



# How To QC A KiloPipeline Per Month?

Michael S. Noble  
The Broad Institute of MIT & Harvard

TCGA Steering Committee Meeting  
Houston, Texas

April 26, 2012



Born of the desire to systematize analyses from The Cancer Genome Atlas pilot and scale their execution to the dozens of remaining diseases to be studied, now sits atop 14 terabytes of TCGA data and reliably executes more than 1000 pipelines per month.

# Data Snapshot

## 2012\_04\_12 stddata Run

Tumor	BCR	Clinical	CN	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BLCA	89	65	58	78	0	32	0	54	0	28
BRCA	859	857	833	858	529	751	0	781	408	507
CESC	110	25	68	0	0	0	0	8	0	36
COADREAD	590	590	575	584	224	83	0	255	399	224
DLBC	27	0	0	0	0	0	0	0	0	0
GBM	595	563	546	287	542	0	491	0	214	276
HNSC	294	255	165	292	0	103	0	89	0	0
KIRC	502	502	490	500	72	469	0	463	454	327
KIRP	135	84	75	117	16	14	0	16	0	0
LAML	202	200	0	192	0	179	0	187	0	199
LGG	144	140	143	0	27	0	0	30	0	0
LIHC	84	55	58	0	0	17	0	28	0	0
LNNH	2	0	0	0	0	0	0	0	0	0
LUAD	372	274	266	347	32	106	0	95	0	147
LUSC	290	272	282	282	154	220	0	202	0	178
OV	592	580	564	551	568	0	564	46	412	316
PAAD	48	0	14	30	0	0	0	0	0	0
PRAD	153	0	100	153	0	0	0	63	0	0
SKCM	253	0	219	240	0	0	0	0	0	0
STAD	162	150	132	133	0	57	0	123	0	133
THCA	274	73	228	230	0	0	0	45	0	0
UCEC	462	425	430	451	54	266	0	359	200	239
PANCANCER	6239	5110	5246	5325	2218	2297	1055	2844	2087	2610

# Data Snapshot

## 2012\_04\_12 stddata Run

Tumor	BCR	Clinical	CN	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BLCA	89	65	58	78	0	32	0	54	0	28
BRCA	859	857	833	858	529	751	0	781	408	507
CESC	110	25	68	0	0	0	0	8	0	36
COADREAD	590	590	575	584	224	83	0	255	399	224
DLBC	27	0	0	0	0	0	0	0	0	0
GBM	595	563	546	287	542	0	491	0	214	276
HNSC	294	255	165	292	0	103	0	89	0	0
KIRC	502	502	490	500	72	469	0	463	454	327
KIRP	135	84	75	117	16	14	0	16	0	0
LAML	202	200	0	192	0	179	0	187	0	199
LGG	144	140	143	0	27	0	0	30	0	0
LIHC	84	55	58	0	0	17	0	28	0	0
LNNH	2	0	0	0	0	0	0	0	0	0
LUAD	372	274	266	347	32	106	0	95	0	147
LUSC	290	272	282	282	154	220	0	202	0	178
OV	592	580	564	551	568	0	564	46	412	316
PAAD	48	0	14	30	0	0	0	0	0	0
PRAD	153	0	100	153	0	0	0	63	0	0
SKCM	253	0	219	240	0	0	0	0	0	0
STAD	162	150	132	133	0	57	0	123	0	133
THCA	274	73	228	230	0	0	0	45	0	0
UCEC	462	425	430	451	54	266	0	359	200	239
PANCANCER	6239	5110	5246	5325	2218	2297	1055	2844	2087	2610

+821  
CopyNumber

# Data Snapshot

2012\_04\_12 stddata Run

Tumor	BCR	Clinical	CN	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BLCA	89	65	58	78	0	32	0	54	0	28
BRCA	859	857	833	858	529	751	0	781	408	507
CESC	110	25	68	0	0	0	0	8	0	36
COADREAD	590	590	575	584	224	83	0	255	399	224
DLBC	27	0	0	0	0	0	0	0	0	0
GBM	595	563	546	287	542	0	491	0	214	276
HNSC	294	255	165	292	0	103	0	89	0	0
KIRC	502	502	490	500	72	469	0	463	454	327
KIRP	135	84	75	117	16	14	0	16	0	0
LAML	202	200	0	192	0	179	0	187	0	199
LGG	144	140	143	0	27	0	0	30	0	0
LIHC	84	55	58	0	0	17	0	28	0	0
LNNH	2	0	0	0	0	0	0	0	0	0
LUAD	372	274	266	347	32	106	0	95	0	147
LUSC	290	272	282	282	154	220	0	202	0	178
OV	592	580	564	551	568	0	564	46	412	316
PAAD	48	0	14	30	0	0	0	0	0	0
PRAD	153	0	100	153	0	0	0	63	0	0
SKCM	253	0	219	240	0	0	0	0	0	0
STAD	162	150	132	133	0	57	0	123	0	133
THCA	274	73	228	230	0	0	0	45	0	0
UCEC	462	425	430	451	54	266	0	359	200	239
PANCANCER	6239	5110	5246	5325	2218	2297	1055	2844	2087	2610

+821  
CopyNumber

+917  
Methylation



# Data Snapshot

New datatype column  
+2087 protein samples

2012\_04\_12 stddata Run

Tumor	BCR	Clinical	CN	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BLCA	89	65	58	78	0	32	0	54	0	28
BRCA	859	857	833	858	529	751	0	781	408	507
CESC	110	25	68	0	0	0	0	8	0	36
COADREAD	590	590	575	584	224	83	0	255	399	224
DLBC	27	0	0	0	0	0	0	0	0	0
GBM	595	563	546	287	542	0	491	0	214	276
HNSC	294	255	165	292	0	103	0	89	0	0
KIRC	502	502	490	500	72	469	0	463	454	327
KIRP	135	84	75	117	16	14	0	16	0	0
LAML	202	200	0	192	0	179	0	187	0	199
LGG	144	140	143	0	27	0	0	30	0	0
LIHC	84	55	58	0	0	17	0	28	0	0
LNNH	2	0	0	0	0	0	0	0	0	0
LUAD	372	274	266	347	32	106	0	95	0	147
LUSC	290	272	282	282	154	220	0	202	0	178
OV	592	580	564	551	568	0	564	46	412	316
PAAD	48	0	14	30	0	0	0	0	0	0
PRAD	153	0	100	153	0	0	0	63	0	0
SKCM	253	0	219	240	0	0	0	0	0	0
STAD	162	150	132	133	0	57	0	123	0	133
THCA	274	73	228	230	0	0	0	45	0	0
UCEC	462	425	430	451	54	266	0	359	200	239
PANCANCER	6239	5110	5246	5325	2218	2297	1055	2844	2087	2610

