THE CANCER GENOME ATLAS

BROAD INSTITUTE

The Broad GDAC Pipeline

Michael S. Noble
GDAC Pipeline Manager

On Behalf of Broad/DFCI GDAC Team :

Lynda Chin, PI
Gaddy Getz, PI
Peter Park
Douglas Voet

Gordon Saksena
Kristian Cibulskis
Rui Jing
Michael Lawrence

Andrey Sivachenko
Carrie Sougnez
John Zhang
Yinghong Xiao

Spring Liu
Hailei Zhang
Sachet Shukla
Terrance Wu

Lihua Zou
Richard Park
Peter Carr
Marc Danie Nazaire

Outline

I. Purpose

II. Flow

III. Data

IV. Analyses

V. Firehose

VI. Examples

VII. Future

Aside: Since You Don't Know Me

- Computational Scientist
- 3 months in cancer genomic analysis @ Broad
- Last 14 Years in astrophysics @ Harvard & MIT
- Managing/developing pipeline & analysis infrastructure
- And publication research/SW for spectral analysis
- For Chandra X-Ray Observatory (11 years in flight)
- Research interests in parallel computing, spectral modeling, data analysis & vizualization, automated code generation, modular/scriptable numerical SW

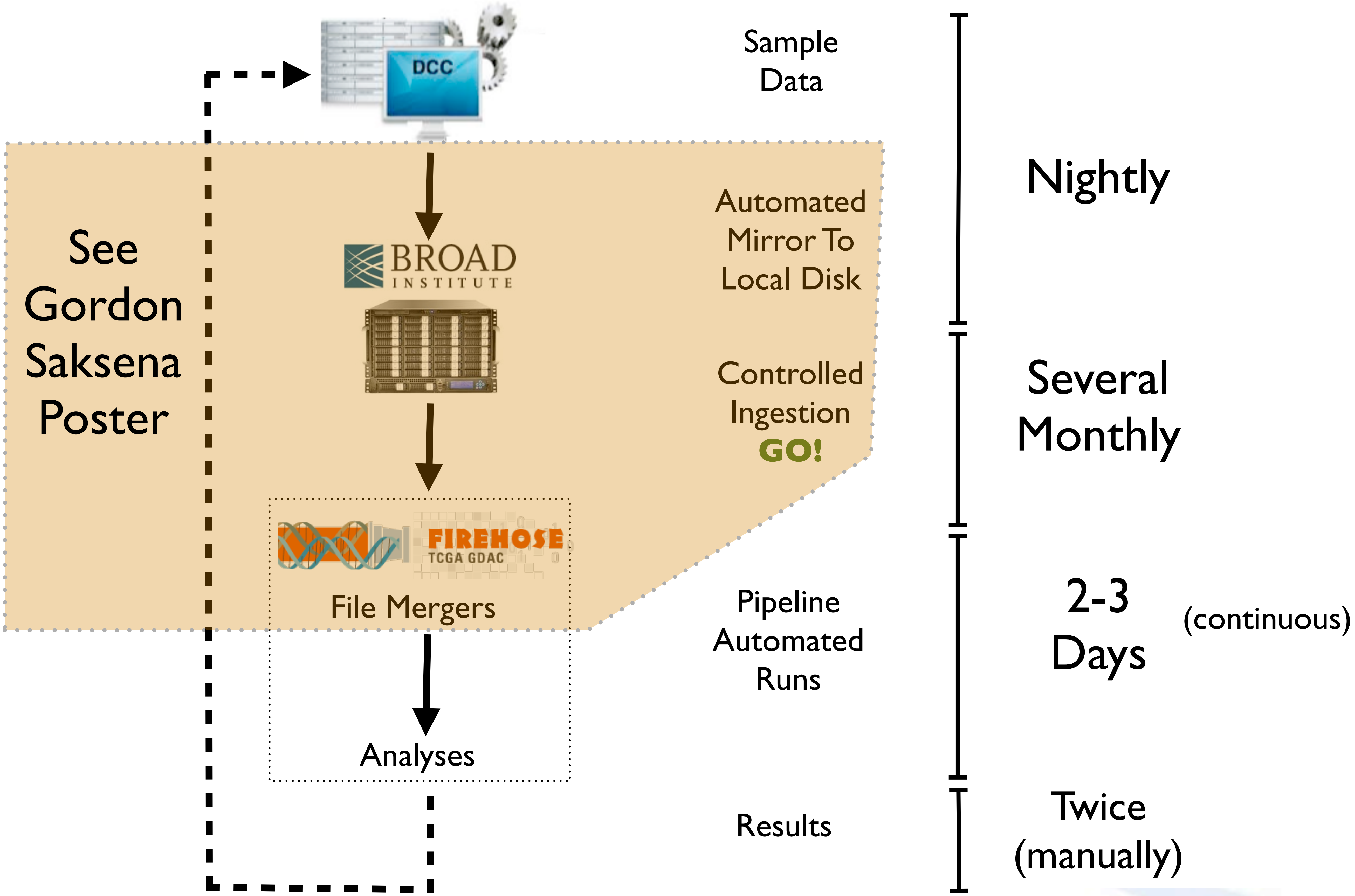
I. Purpose

Coordinate the flow of massive, terabyte-scale genomic datasets through scores of quantitative algorithms.

