The Broad Institute GDAC Pipeline

Michael S. Noble
April 27, 2011

Outline

I. Why

II. Past

III. Present

IV. Future

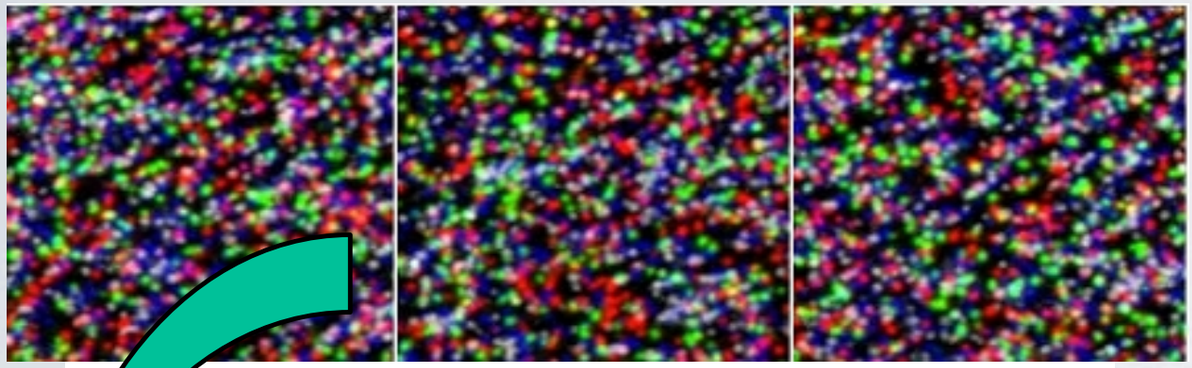
I : Why

Apart from the fact that
we love our families and friends.

FLOOD OF DATA & ALGORITHMS

- Thousands of samples: 19 tumor types + clinical
- 20+ analyses comprised of scores of modules
- From 19 TCGA centers nationwide
- **TODAY ... AND EVOLVING DAILY**
- Standards and Coordination **NIGHTMARE**

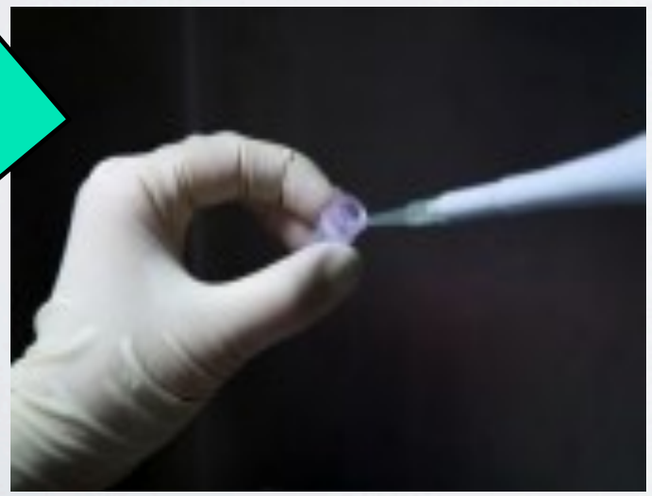
GET DATA WHERE IT BELONGS: INTO BIOLOGIST HANDS



Type	Biospecimen	Any_Level_1	Clinical	CNA	Methylation	mRNA	miR	MAF
CA	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								139
GBM								199
HNSC								0
KIRC								0
KIRP								0
LAML								135
LGG								0
LIHC								0
LUAD								0
LUSC								0
OV								383
PRAD								0
READ								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
Full	3853	2347	2119	2484	1991	2014	1159	856

April 21, 2011 Snapshot

A total of 2347 cases from 19 tumor types, with clinical annotations and up to 6 different genomic data types



FIREHOSE
Broad GDAC

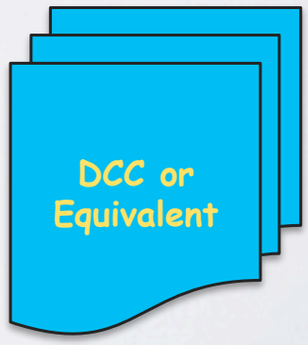
Data Aggregation & Normalization

GBM OV ... THYR UTER

Algorithmic Analyses

Methylation
Mutation Significance
Expression
GISTIC
Clustering
Clinical
....
Correlations

