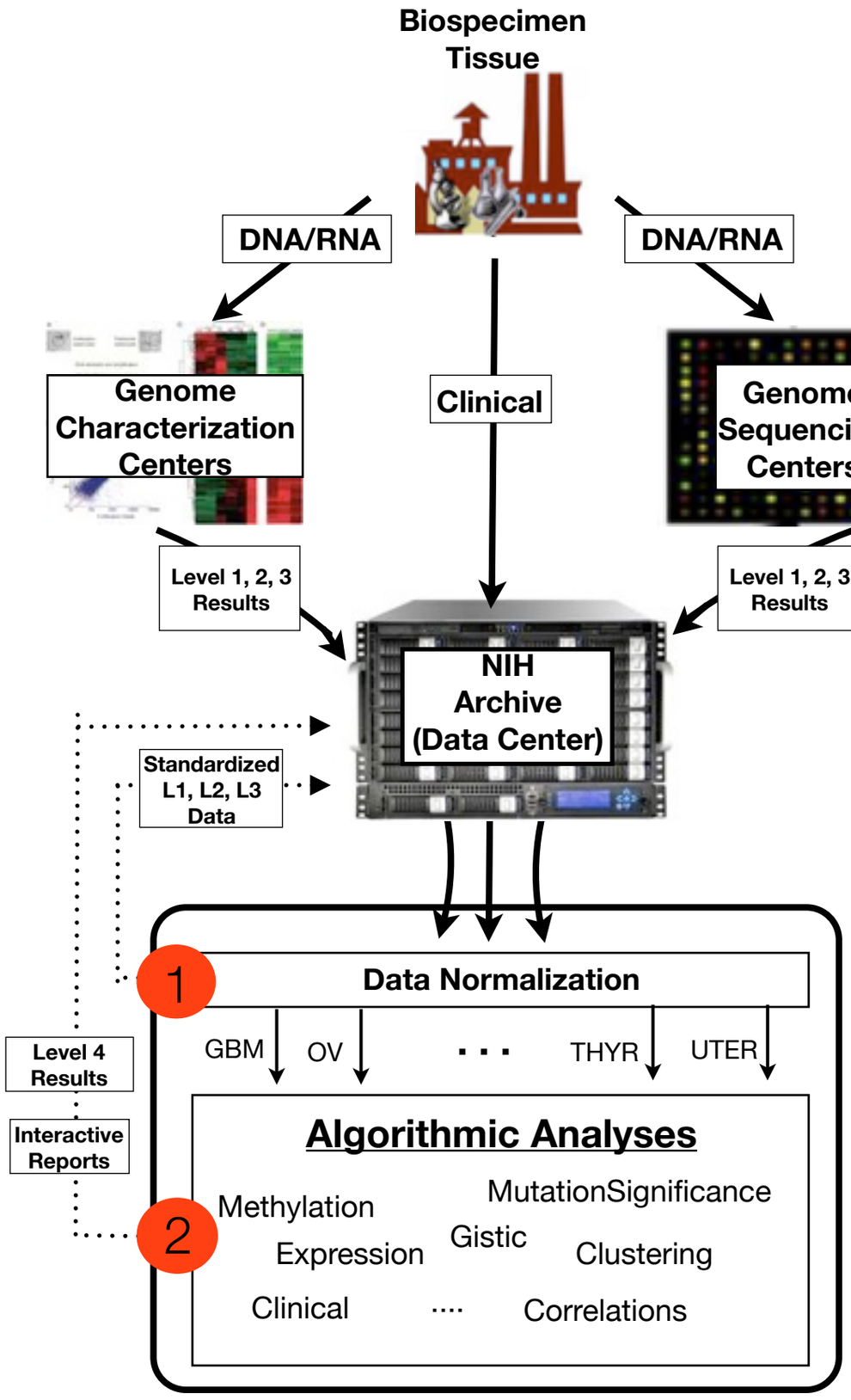


Example: March Colorectal Workshop

CURRENT VIEW OF FIREHOSE IN TCGA



Monthly Aggregation Point for “Standard Analyses”



BUT

"Standard Analyses"
is meaningless without
standardized data.

Therefore, to facilitate 2
Firehose internally does 1

With little additional work
1 can ALSO address
the Babel problem.