With the aims of automation, high throughput, and nearly turnkey reproducibility.

While facilitating research & discovery.

II. Flow



See Gordon Saksena Poster

File Mergers

Analyses

Sample Data

Automated Mirror To Local Disk

Controlled Ingestion **GO!**

Pipeline Automated Runs

Results

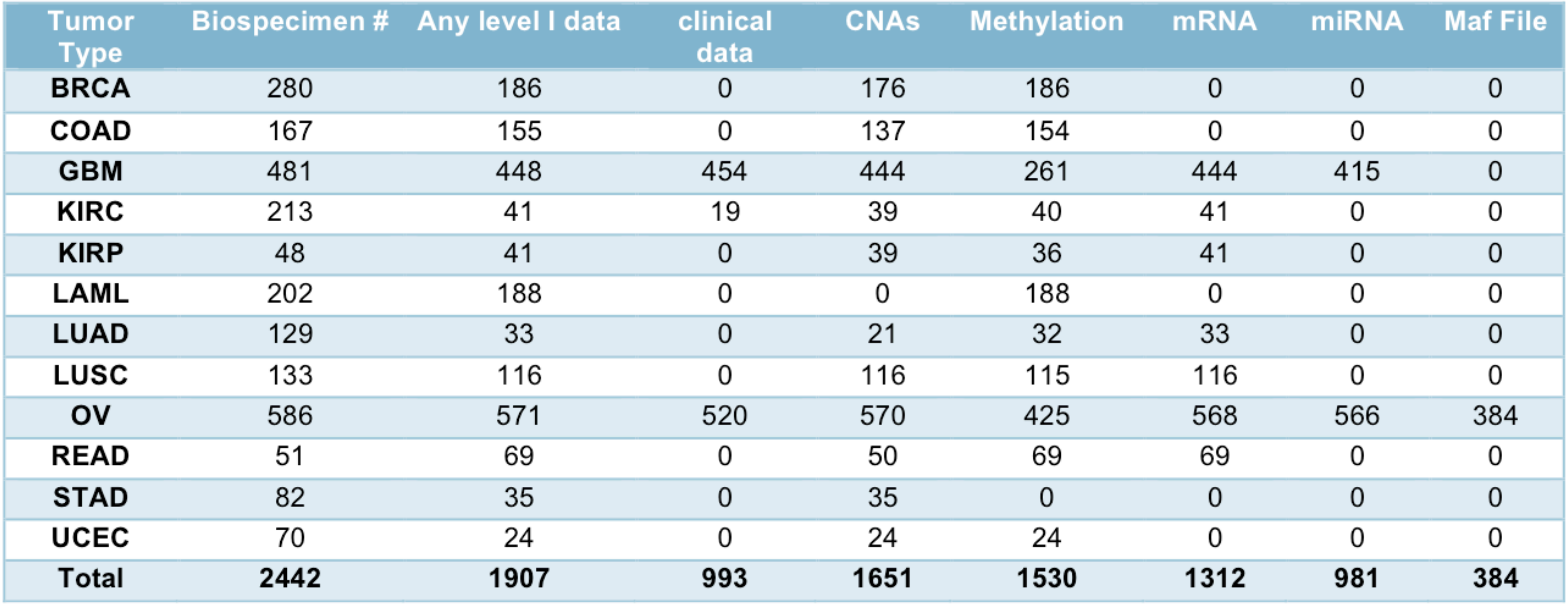
Nightly

Several Monthly

2-3 Days (continuous)

Twice (manually)