Biologist-friendly Reports



Aggregated and normalized data files

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

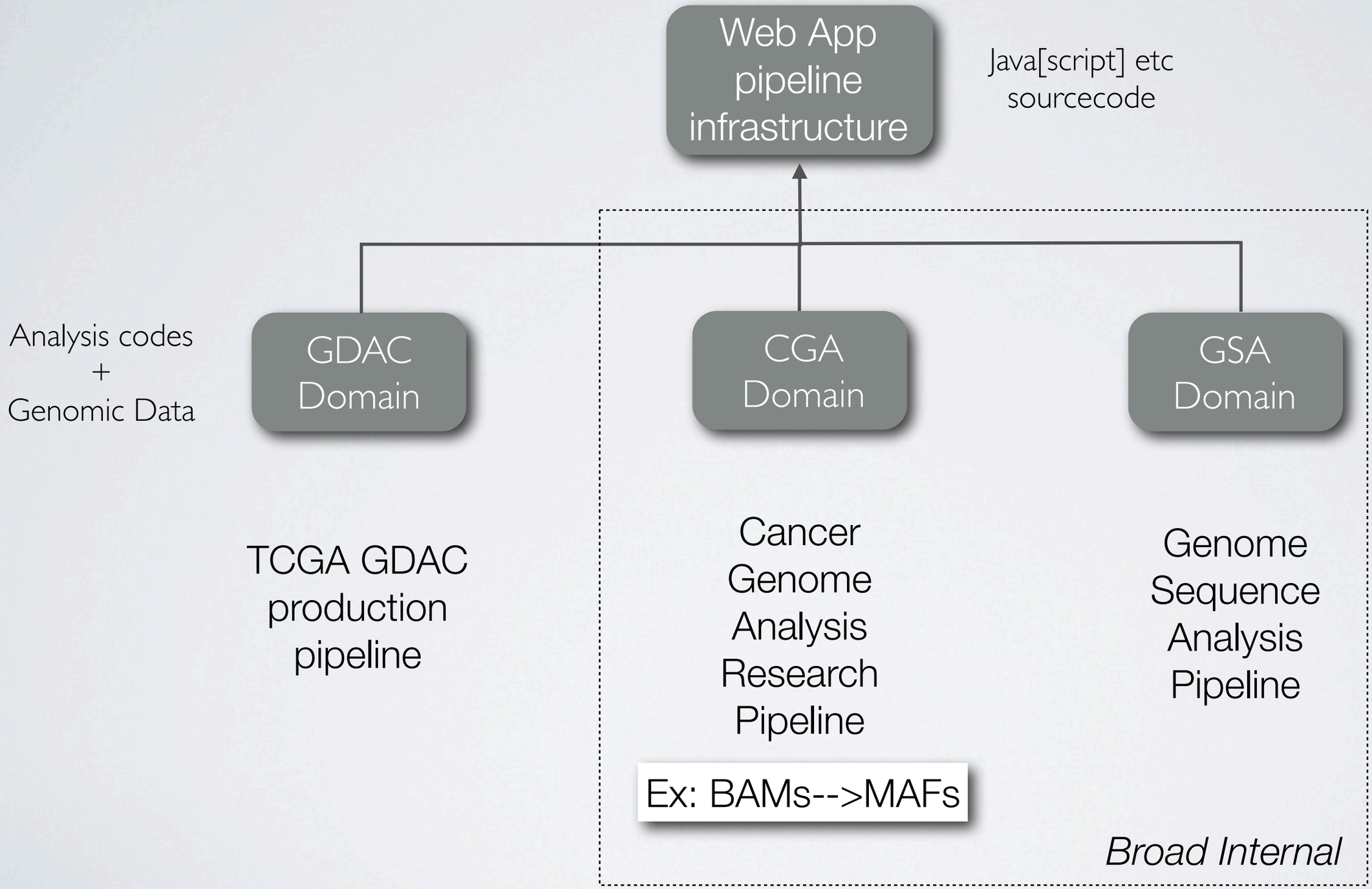
*Lists of participants and their affiliations appear at the end of the paper.

FIREHOSE
Broad GDAC

Operational 6 months

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

ASIDE: WHAT IS FIREHOSE?