+821  
CopyNumber

+917  
Methylation

# GDAC Session Yesterday

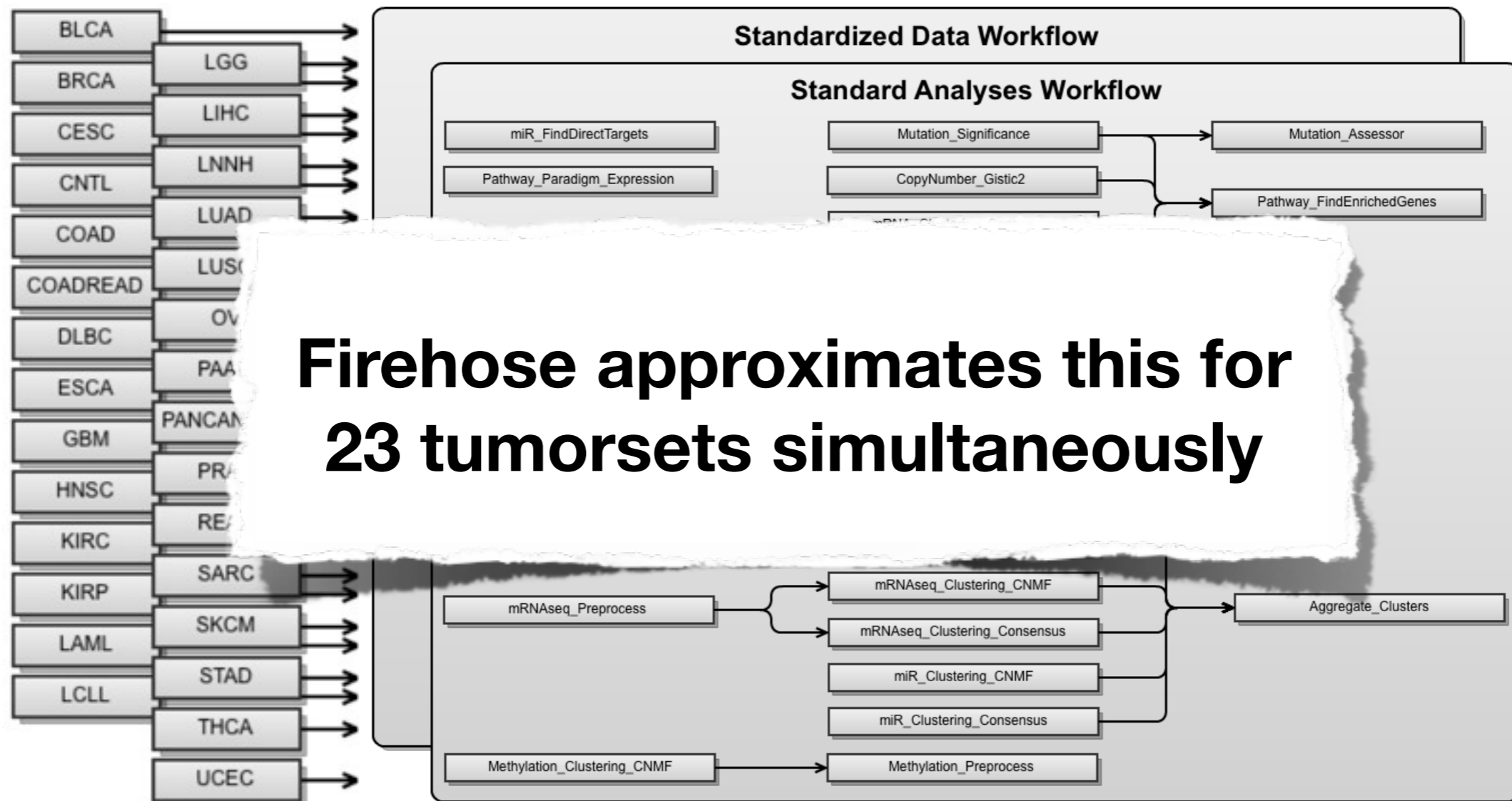
---

Recommendation: formalize AWG co-chair role ***for each tumor type***  
Data Coordinator: ensure best possible data/analysis outcome.

YES!

# GDAC Session Yesterday

Recommendation: formalize AWG co-chair role ***for each tumor type***  
Data Coordinator: ensure best possible data/analysis outcome.





# Volume, Change & Complexity

analyses: 26 x 23 tumor sets / month = 598  
stddata: 273 platforms over 23 tumorsets x 2/month = 546



# Volume, Change & Complexity

analyses: 26 x 23 tumor sets / month = 598  
stddata: 273 platforms over 23 tumorsets x 2/month = 546

- Not even counting RPPA: ingested & almost ready
- CPTAC Proteomics Consortium interested, too
- Nothing on this scale ever attempted before?
- But worthless if we cannot establish scientific credibility

# Enormous QC Challenge

```
graph TD; A[Enormous QC Challenge] --> B[Computing Infrastructure]; A --> C[Scientific Veracity]
```

Computing  
Infrastructure

Scientific  
Veracity

# Enormous QC Challenge

## Computing Infrastructure

Firehose  
Genepattern  
HPC system (LSF)  
Unix filesystems  
DCC Mirroring  
Normalization  
Control Scripts

Website  
Dashboards  
Submission  
etc

## Scientific Veracity

### Data

How much?  
Formatted ok?  
New platforms?  
Combine platforms?

### Algorithms

What knob settings?  
New versions?  
Credible results?  
Wired together properly?  
Reports ok?

# Scale preclude exhaustive inspection

---

# Scale preclude exhaustive inspection

---

## How We Cope

1. Automate
2. Aggregate
3. Clarify
4. Simplify



# Scale preclude exhaustive inspection










---

## How We Cope

1. Automate
2. Aggregate
3. Clarify
4. Simplify










Now some non-trivial examples ...

# Automate: Continuous Unit Testing

Plan	Build	Completed	Tests	Reason
<a href="#">GDAC Ingestor</a>	 #260	3 hours ago	26 passed	Manual build by <a href="#">Daniel DiCara</a>
<a href="#">Module Bam Realign BWA</a>	 #75	5 months ago	10 passed	Dependant of <a href="#">CGA-NB-81</a>
<a href="#">Module Create Merge Data Files SDRF</a>	 #20	1 day ago	5 passed	<a href="#">Updated by Daniel DiCara</a>
<a href="#">Module Iterative Scatter Gather Test</a>	 #57	5 months ago	3 passed	Dependant of <a href="#">CGA-NB-81</a>
<a href="#">Module PVCA Aggregator</a>	 #14	4 months ago	3 passed	<a href="#">Updated by Daniel DiCara</a>
<a href="#">Pipeline GISTIC 2</a>	 #366	19 hours ago	21 passed	<a href="#">Updated by Chip Stewart</a>
<a href="#">Pipeline Mutation Significance</a>	 #374	19 hours ago	2 of 6 failed	<a href="#">Updated by Chip Stewart</a>
<a href="#">Pipeline PARADIGM</a>	 #172	1 month ago	19 passed	Manual build by <a href="#">Daniel DiCara</a>
<a href="#">Pipeline PVCA</a>	 #101	1 month ago	12 passed	Manual build by <a href="#">Daniel DiCara</a>

- Implemented in Bamboo framework
- Regression tests run automatically
- ***Immediately*** when changes checked in for covered tools
- No need for CompBios / BInfs to explicitly run