III. Data



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**

November 5 Analysis Run : 12 tumor types

www.broadinstitute.org/~gdac/TumorDataSummary.png

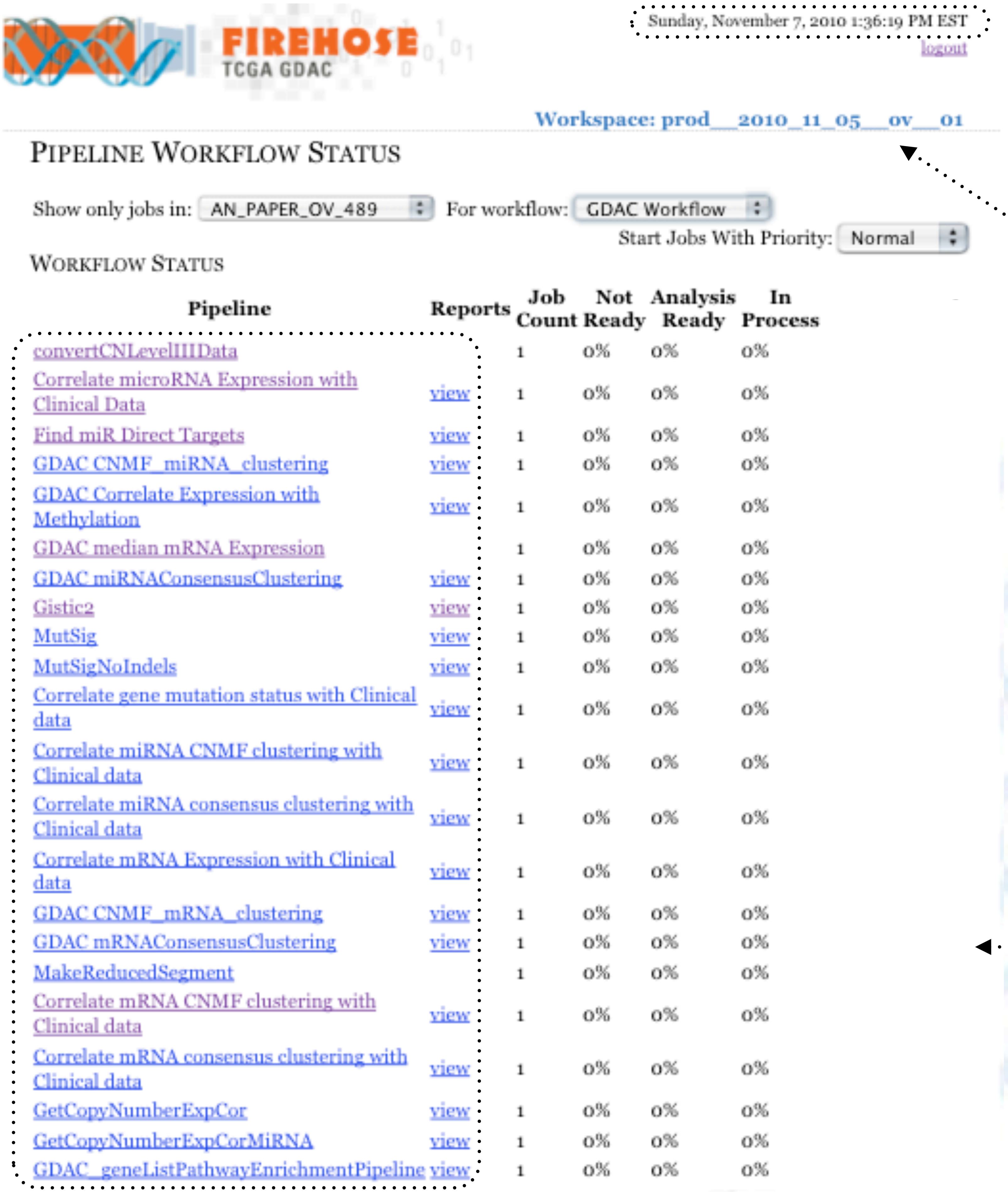
- Daily auto-mirror DCC → Broad local disk
- ***Partition:*** to one sample per file (part of normalization)
- Daily ingestion into FireHose DEV & PROD workspaces
- Controlled ingestion into production analysis: press GO
- Date-stamped workspaces created: inherit from PROD
- Currently 13 per run: one per tumor + "ALL"

  prod_2010_11_05_ov_01 ← Pass 1 : DNU list applied
  prod_2010_11_05_gbm_00
  prod_2010_11_05_lusc_00 ← Pass 0 : full individual set
  ...

- ***Selection:*** filtered (by DNU list) samples merged ...
- Into files whose names seed L3 input ***annotations***

See Gordon Saksena Poster

IV. Analyses



Sunday, November 7, 2010 1:36:19 PM EST

Workspace: prod__2010_11_05__ov__01

PIPELINE WORKFLOW STATUS

Show only jobs in: AN_PAPER_OV_489 For workflow: GDAC Workflow

Start Jobs With Priority: Normal

WORKFLOW STATUS

Pipeline	Reports	Job Count	Not Ready	Analysis Ready	In Process
convertCNLevelIIIData		1	0%	0%	0%
Correlate microRNA Expression with Clinical Data	view	1	0%	0%	0%
Find miR Direct Targets	view	1	0%	0%	0%
GDAC CNMF_miRNA_clustering	view	1	0%	0%	0%
GDAC Correlate Expression with Methylation	view	1	0%	0%	0%
GDAC median mRNA Expression		1	0%	0%	0%
GDAC miRNAConsensusClustering	view	1	0%	0%	0%
Gistic2	view	1	0%	0%	0%
MutSig	view	1	0%	0%	0%
MutSigNoIndels	view	1	0%	0%	0%
Correlate gene mutation status with Clinical data	view	1	0%	0%	0%
Correlate miRNA CNMF clustering with Clinical data	view	1	0%	0%	0%
Correlate miRNA consensus clustering with Clinical data	view	1	0%	0%	0%
Correlate mRNA Expression with Clinical data	view	1	0%	0%	0%
GDAC CNMF_mRNA_clustering	view	1	0%	0%	0%
GDAC mRNAConsensusClustering	view	1	0%	0%	0%
MakeReducedSegment		1	0%	0%	0%
Correlate mRNA CNMF clustering with Clinical data	view	1	0%	0%	0%
Correlate mRNA consensus clustering with Clinical data	view	1	0%	0%	0%
GetCopyNumberExpCor	view	1	0%	0%	0%
GetCopyNumberExpCorMiRNA	view	1	0%	0%	0%
GDAC_geneListPathwayEnrichmentPipeline	view	1	0%	0%	0%

22 Pipelines

Nov 5th Run

Ovarian

48 Hours: Ingest to Completion

100% Success

(Uploaded to DCC)

