FIREHOSE
Broad GDAC

Broad/Harvard GDAC Analysis Pipeline

Nils Gehlenborg | Harvard Medical School & Broad Institute

The Cancer Genome Atlas Steering Committee Meeting | April 27-29, 2011

FIREHOSE
Broad GDAC

Broad/Harvard GDAC Analysis Pipeline

Nils Gehlenborg | Harvard Medical School & Broad Institute

With help from

Michael S Noble
Douglas Voet

Lihua Zou
Gordon Saksena
Pei Lin

Hailei Zhang
John Zhang
Richard Park
Sachet Shukla
Spring Liu
Terrence Wu

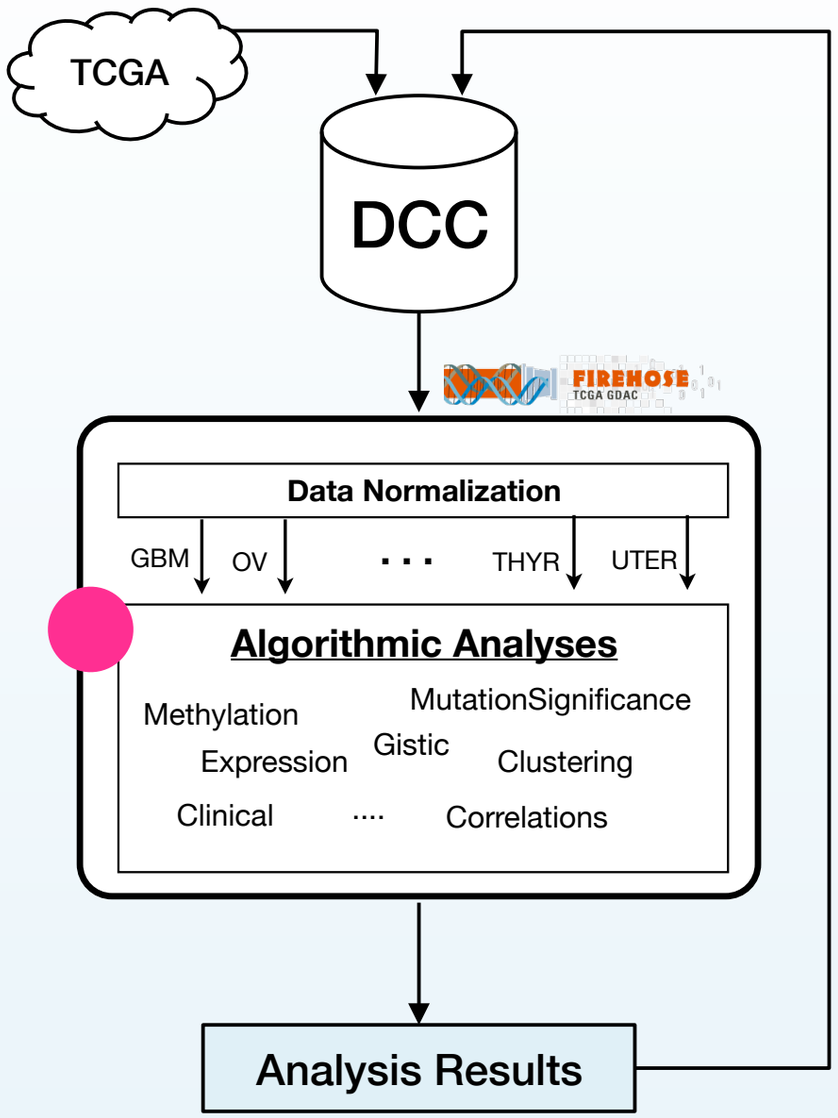
Peter J Park
Gad Getz
Lynda Chin

The Cancer Genome Atlas Steering Committee Meeting | April 27-29, 2011

FIREHOSE
Broad GDAC

Algorithms

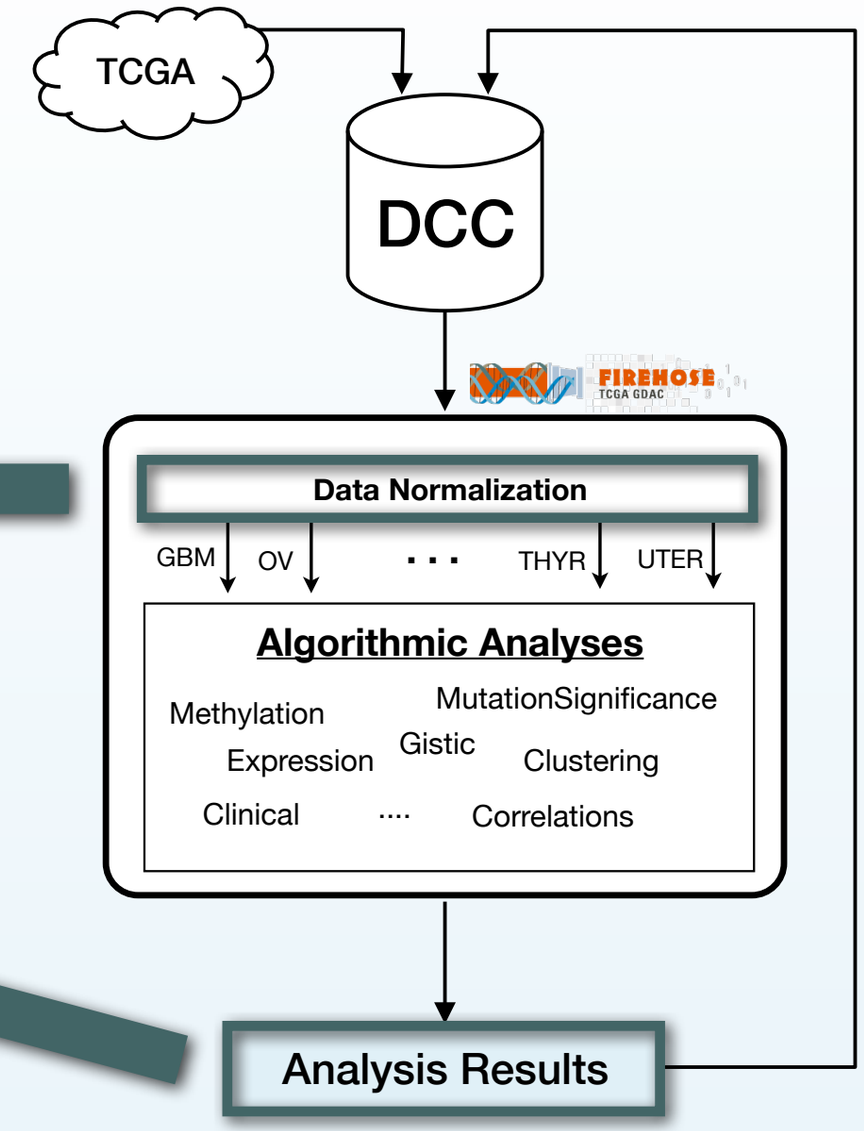
- **miR:** clustering, correlations
- **mRNA:** clustering, correlation
- **Methylation:** correlations
- **Clinical:** correlations
- **CNA:** GISTIC2, correlations
- **Mutations:** MutSig, MutationAssessor, correlations
- **Pathways:** Paradigm, EnrichedGenes