# Automate: Continuous Unit Testing

Plan	Build	Completed	Tests	Reason
<a href="#">GDAC Ingestor</a>	 #260	3 hours ago	26 passed	Manual build by <a href="#">Daniel DiCara</a>
<a href="#">Module Bam Realign BWA</a>	 #75	5 months ago	10 passed	Dependant of <a href="#">CGA-NB-81</a>
<a href="#">Module Create Merge Data Files SDRF</a>	 #20	1 day ago	5 passed	<a href="#">Updated by Daniel DiCara</a>
<a href="#">Module Iterative Scatter Gather Test</a>	 #57	5 months ago	3 passed	Dependant of <a href="#">CGA-NB-81</a>
<a href="#">Module PVCA Aggregator</a>	 #14	4 months ago	3 passed	<a href="#">Updated by Daniel DiCara</a>
<a href="#">Pipeline GISTIC 2</a>	 #366	19 hours ago	21 passed	<a href="#">Updated by Chip Stewart</a>
<a href="#">Pipeline Mutation Significance</a>	 #374	19 hours ago	2 of 6 failed	<a href="#">Updated by Chip Stewart</a>
<a href="#">Pipeline PARADIGM</a>	 #172	1 month ago	19 passed	Manual build by <a href="#">Daniel DiCara</a>
<a href="#">Pipeline PVCA</a>	 #101	1 month ago	12 passed	Manual build by <a href="#">Daniel DiCara</a>

- Implemented in Bamboo framework
- Regression tests run automatically
- ***Immediately*** when changes checked in for covered tools
- No need for CompBios / BInfs to explicitly run

***“Seems to find bugs before I’ve checked in the code!”***

# Automate: Continuous Unit Testing

Plan	Build	Completed	Tests	Reason
GDAC Ingestor	✔ #260	3 hours ago	26 passed	Manual build by Daniel DiCara
Module Bam Realign BWA	✔ #75	5 months ago	10 passed	Dependant of CGA-NB-81
Module Creat				el DiCara
Module Iterat				A-NB-81
Module PVCA				el DiCara
Pipeline GIST				Stewart
Pipeline Mutation Significance	❗ #374	19 hours ago	2 of 6 failed	Updated by Chip Stewart
Pipeline PARADIGM	✔ #172	1 month ago	19 passed	Manual build by Daniel DiCara
Pipeline PVCA	✔ #101	1 month ago	12 passed	Manual build by Daniel DiCara

**\*** Gradual but steady cultural change

*Where's the Real Bottleneck in Scientific Computing?*

[www.americanscientist.org](http://www.americanscientist.org) Jan/Feb 2006

- Implemented in Bamboo framework
- Regression tests run automatically
- **Immediately** when changes checked in for covered tools
- No need for CompBios / BInfs to explicitly run **\***

***“Seems to find bugs before I’ve checked in the code!”***

# But Automation Not Enough

When N things break : manually diagnose / fix

Or automation N/A : eyeball clusters/plots



# But Automation Not Enough

task	BLCA	BRCA	CESC	COADREAD	DLBC	GBM	HNSC	KIRC	KIRP	LAML	LGG	LIHC	LNNH	LUAD	LUSC	OV	PAAD	PANCANCER	PRAD	SKCM	STAD	THCA	UCEC
Aggregate_Clusters	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CopyNumber_GeneBySample	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CopyNumber_Gistic2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CopyNumber_Preprocess	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0
Correlate_CopyNumber_vs_miR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Correlate_CopyNumber_vs_mRNA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Correlate_Methylation_vs_mRNA	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Methylation_Clustering_CNMF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Methylation_Preprocess	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
miRseq_Clustering_CNMF	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
miRseq_Clustering_Consensus	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
miRseq_Preprocess	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
miR_Clustering_CNMF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
miR_Clustering_Consensus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
miR_FindDirectTargets	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mRNAseq_Clustering_CNMF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mRNAseq_Clustering_Consensus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mRNAseq_Preprocess	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0
mRNA_Clustering_CNMF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
mRNA_Clustering_Consensus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
mRNA_Preprocess_Median	1	0	1	0	1	0	1	0	0	1	0	1	1	0	0	0	1	0	1	1	1	1	0
Mutation_Assessor	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mutation_Significance	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Pathway_FindEnrichedGenes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pathway_Paradigm_Expression	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Pathway_Paradigm_Expression_CopyNumber	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1
Pathway_Paradigm_Lite	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Aggregate: failure dashboard for all tumors & analyses  
 Clarify: instant synoptic view of entire run

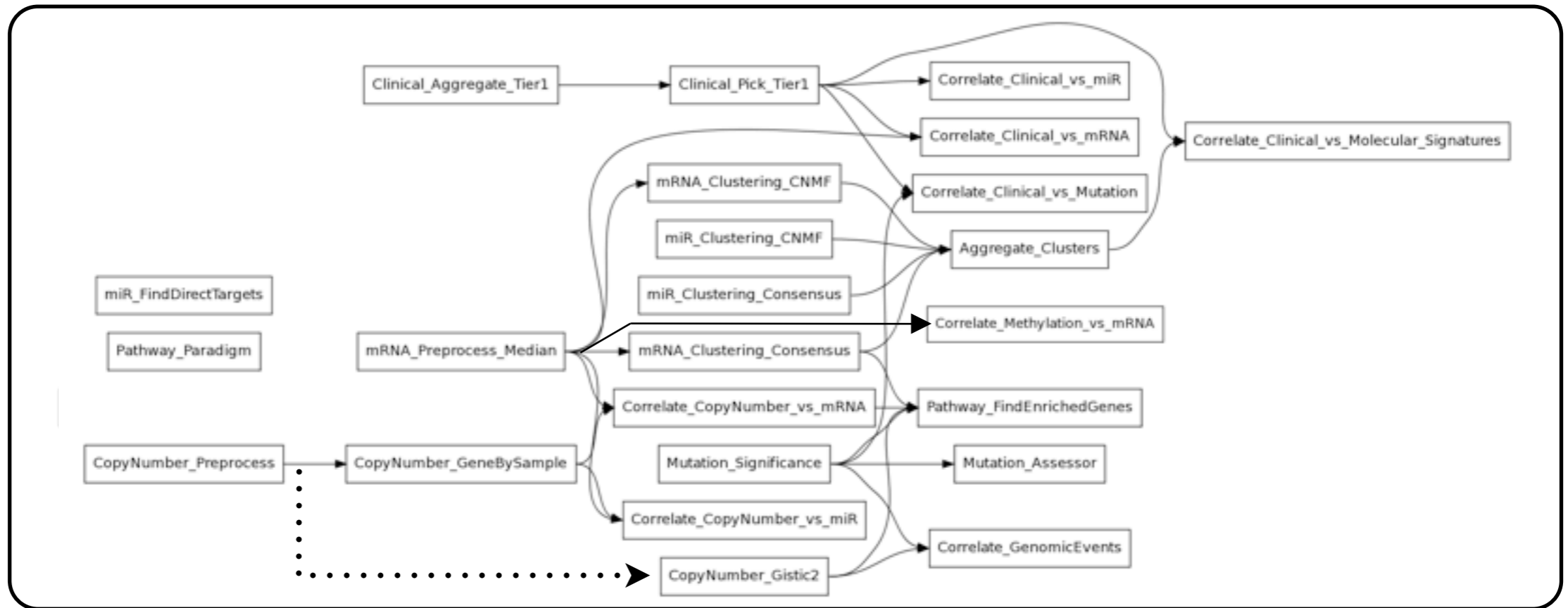
Saves 100s of clicks through Firehose GUI