- Constitutes signficant portion of OV manuscript
- Regenerated in days, automatically
- With <u>novel results</u> included in resubmitted OV ms!
- Contrast to manual effort, by teams over months.
- Results + reports uploaded to DCC
- Manually, but automated + SDRF in works

- 23 analysis pipelines currently installed, including:
  - MutSig (with and w/out indels)
  - Gistic2
  - 7 clinical correlations
  - methylation VS expression
  - gene list pathway enrichment
  - multiple clusterings
  - PARADIGM (lite)

External Module
Benz/Vaske
UCSC

- Covering all data types:
  - mRNA, miRNA Expression
  - methylation
  - copy number
  - mutation
  - clinical

- Collected into automated workflow
- Run against each of 12 tumor types with extant data
- Majority have reports

- Basic pre-defined integrative analyses

- 2 data types in most cases

- Include single data type analysis (level IV) when required for integrative analysis

- Intermediate data files available for use in algorithms at other centers

Example

Gene-centric summary table output from PL (one per datatype) can be fed to Oncoprint at MSKCC's cBio pathway portal

In The Queue

PARADIGM Lite	Sam NG, UCSC	Integrated; need reports for runs
NetBox	Cerami et al MSKCC	TBD
ICluster	ditto	TBD
RNA-Seq	A. Sivachenko Broad	TBD
Co-occurrence mutual exclusivity	J. Weinstein M.D. Anderson	TBD

Putting New Codes In

- Source code not private (published/open/available)
- Tested on TCGA data, preferably multiple tumors
- Runnable from Unix
- Drivable by command line args
- Meaning essentially any language is OK, even proprietary runtimes (but only MatLab so far)
- Library ok, but need executable wrapper
- Then contact us

Autonomy would be ideal: you put your codes in yourself
Security / authentication / IRB privacy issues
Under active consideration

V. FIREHOSE TCGA GDAC

- Version control for computational experiments
- On steroids :

  Capable of generating 100K reproducible
  LSF jobs in just seconds

- Portable: Java implementation, browser UI
- ROBUST research tool: used daily by scores
- Evolving to GDAC production use case

- Currently runs only @ Broad (protected data)

- VPN for off-site access:

Daily use by DFCI contributors

Daily use by Broadies@Home

- We proposed to build a TCGA-wide VPN to run Firehose, allowing entire TCGA to directly install tools and interact with results

--> This was not funded.

Bad Old Days : Manual Experiment Mgmt

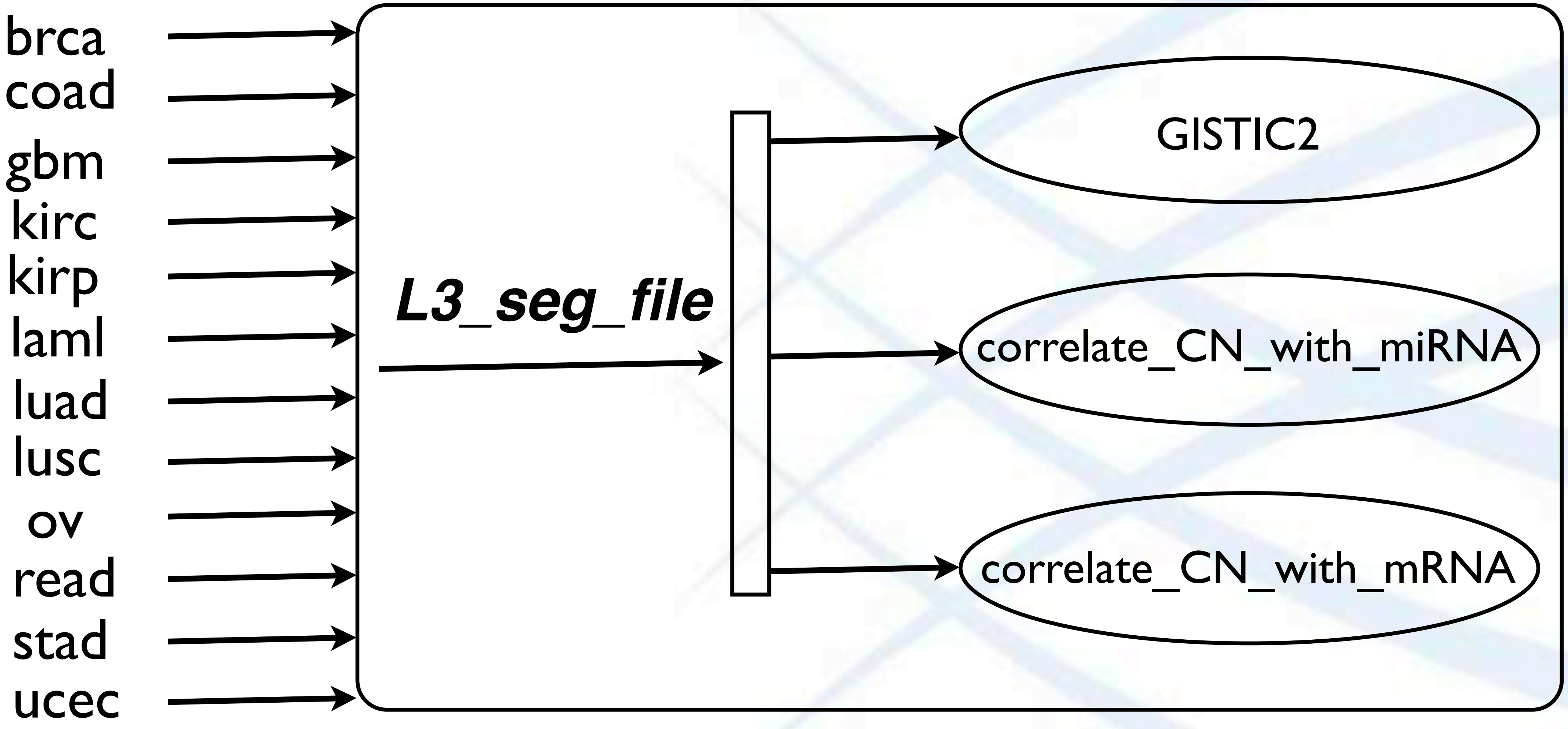
```
% mkdir  my_GISTIC_run_Nov_10_OV

% cd

% ftp JAMBOREE.nih.gov

% tar xvjf ...

% run
```