FIREHOSE
Broad GDAC

1 **Level 3 Data**

2 **Level 4 Data**



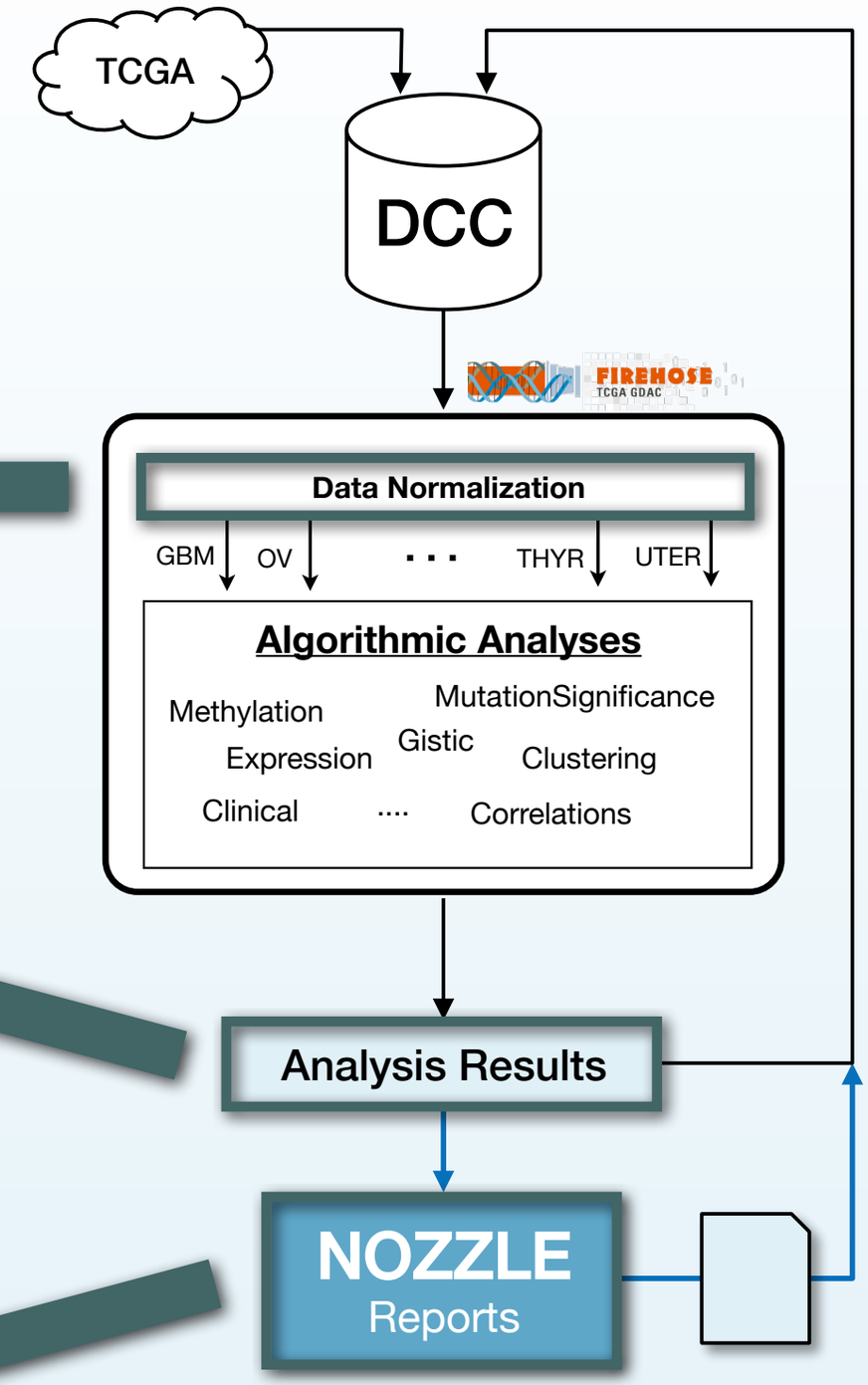
Scalable solution for standard analyses

FIREHOSE
Broad GDAC

1 **Level 3 Data**
for computers

2 **Level 4 Data**
for computers

3 **Reports**
for cancer biologists!



1. TCGA Analysis Working Groups
2. Cancer Biology Community

FIREHOSE
Broad GDAC

Reports: Goal

Make *effective* interpretation of analysis results as *efficient* as possible.