# Aggregate & Automate: Integration Testing

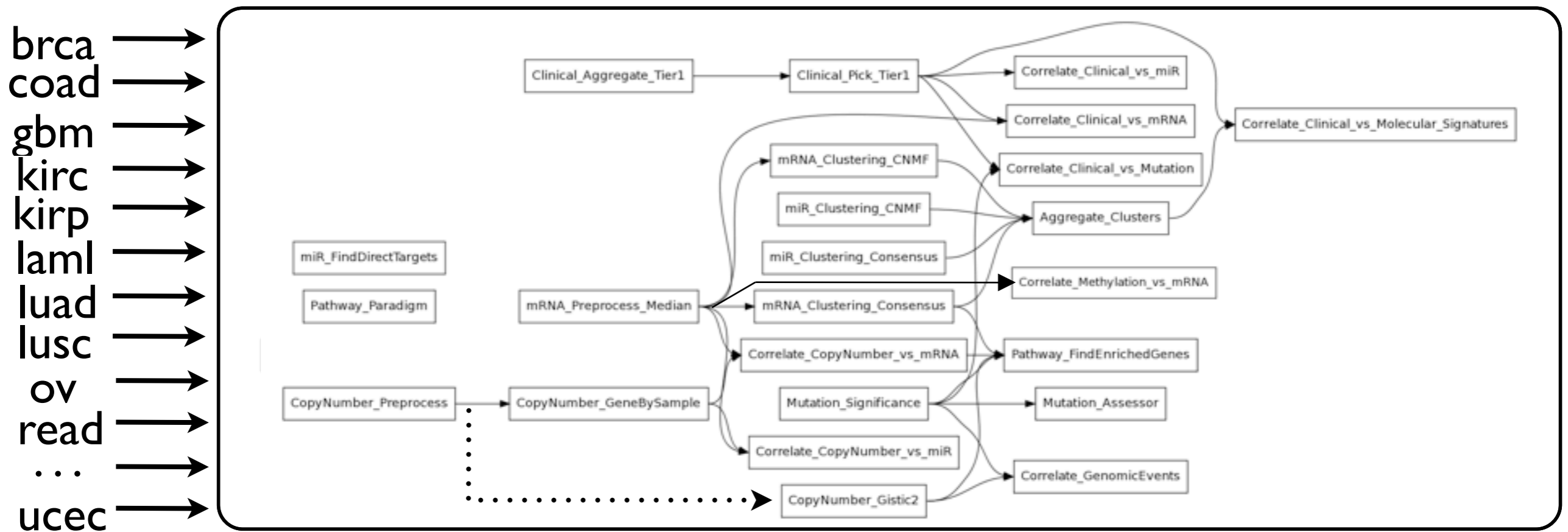
- Unit tests must be predictable
- Which means stable inputs and outputs
- Essentially mandates old data
- What about realtime? More current/live data?

# Aggregate & Automate: Integration Testing



Establish that code changes play nice with rest of system

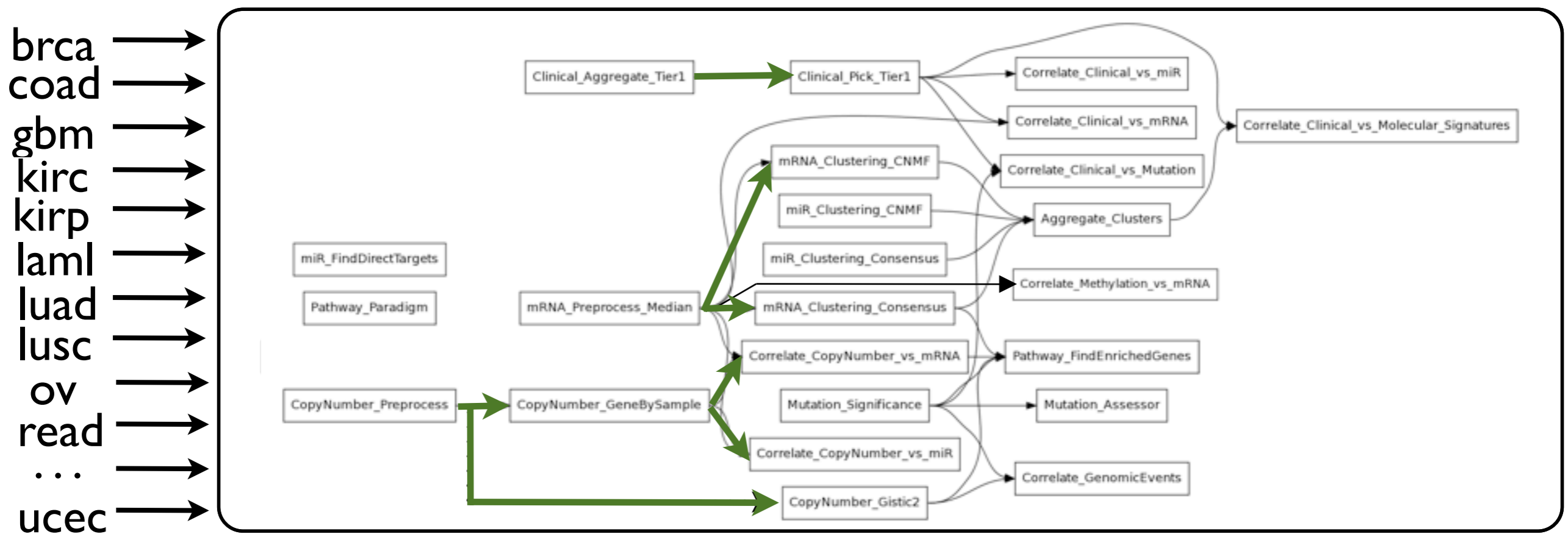
# Aggregate & Automate: Integration Testing