Then do it again Nov 15, 19, ...
Then forget ... and search, search, search
Then repeat everything for
GBM, LUSC, LAML, ...

Then multiply by 10, 20, 30 researchers ...

Enter FireHose Annotations

- Logical identifier for datum: input or output
- Abstracts file system knowledge from algorithms
- Transparent multiplexing across TCGA tumor types



- All “for free” once algorithm in FIREHOSE TCGA GDAC

- One learns to care less about directories ...

- And LSF parallel job dispatching, etc ...

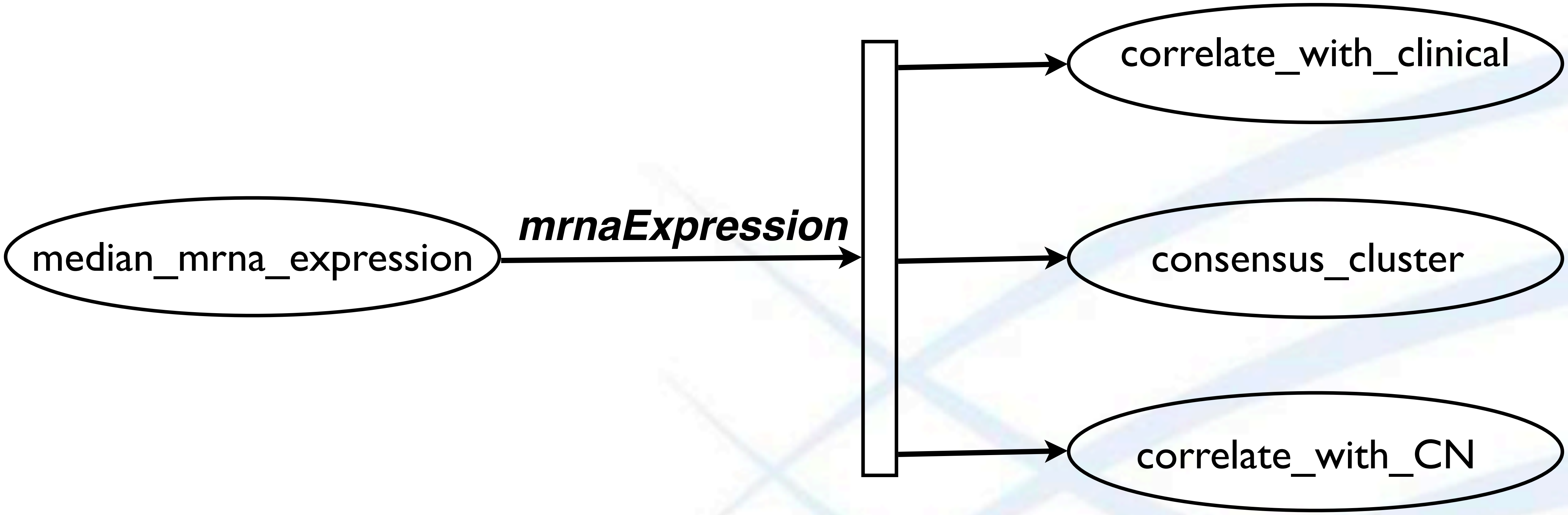
- FH manages the nuisance details

- Still challenging, even on dedicated infrastructure with hundreds/thousands of nodes at our disposal

- Can be distributed (DCC connection is,)

- But devil-in-details work remains for runs across institutional boundaries, esp in compliance with privacy requirements