II : Past

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

III : Present

APRIL 2011

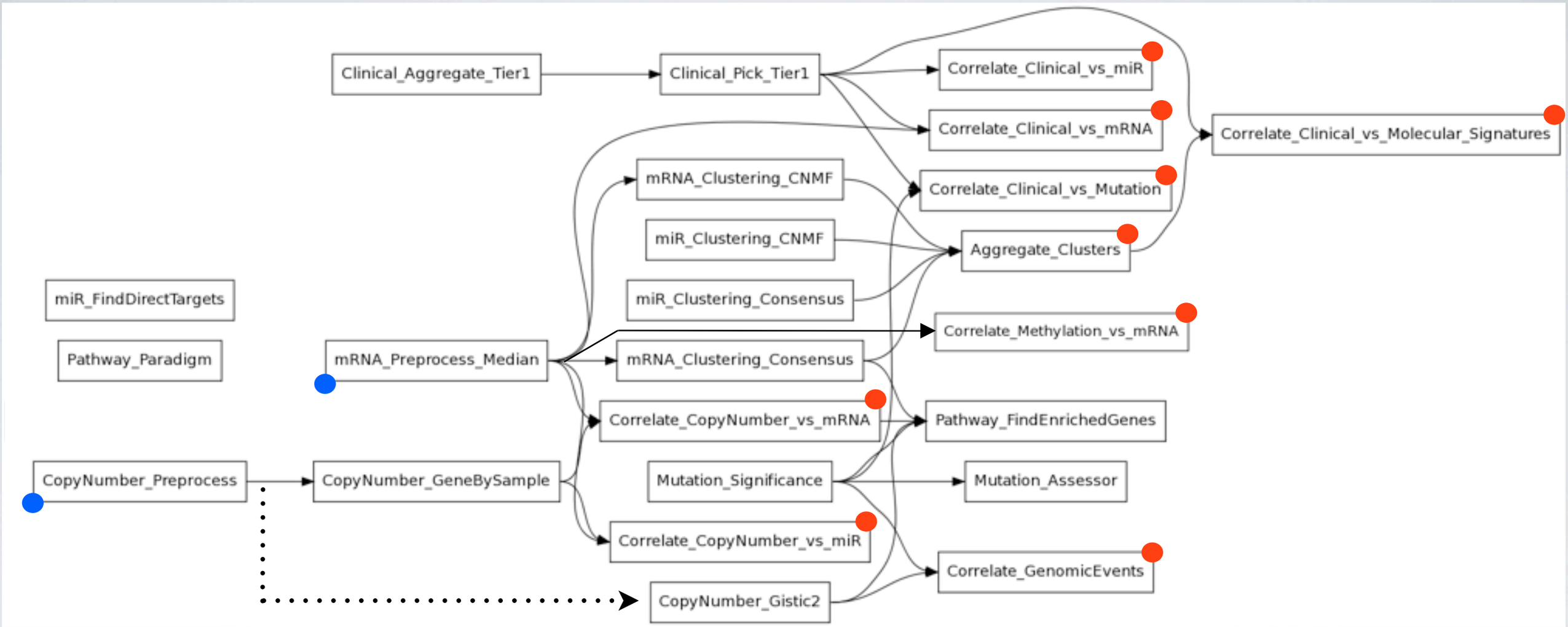
TumorType	Biospecimen	Any_Level_1	Clinical	CNA	Methylation	mRNA	miR	MAF
BLCA	26	12	9	9	0	0	0	0
BRCA	647	390	353	375	186	434	0	0
CESC	23	8	5	8	0	0	0	0
COAD	245	151	207	182	167	155	0	88
COADREAD	338	203	285	253	236	224	0	139
GBM	508	476	465	466	288	506	415	199
HNSC	59	59	0	57	0	0	0	0
KIRC	460	347	192	345	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

Operational Progress

Nov 2010	April 2011
Increase transparency	http://gdac.broadinstitute.org
Promote communication	gdac@broadinstitute.org
Improve automation	Firehose now programmable via growing web services
Improve clarity & rigor	Consistent pipeline nomenclature
Improve reports	Nozzle
Systematize DCC loopback	Auto SDRF packaging, upload
Grow staff to support operations	3 new SWEs, 2 bioinformaticists
Lower entry/maintenance barriers	Hydrant (in progress)

Analysis Workflow



24 pipelines X 21 tumor sets per run.

- Data Mediators : abstract platform details from algorithms
- Integrative Analyses : correlations across data types

IV : Future

So What?

ALL THIS HARD WORK IS POINTLESS ...

If we do not get ***uniform*** data & analyses ...

Into analyst & ultimately biologist hands ...

In ***timely*** fashion ...

And ***comprehensible*** form.