Establish that code changes play nice with rest of system

Across all datasets

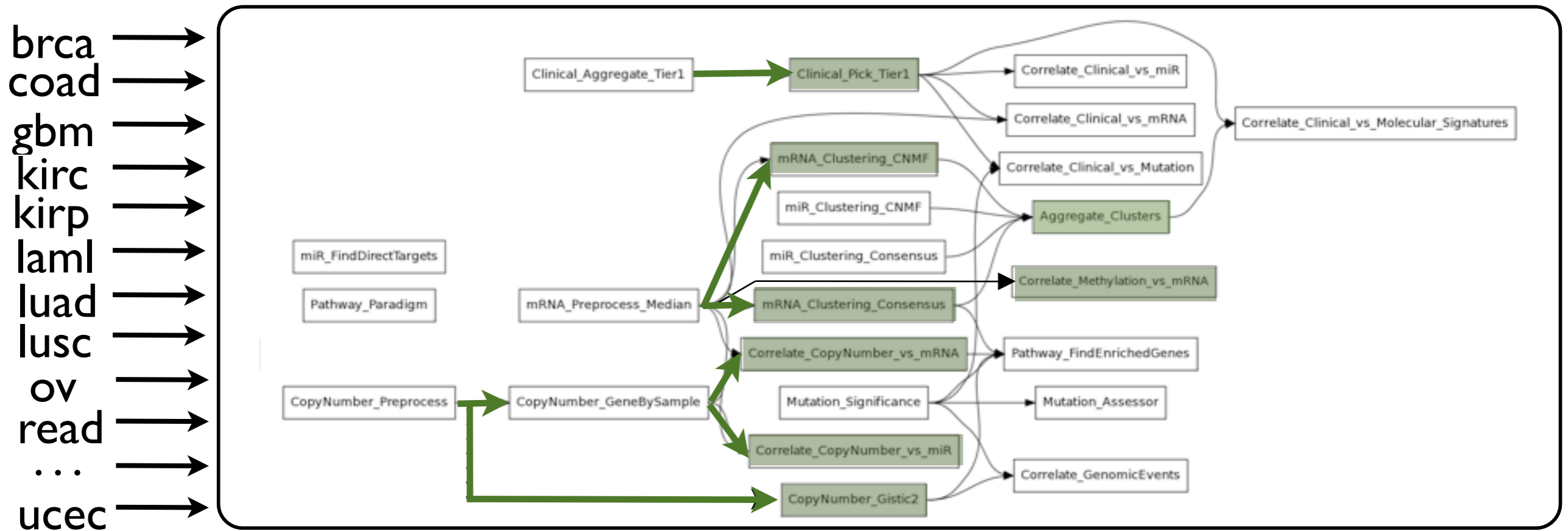
# Aggregate & Automate: Integration Testing



Establish that code changes play nice with rest of system

Across all datasets  
With O's correctly wired to I's

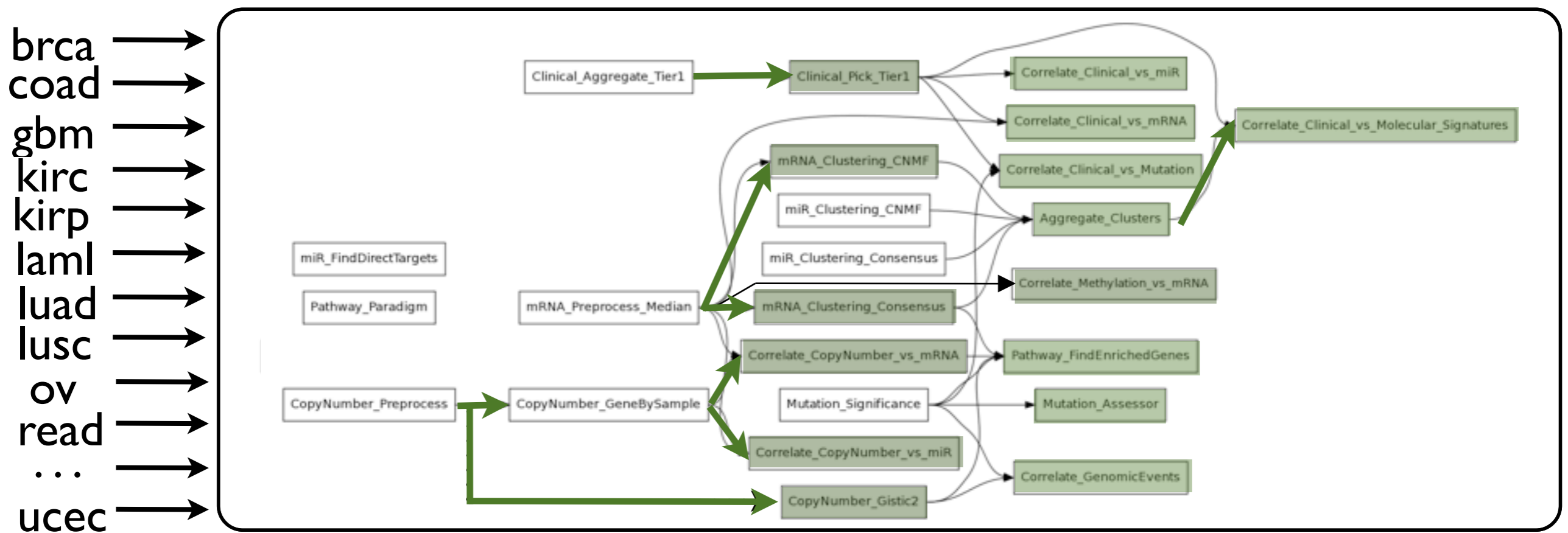
# Aggregate & Automate: Integration Testing



Establish that code changes play nice with rest of system

Across all datasets  
Downstream dependents ***correctly read*** outputs  
With O's correctly wired to I's

# Aggregate & Automate: Integration Testing



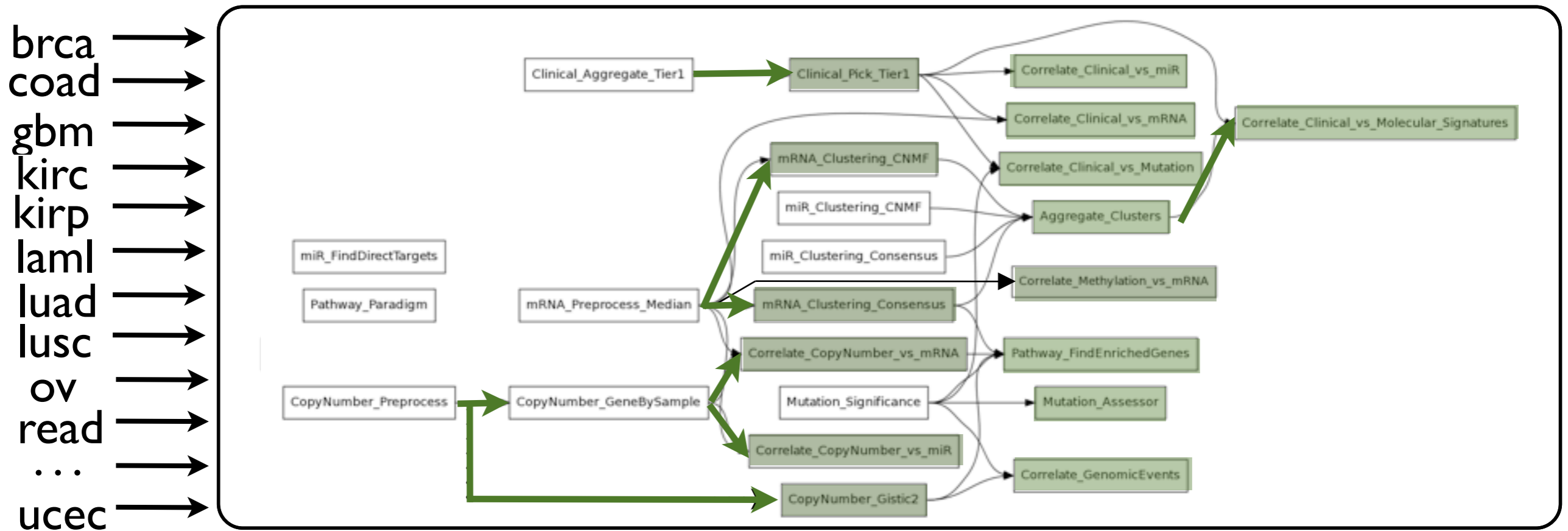
Establish that code changes play nice with rest of system