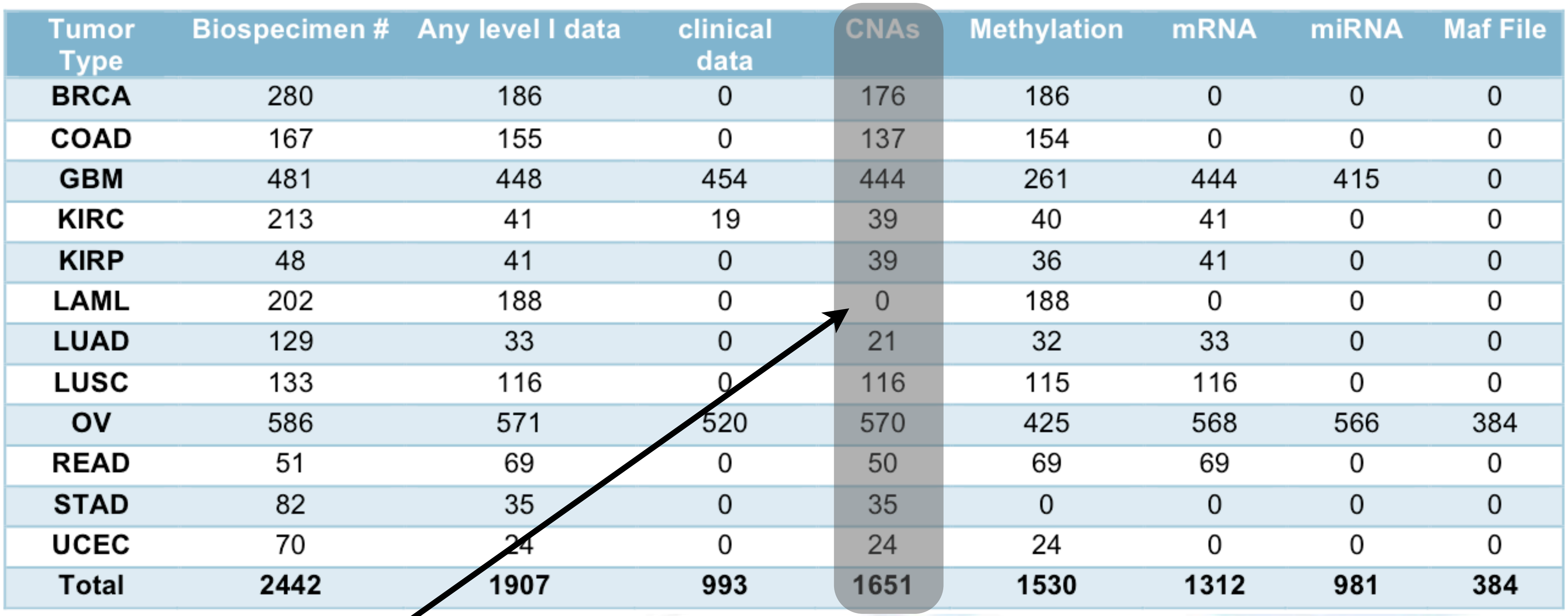
Also elegantly enforces workflow DAG constraints



FH will not run latter 3 modules until ***mrnaExpression*** annotation populated with value from first module

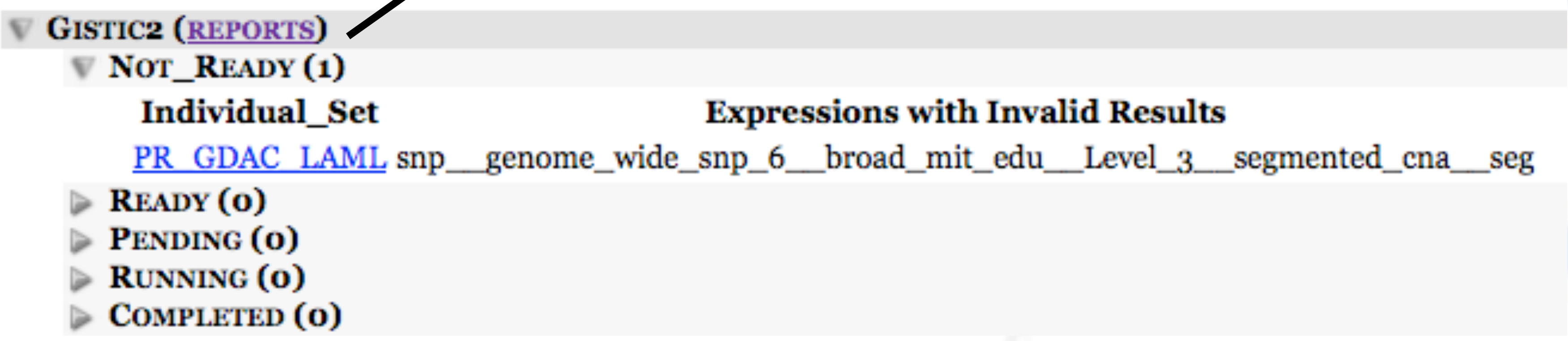
VI. Example

- Gistic2 attempted for all 12 tumor types in 11/5



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**

Plenty of CNA data



Gistic2 (REPORTS)

Not_Ready (1)

Individual_Set	Expressions with Invalid Results
PR_GDAC_LAML	snp__genome_wide_snp_6__broad_mit_edu__Level_3__segmented_cna__seg

Ready (0)