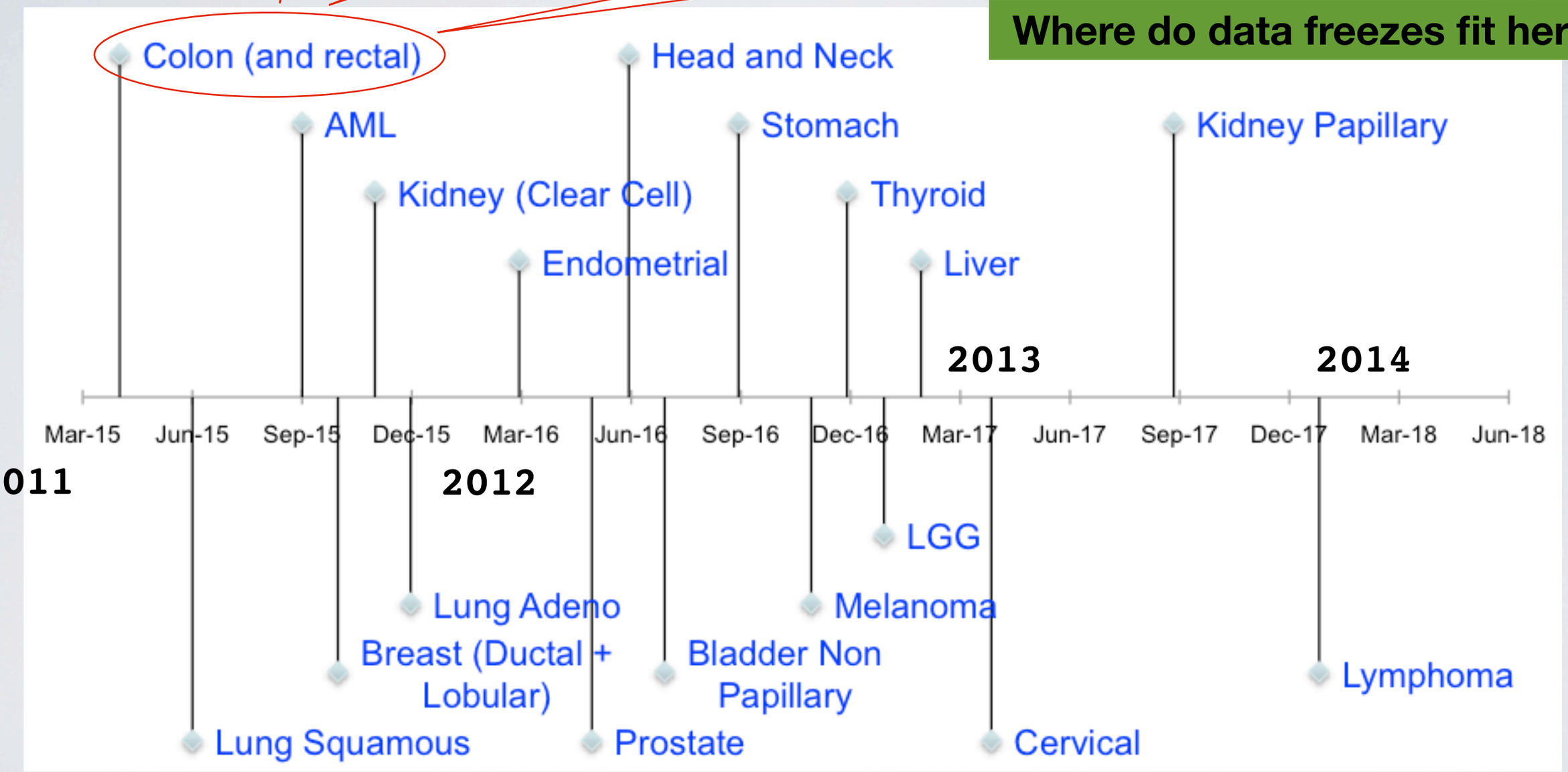
Hint: we're not there yet.

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

Participant Comment:
“We can’t do it this way for 19 more tumor types”

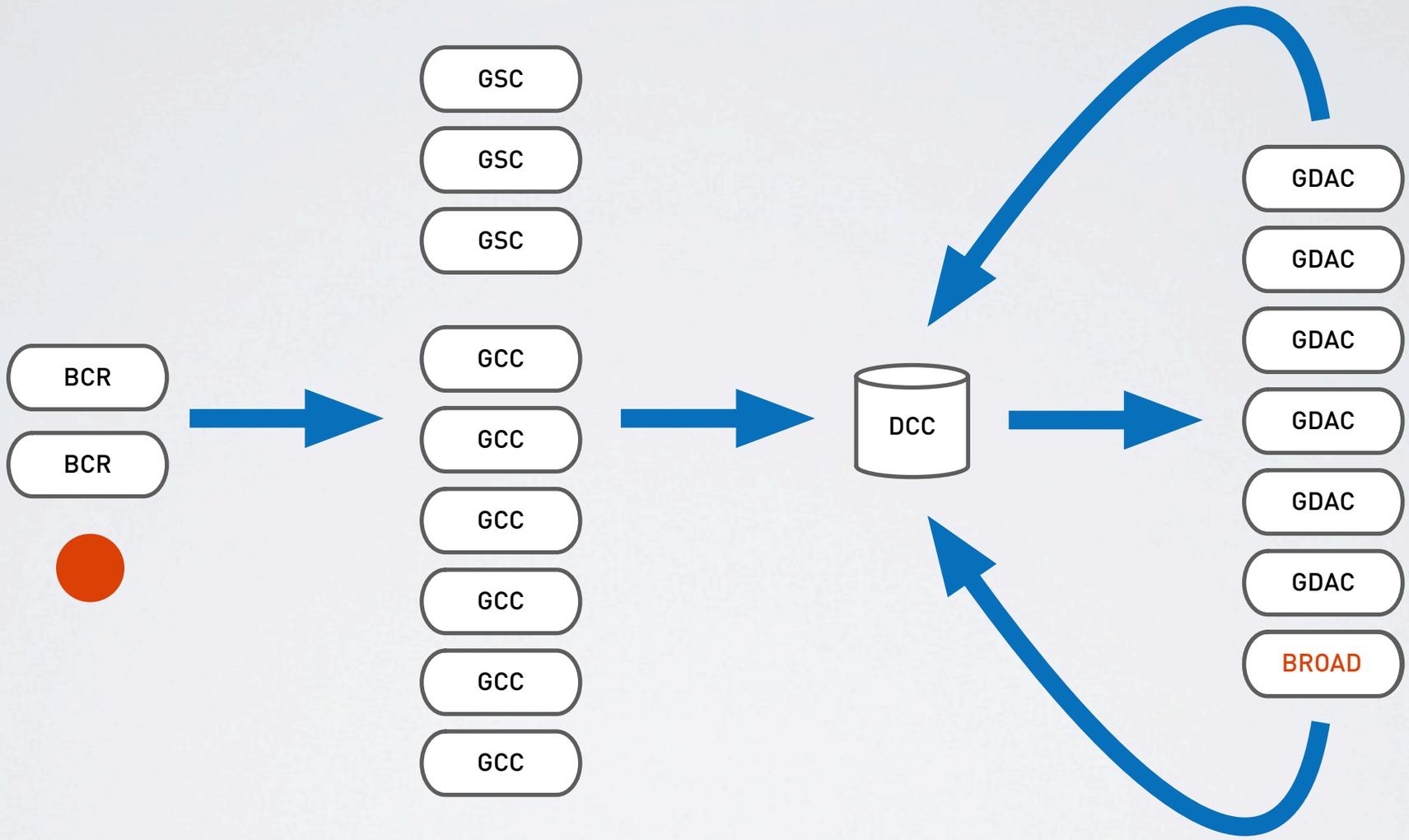
Firehose missed workshop by ~1 day …
Despite weekends & nights by several groups