Across all datasets  
With O's correctly wired to I's

Downstream dependents ***correctly read*** outputs  
And remainder of workflow runs to completion



# Aggregate & Automate: Integration Testing



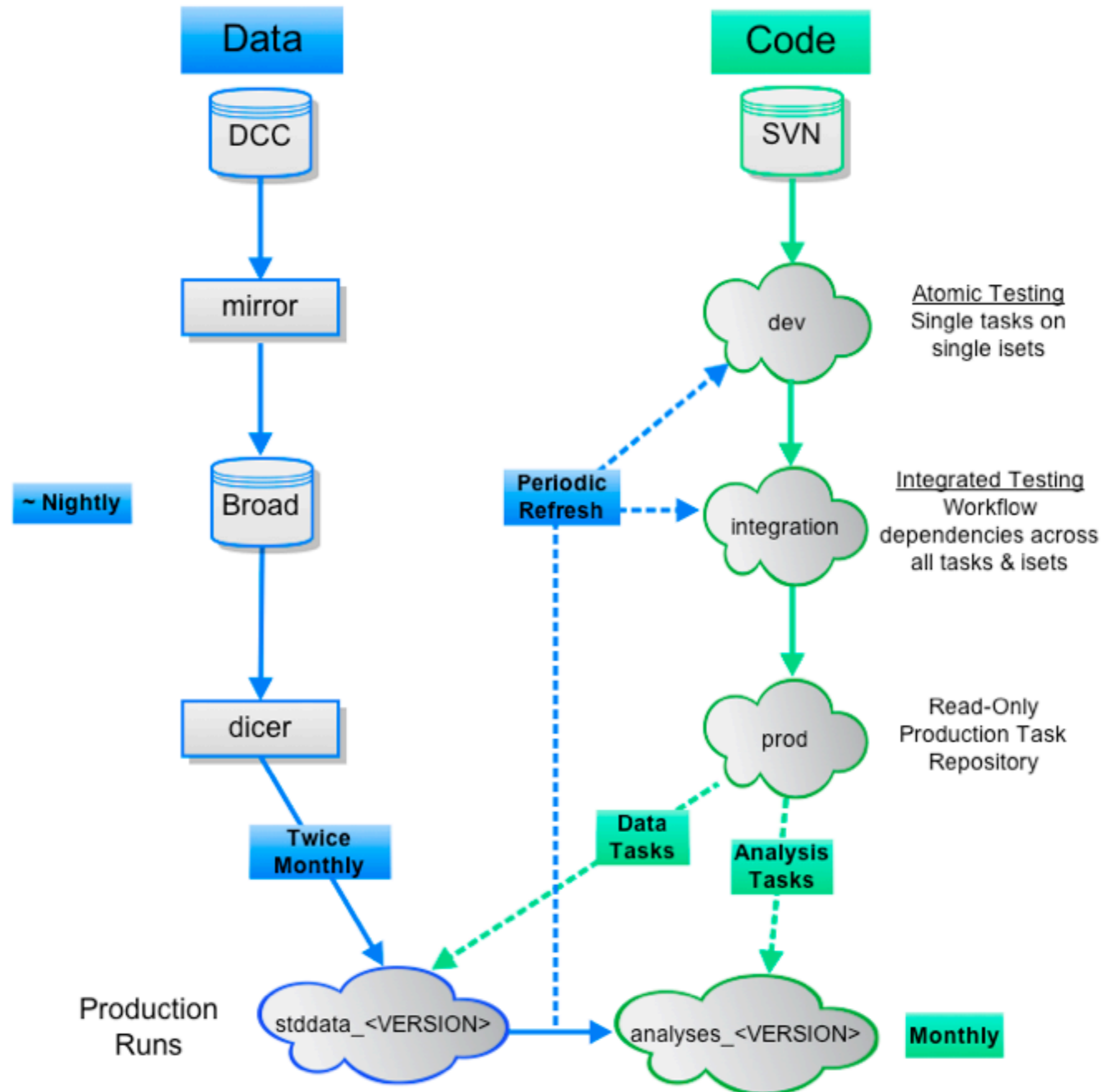
Establish that code changes play nice with rest of system

Across all datasets  
With O's correctly wired to I's

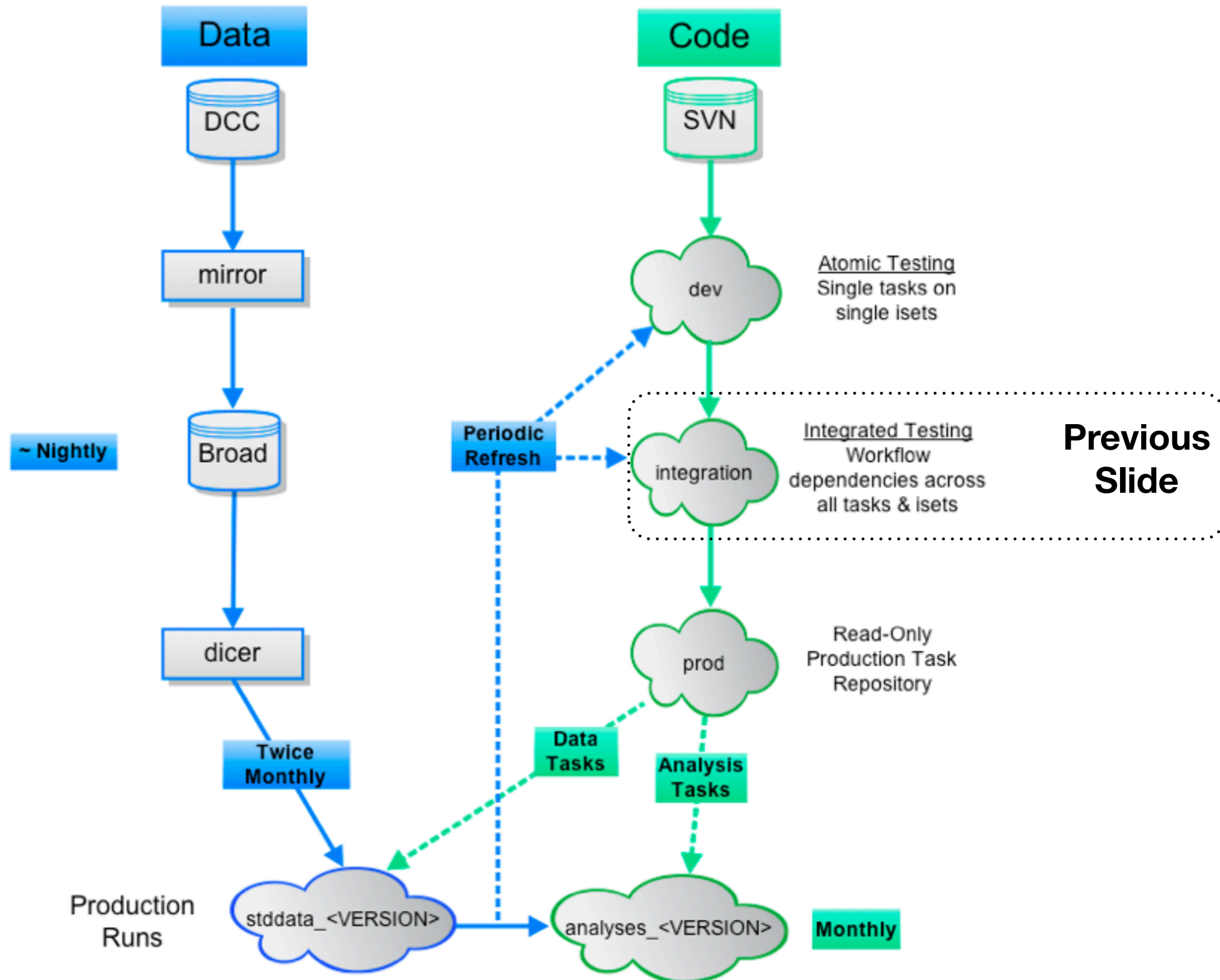
Downstream dependents ***correctly read*** outputs  
And remainder of workflow runs to completion