Adding new value to TCGA,
as "FH data norming" service

How Would This Play Out?

Current

- Standard Runs: analysis of all tumor data, monthly

New

- Data Runs: norming service fosters TCGA-wide "standard view" of data; ad-hoc analyses use versioned datasets

New

- TOO Runs: analysis targets of opportunity, e.g. for coordinated activity such as AWG workshops

- Example: 2 runs performed in April 2011

  - Standard run 4/21/2011
  - TOO for May 2 LUNG workshop in NC (which largely served intended purpose)

Where We Need More Feedback

- Clearer schedule of AWG activities

- Better sense of analyses to perform & data (sub)groupings
  - ✓ We are starting to write “individual set service”
  - ✓ For easy subsetting/aggregating of individuals & samples
  - ✓ Without needing Firehose login credentials
  - ✓ Will also appear in TAP: TCGA Analysis Portal

- Example: potential colorectal analyses (A. Bass & MSKCC et al)
  - ✓ 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

Appendix: How Does FH Norm Data?

- Daily automatic mirror from DCC to Broad
- **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 ...

See Gordon Saksena poster from Nov 2010 F2F
(next slide, or online at gdac.broadinstitute.org)

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

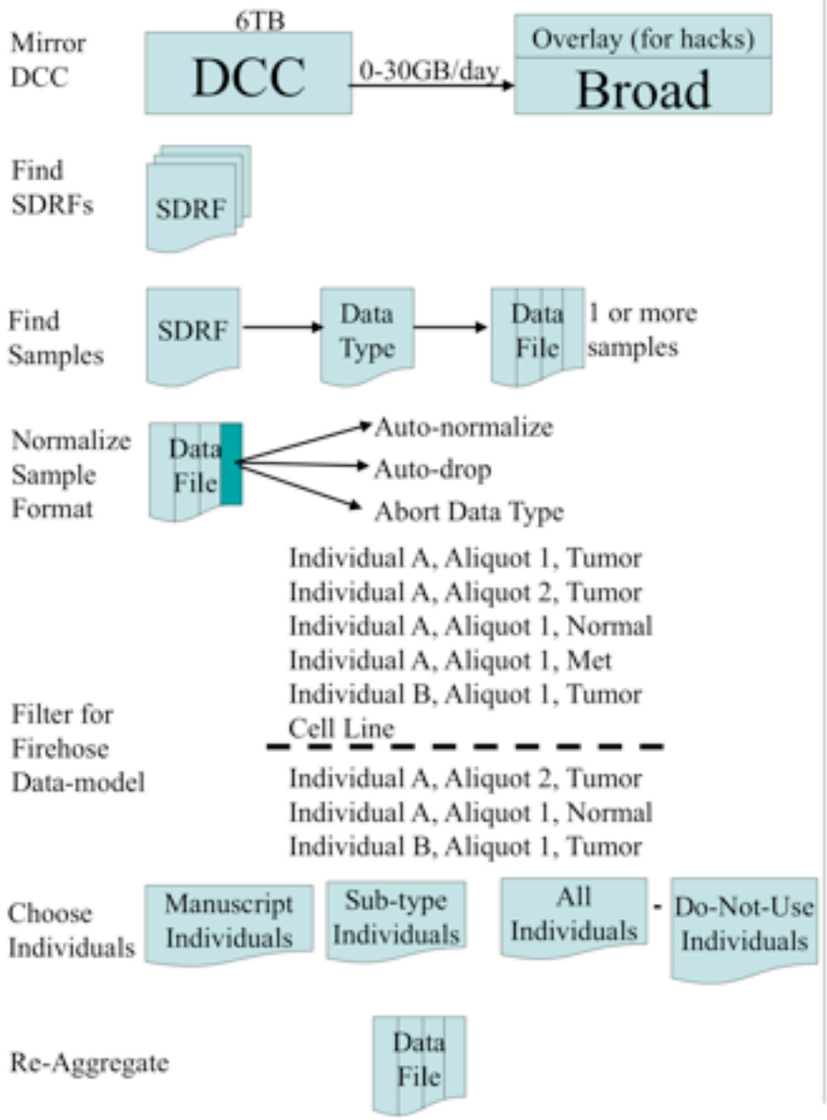
XML -> 2 column parameter-value pair

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