Where do data freezes fit here?

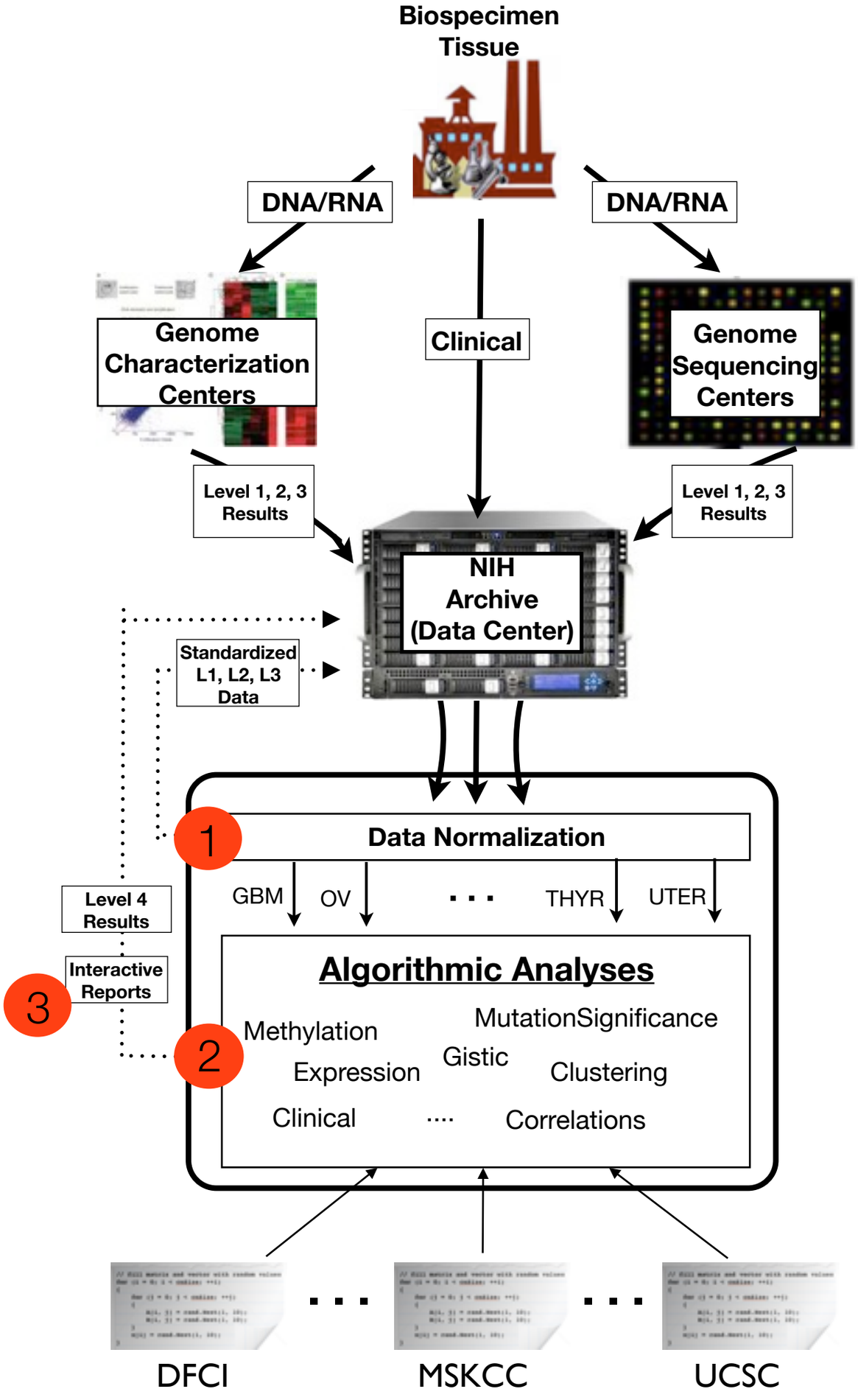


TCGA Phase II Tumor Projects Timeline

TCGA : Complex National-Scale Data Flow



The life cycle of a sample



More Detail

FIREHOSE TCGA GDAC

Aggregation & Standardization Point For TCGA Data & Analyses

1 Uniform

Core data & analyses should be standardized

Everyone agrees ...

... so, uh, why aren’t they?

1 Data Normalization

- 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 must , too.

1 Data Normalization

- Normed data is posted to DCC
- But Broad needs to make SDRFs
- And provide companion reports
- Likely by summer ...

(See Gordon Saksena Poster)

2

Timely

- Switching to multiple runs per month
- Default to bi-weekly
- But look for TOO: Targets Of Opportunity
- Such as manuscripts
- Or AWG workshops

2

Timely

- Predefined analyses: baselines for AWG work

"We can't do it this way for 19 more tumor types"

- Federated: DAG already demonstrated (cBIO)
- Loops: nice, but ...
- **DCC-served results** accessibility ...
- ... and **NGS/RNA-Seq** are **much** higher prios