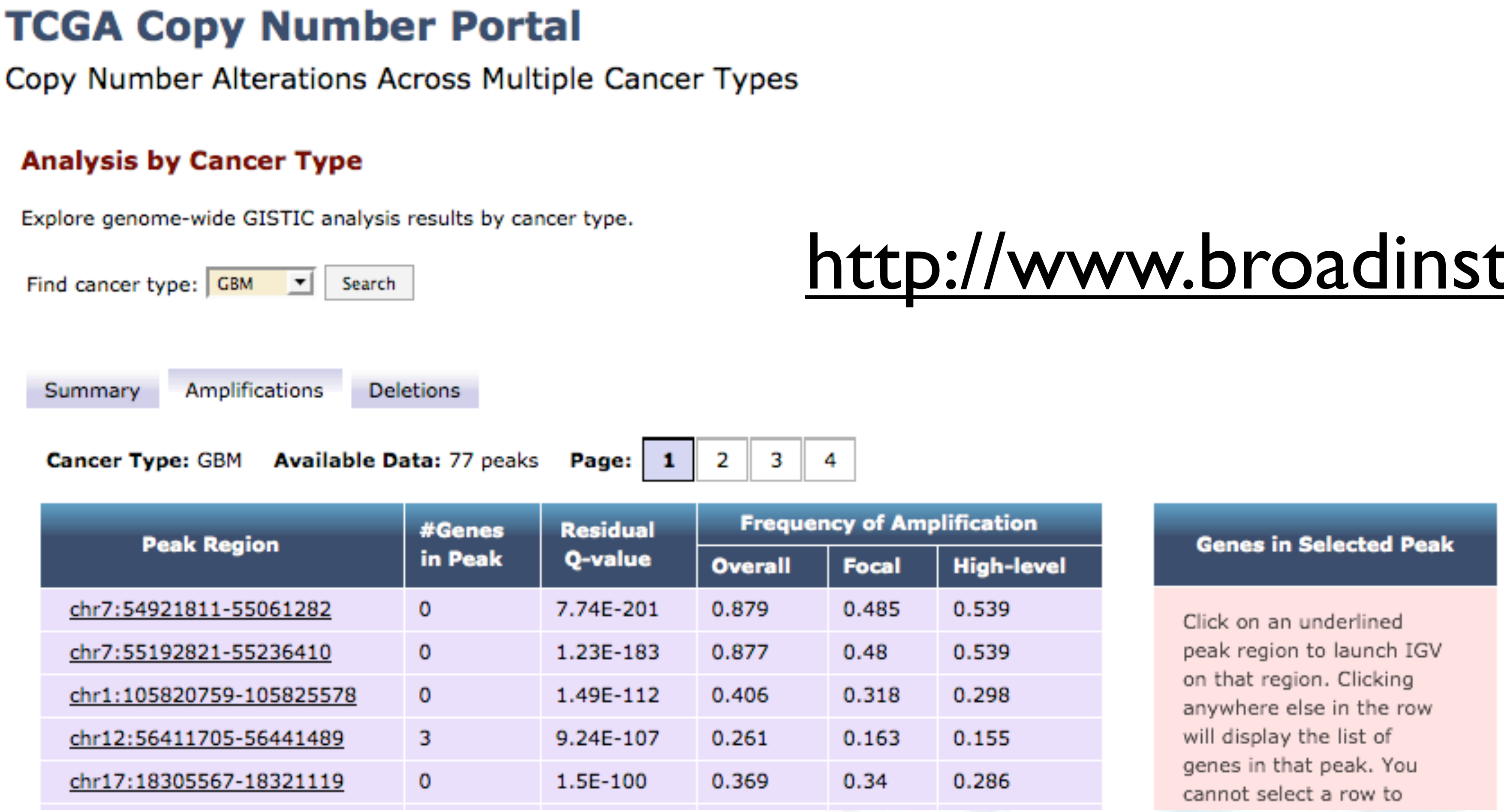
Pending (0)

Running (0)

Completed (0)

Only LAML did not run

- Results for 4 tumor types (OV, GBM, Breast, Colon) then injected into Tumorscape portal



TCGA Copy Number Portal

Copy Number Alterations Across Multiple Cancer Types

Analysis by Cancer Type

Explore genome-wide GISTIC analysis results by cancer type.

Find cancer type: GBM ▾ Search

Cancer Type: GBM **Available Data:** 77 peaks **Page:** 1 2 3 4

Peak Region	#Genes in Peak	Residual Q-value	Frequency of Amplification		
			Overall	Focal	High-level
chr7:54921811-55061282	0	7.74E-201	0.879	0.485	0.539
chr7:55192821-55236410	0	1.23E-183	0.877	0.48	0.539
chr1:105820759-105825578	0	1.49E-112	0.406	0.318	0.298
chr12:56411705-56441489	3	9.24E-107	0.261	0.163	0.155
chr17:18305567-18321119	0	1.5E-100	0.369	0.34	0.286

Genes in Selected Peak

Click on an underlined peak region to launch IGV on that region. Clicking anywhere else in the row will display the list of genes in that peak. You cannot select a row to

http://www.broadinstitute.org/tcga

- See Andrew Cherniack poster for details

- With thanks To: Steve Schumacher, Reid Pinchback, Rameen Beoukhim, Matthew Meyerson

Gistic Pipeline Assessment

- Overall results look very similar to OV manuscript, including all of our reported findings