Using same automation infrastructure as production runs.

# Clarify : Internal Process Flow

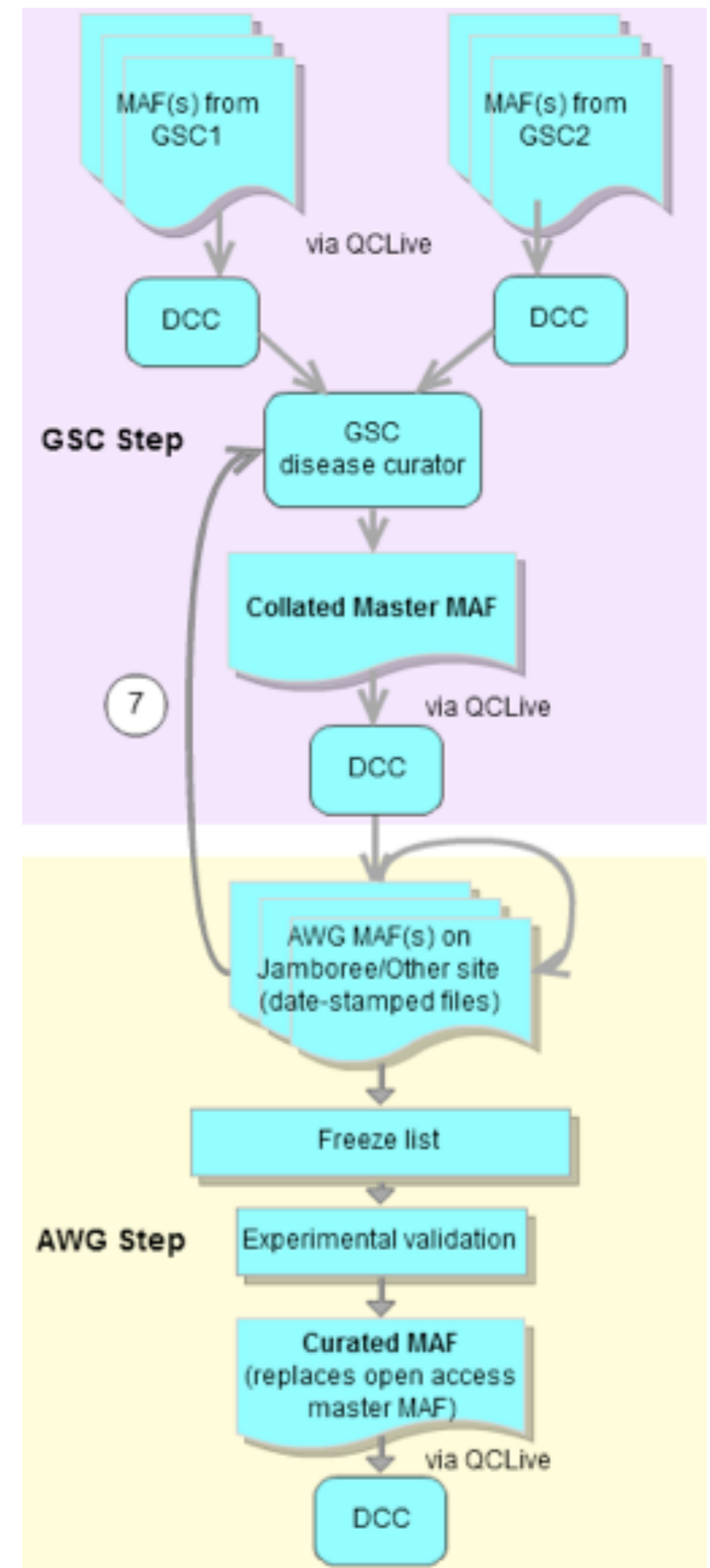


# Clarify : Internal Process Flow



# Clarify & Simplify: TCGA MAF WorkFlow

Nikki Schultz  
Broad Institute  
Prachi Kothiyal  
Heidi Sofia &  
Many Others



# Simplify : **firehose\_get**

2012\_03\_21 stddata Run

ReleaseNotes	# Datasets	% Processed	Download
BLCA	8	100%	Open Protected
BRCA	16	100%	Open Protected
CESC	7	100%	Open Protected
COADREAD	14	100%	Open Protected
GBM	21	100%	Open Protected
HNSC	7	100%	Open Protected
KIRC	7	100%	Open Protected
KIRP	7	100%	Open Protected
LAML	7	100%	Open Protected
LGG	14	100%	Open Protected
LHC	7	100%	Open Protected
LUAD	7	100%	Open Protected
LUSC	7	100%	Open Protected
OV	7	100%	Open Protected
PAAD	3	100%	Open Protected
PRAD	5	100%	Open Protected
SKCM	1	100%	Open Protected
STAD	14	100%	Open Protected
THCA	7	100%	Open Protected
UCEC	16	100%	Open Protected
PANCANCER	34	85%	Open Protected

stddata  
dashboard

2012\_03\_21 analyses Run

AnalysisReport	# Pipelines	% Successful	Download
BRCA	23	100%	Open Protected
COADREAD	23	100%	Open Protected
GBM	21	100%	Open Protected
LGG	14	100%	Open Protected
LUSC	23	100%	Open Protected
OV	7	100%	Open Protected
KIRC	7	100%	Open Protected
LUAD	7	100%	Open Protected
UCEC	7	100%	Open Protected
STAD	7	100%	Open Protected
KIRP	7	100%	Open Protected
PRAD	7	100%	Open Protected
THCA	7	100%	Open Protected
LAML	7	100%	Open Protected
BLCA	7	78%	Open Protected
HNSC	7	78%	Open Protected
LHC	7	78%	Open Protected
CESC	6	60%	Open Protected
PAAD	3	60%	Open Protected
PANCANCER	11	58%	Open Protected

analyses  
dashboard

## Short 10k script

- Download all or parts
- Of data or analyses runs
- Open access : no password
- Select by run type & date
- Subselect by tumor type
- Or analyses type / name
- See what runs we did
- Or what tasks in each run