3

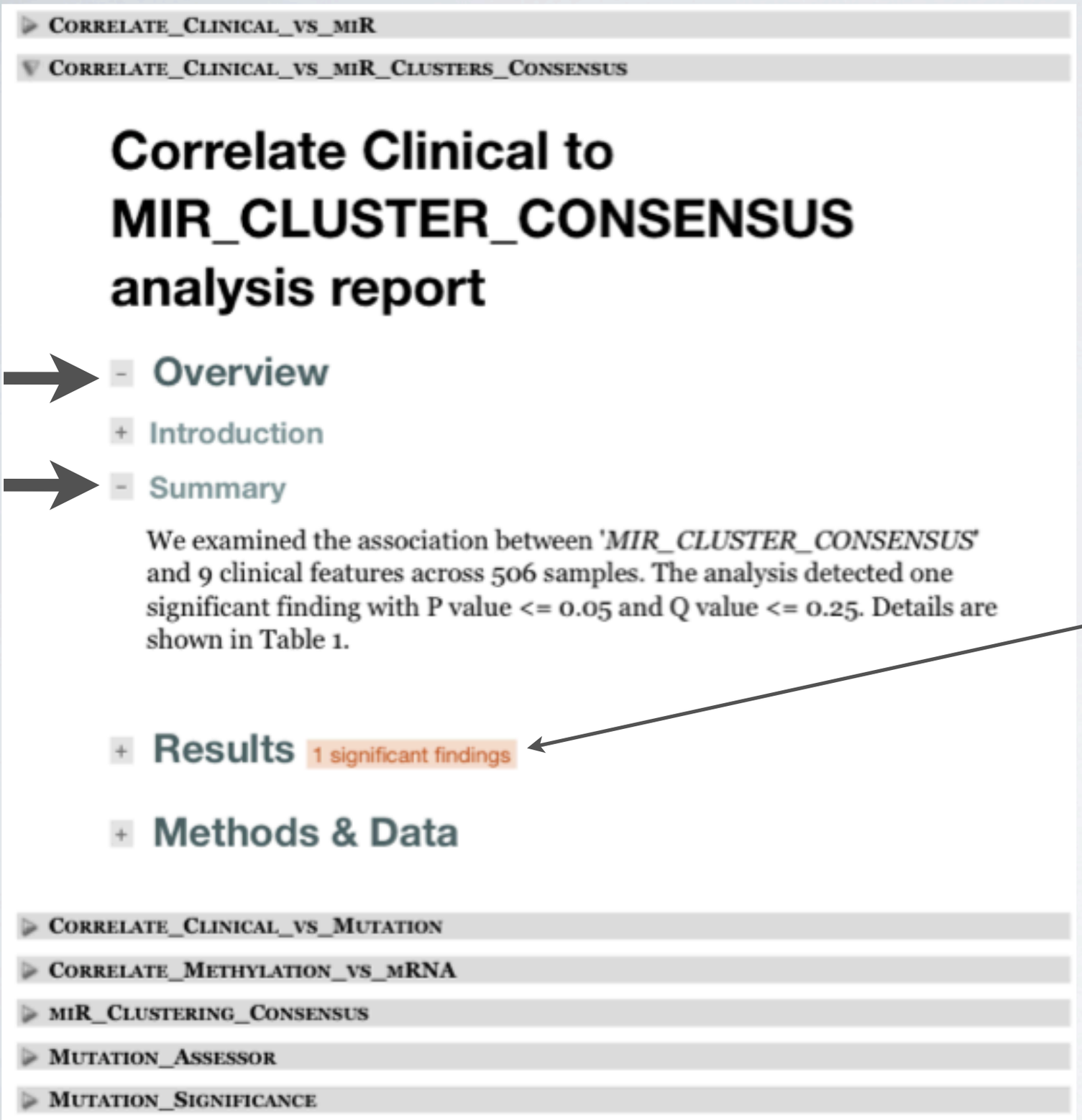
Comprehensible

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. Exposing **no knowledge** of technologies used to render (like HTML).

Producers focus on science content, not HTML syntax.

Nozzle : PAN-CANCER Dataset Example



- Standard visual format for ALL pipelines
- No HTML coding : simple R calls
- 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_VS_MIR

▼ CORRELATE_CLINICAL_VS_MIR_CLUSTERS_CONSENSUS

Correlate Clinical to MIR_CLUSTER_CONSENSUS analysis report

− Overview

+ Introduction

− Summary

We examined the association between '*MIR_CLUSTER_CONSENSUS*' and 9 clinical features across 506 samples. The analysis detected one significant finding with P value <= 0.05 and Q value <= 0.25. Details are shown in Table 1.