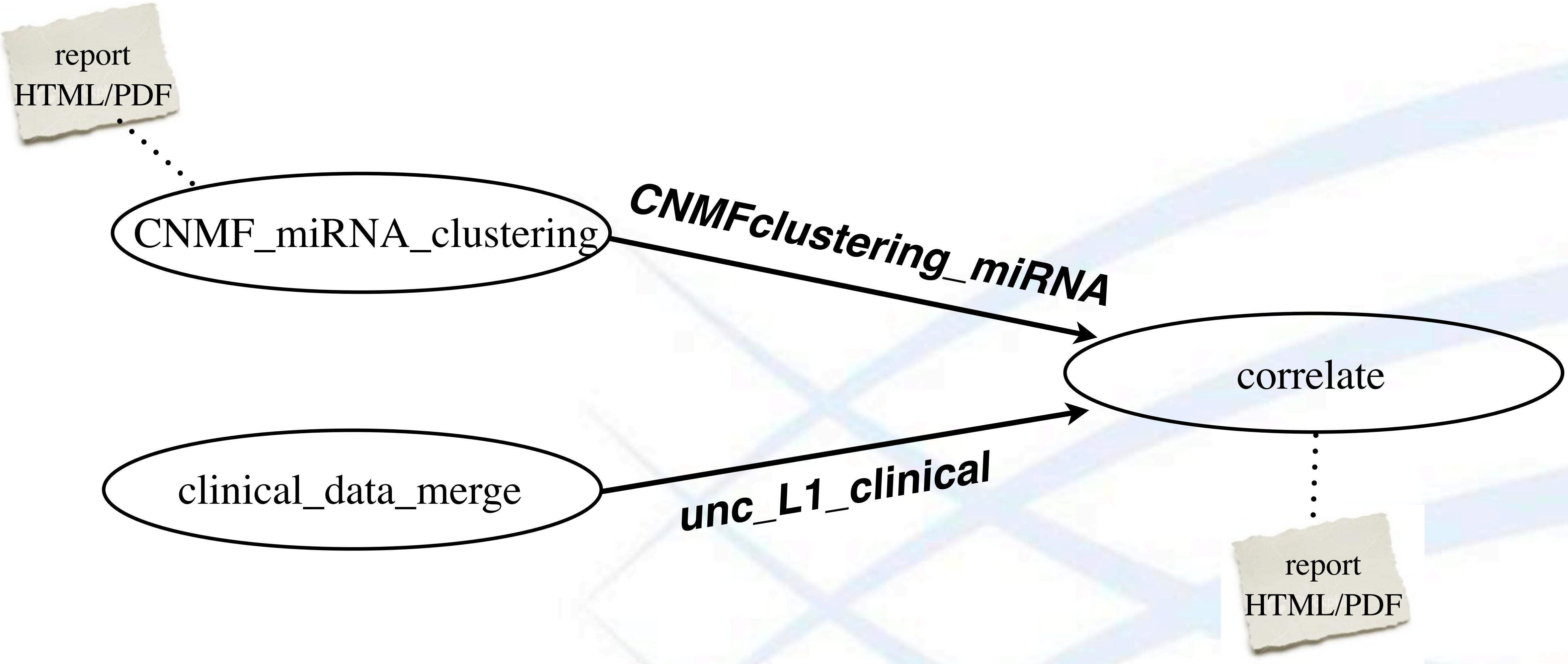
- Diffs due to pipeline using SNP 6.0 data (482 samples) with CNV list not yet reflecting that

- VERSUS manuscript using Agilent data (489 samples) filtered against MSKCC/Agilent CNVs

- Focusing on SNP6.0 to accomodate future tumor types in which there are only SNP6.0 data

See Gaddy for more details.

Example II : Integrative Analysis



- Reports bundled with outputs: HTML default, PDF optional
- Summary Format: still need work, but converging
- All packaged & uploaded to DCC from 11/5 OV run

See For Yourself

Inside FireHose

Live without a net
But with a network

miRNA CNMF clustering

correlated with clinical

View on laptop

Meek & timid
with no network

miRNA CNMF clustering

correlated with clinical

(switch to browser)

caftps.nci.nih.gov /users/gdacbroad

VII. Future

Task not simple ⟶ not supposed to be easy

- Datasets are gigantic and algorithms evolving
- Privacy necessary but burdensome constraint
- But significant progress demonstrated
- The beast of complexity being tamed ...

- Powerful system in place
- With strong conceptual foundation
- Producing tangible results
- Easily chew up 100 TB in few weeks

Forward March

- Increase transparency

Public Dashboard

Tumor	Samples	Pipeline	Status
gbm	454	xyz	fail
ov	520	abc	pass

- Continue to widen usage & lower entry barriers
- Continue adaptation to GDAC production use case

Rigorous pipeline/annotation nomenclature
No hacks to accommodate missing or ill-formatted data
Improve reports
Automatic SDRF-based upload to DCC

- Continue improving automation: scriptable control
- Continue fruitful interaction beyond our walls

Growing staff now more able to translate discussion to actions

- Hope barcode --> UUID not dark storm on horizon?

Cancer now on borrowed time ...
days are numbered.

Thank You!