# Simplify : **firehose\_get**

2012\_03\_21 stddata Run

ReleaseNotes	# Datasets	% Processed	Download
BLCA	8	100%	Open Protected
BRCA	16	100%	Open Protected
CESC	7	100%	Open Protected
COADREAD	14	100%	Open Protected
GBM	21	100%	Open Protected
HNSC	7	100%	Open Protected
KIRC	7	100%	Open Protected
KIRP	7	100%	Open Protected
LAML	7	100%	Open Protected
LGG	14	100%	Open Protected
LHC	7	100%	Open Protected
LUAD	7	100%	Open Protected
LUSC	7	100%	Open Protected
OV	7	100%	Open Protected
PAAD	3	100%	Open Protected
PRAD	5	100%	Open Protected
SKCM	1	100%	Open Protected
STAD	14	100%	Open Protected
THCA	7	100%	Open Protected
UCEC	16	100%	Open Protected
PANCANCER	34	85%	Open Protected

stddata  
dashboard

2012\_03\_21 analyses Run

AnalysisReport	# Pipelines	% Successful	Download
BRCA	23	100%	Open Protected
COADREAD	23	100%	Open Protected
GBM	21	100%	Open Protected
LGG	14	100%	Open Protected
LUSC	23	100%	Open Protected
OV	7	100%	Open Protected
KIRC	7	100%	Open Protected
LUAD	7	100%	Open Protected
UCEC	7	100%	Open Protected
STAD	7	100%	Open Protected
KIRP	7	100%	Open Protected
PRAD	7	100%	Open Protected
THCA	7	100%	Open Protected
LAML	7	100%	Open Protected
BLCA	7	78%	Open Protected
HNSC	7	78%	Open Protected
LHC	7	78%	Open Protected
CESC	6	60%	Open Protected
PAAD	3	60%	Open Protected
PANCANCER	11	58%	Open Protected

analyses  
dashboard

## Short 10k script

- Download all or parts
- Of data or analyses runs
- Open access : no password
- Select by run type & date
- Subselect by tumor type
- Or analyses type / name
- See what runs we did
- Or what tasks in each run

## Easier

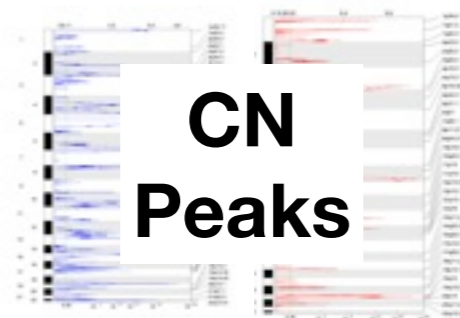
- **more eyeballs**
- **higher quality**
- **better science**

# Clarify & Simplify : Quicklook Summaries

## OV Analysis Summary

This is the analysis overview for Firehose run "21 March 2012".

There were 547 tumor samples used in this analysis: 29 sign



### Significantly Mutated Genes

Table 3. A Ranked List of Significantly Mutated Genes. Number of significant genes found: 24. Number of genes displayed: 35.

rank	gene	description	N	h	spat	mtte	mtll	mt	mt2	mt3	mt4	mt5	mt6	P	Q
1	TP53	tumor protein p53	493241	279	276	142	3	47	21	29	51	104	0	<1.00e-10	<1.7e-11
2	BRCA1	breast cancer 1, early onset	270441	12	12	12	0	0	0	1	0	11	0	1.76e-08	0.00002
3	RB1	retinoblastoma 1 (including osteosarcoma)	0	0	0	0	0	0	0	0	0	0	0	7.58e-07	0.0043
4	NF1	neurofibromin 1 (neurofibromatosis type 1; von Recklinghausen disease, NF1 disease)	0	0	0	0	0	0	0	0	0	0	0	1.69e-06	0.0027
5	CSMD1	CUB and Sushi multiple domain 1	0	0	0	0	0	0	0	0	0	0	0	1.66e-06	0.0027
6	FAT2	FAT tumor suppressor homolog 2 (Drosophila)	270441	20	18	20	1	4	2	3	10	1	0	0.000011	0.032

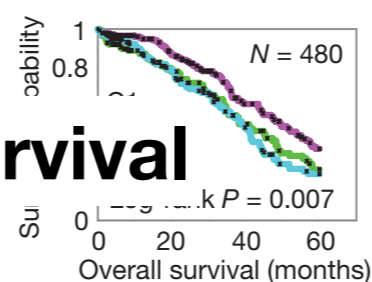
## SMGs

is NMF using genes was rs. We for k = 2 genetic on.

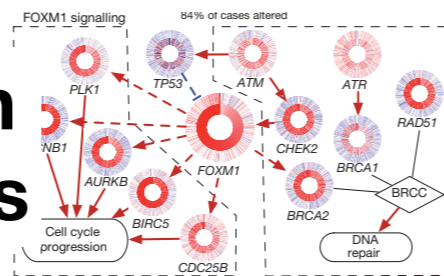


ster P

## Survival



## Path Ways



- Pulse of entire tumor analyses
- In key representational figures
- With short accompanying text
- Executive summary from
- 50+ page Nozzle reports
- Auto-excerpted from them

Human interpretation needed, but not automate-able  
This distills into most concise aggregate form



## Related Posters

- Poster : *Engineering Firehose* (DiCara et al)
- Poster : *RNA-Seq in Firehose* (Zhang et al)
- Poster : *GDAC Interoperability* (Cerami et al)
- Poster : *Broad SNP6 Pipeline* (Saksena et al)

# Acknowledgements

PI: **Lynda Chin, Gaddy Getz**

## **Broad**

**Michael Noble**

**Douglas Voet**

**Gordon Saksena**

**Dan DiCara**

Kristian Cibulskis

Juok Cho

Rui Jing

Michael Lawrence

Lee Lichtenstein

Pei Lin

Spring Liu

Aaron McKenna

Sachet Shukla

Raktim Sinha

Andrey Sivachenko

Carrie Sougnez

Petar Stojanov

Lihua Zhou

Hailei Zhang

Robert Zupko

## **Belfler-DFCI/MDACC**

Yonghong Xiao

Juinhua Zhang

Terrence Wu

## **IGV & GenePattern teams @ Broad**

Jill Mesirov

Michael Reich

Peter Carr

Marc-Danie Nazaire

Jim Robinson

Helga Thorvaldsdottir

## **Harvard**

**Peter Park**

**Nils Gehlenborg**

Semin Lee

Richard Park

**Matthew Meyerson**

Todd Golub

Eric Lander