+ Results 1 significant findings

+ Methods & Data

▶ CORRELATE_CLINICAL_VS_MUTATION

▶ CORRELATE_METHYLATION_VS_MRNA

▶ MIR_CLUSTERING_CONSENSUS

▶ MUTATION_ASSESSOR

▶ MUTATION_SIGNIFICANCE

Interactivity: Drill Down To Significant Findings

View Summary Tables

Or Fully Expanded

Correlate Clinical to MIR_CLUSTER_CONSENSUS analysis report

Poor Man's First Draft Methods Paper

+ Overview

- Results

- Overview of the results_

GET FULL TABLE

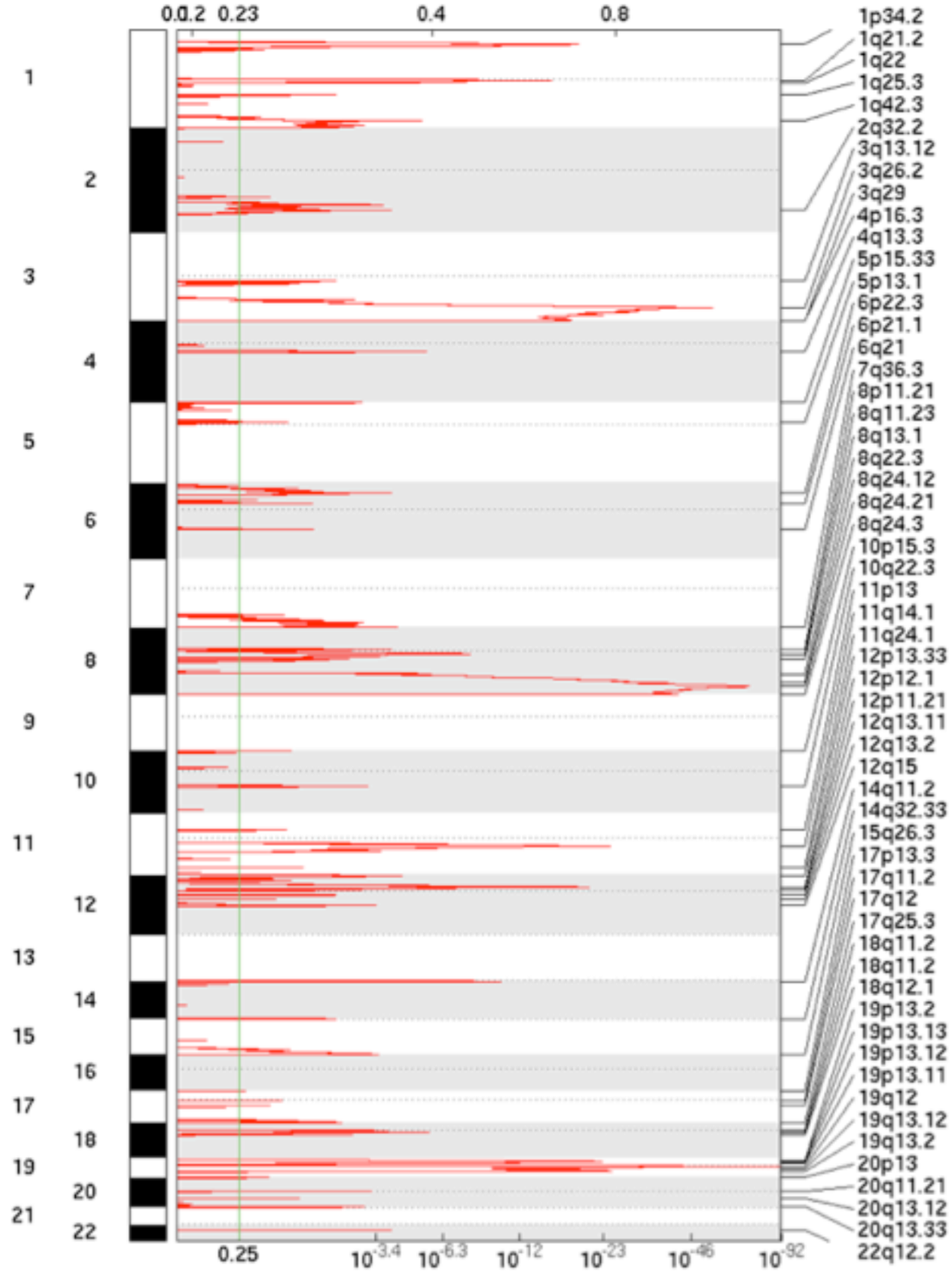
Table 1. Overview of the association results between 1 clustering variables and 9 clinical features. Shown in the table are P values (Q values). Thresholded by P value <= 0.05 and Q value <= 0.25, one significant finding detected.

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

See Nils Gehlenborg poster.

Figure 1. Genomic positions of amplified regions: the X axis represents the normalized amplification signals (top) and significance by q-value (bottom). The green line represents the significance cutoff at q-value=0.25



Thumbnail Figures

Zoom In Actively On Same Page

3

Comprehensible

- We must do better versioning
- Ex: what version of Gistic are you running?
  - gistic_version() = 2.01.<SVN_REVISION>
  - Relatively new, but in stdout log
- What about MutSig, MutationAssessor, ... ????

3 Comprehensible

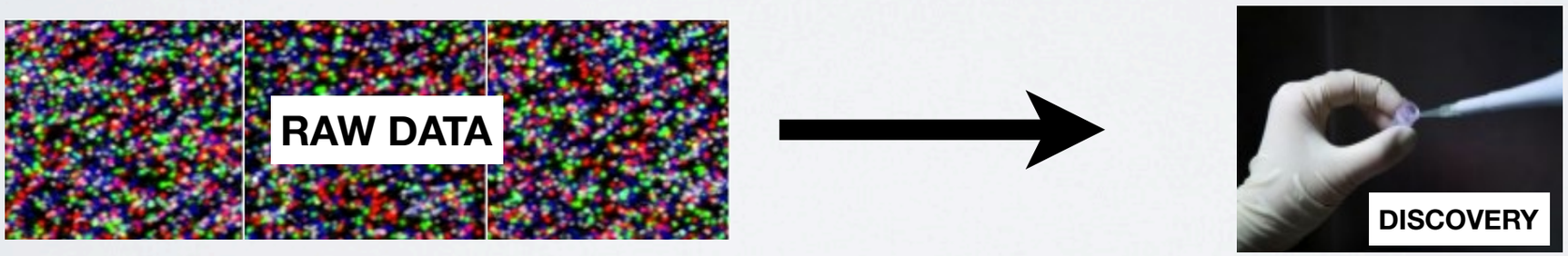
- Version in stdout log is **good**, but ...
- Visible in summary section of report **BETTER**
- All analysis modules should do same: most don’t
- Not done by “make install” : vacuous version changes
- But rather at code checkin time

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.

Take Aways

Data : still fragile, but we are bulletproofing

- Turn off clinical mirroring 2-3 days before run
- Introducing QC mechanism to perform daily clinical validation
- Volume: and we're not even dealing with RNA-seq yet!
- Quality: not possible to V-n-V 24 analyses x 21 tumor sets

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

Algorithms : yes, more == GOOD

- First finish normed data & results for **existing**?
- Hydrant will simplify integration process considerably
- Introducing **Integration workspace** for production stability
- And more clearly discernible analysis versioning

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

Acknowlegements

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

Belfler-DFCI
Yonghong Xiao
Juinhua Zhang
Spring Liu
Sachet Shukla
Hailei Zhang
Terrence Wu

Harvard
Peter Park
Nils Gehlenborg
Semin Lee
Richard Park

IGV & GenePattern teams @ Broad
Jill Mesirov
Michael Reich
Peter Carr
Marc-Danie Nazaire
Jim Robinson
Helga Thorvaldsdottir

Matthew Meyerson
Todd Golub
Eric Lander




The End!