FIREHOSE
Broad GDAC

Reports: Challenges

1. **Number of analyses**
   ~25 modules × ~20 tumor types ≈ 500
2. **Amount and depth of results**
3. **Spectrum of consumers**

FIREHOSE
Broad GDAC

Reports: Solution

1 Common report structure and layout

1. Overview, 2. Results, 3. Methods & Data

2 Hierarchical organization of content

Overview first, details on demand

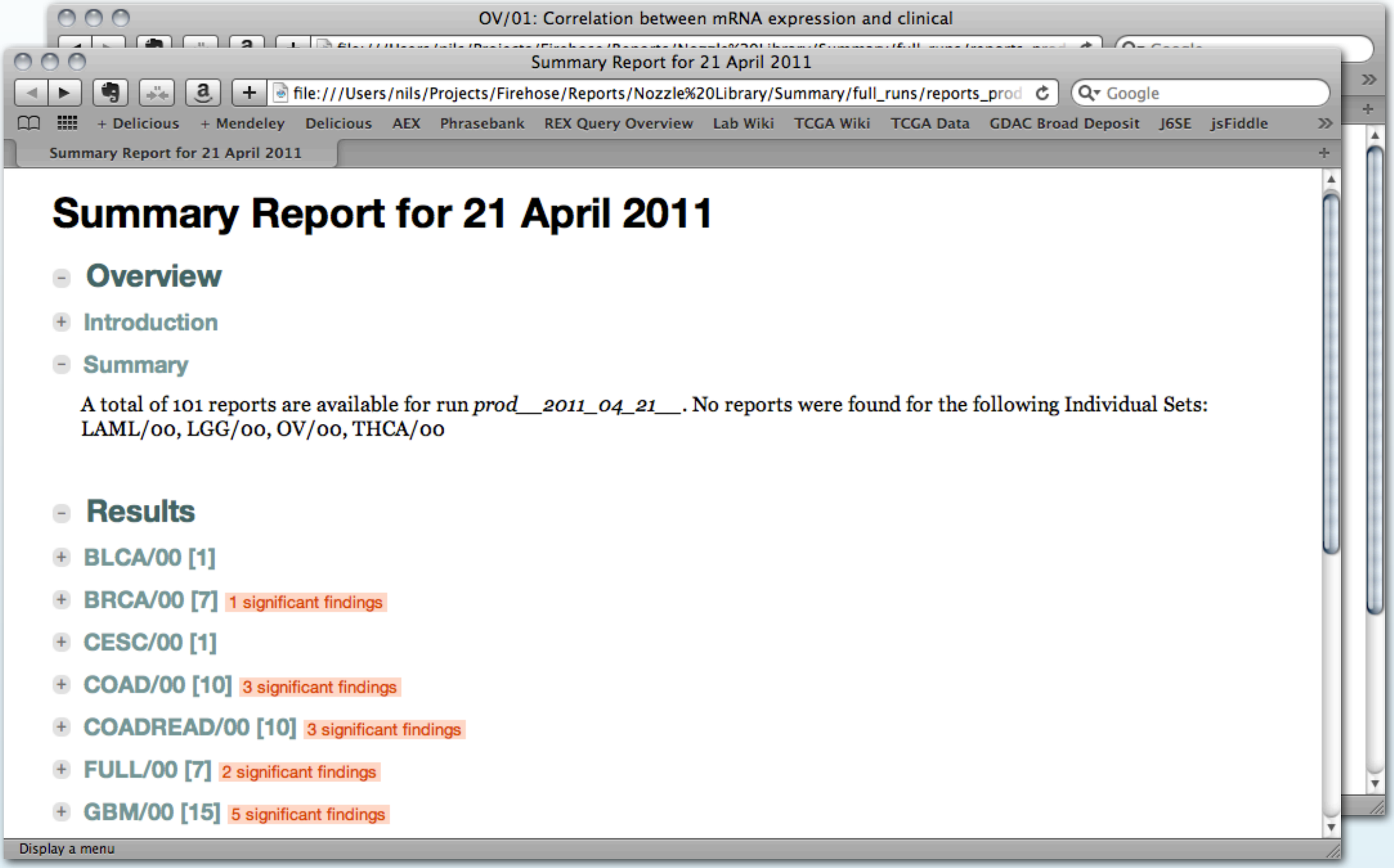
3 Highlighting of significant findings

4 Implementation in *Nozzle* R package

See my poster for details

FIREHOSE
Broad GDAC

Reports: Demo



FIREHOSE
Broad GDAC

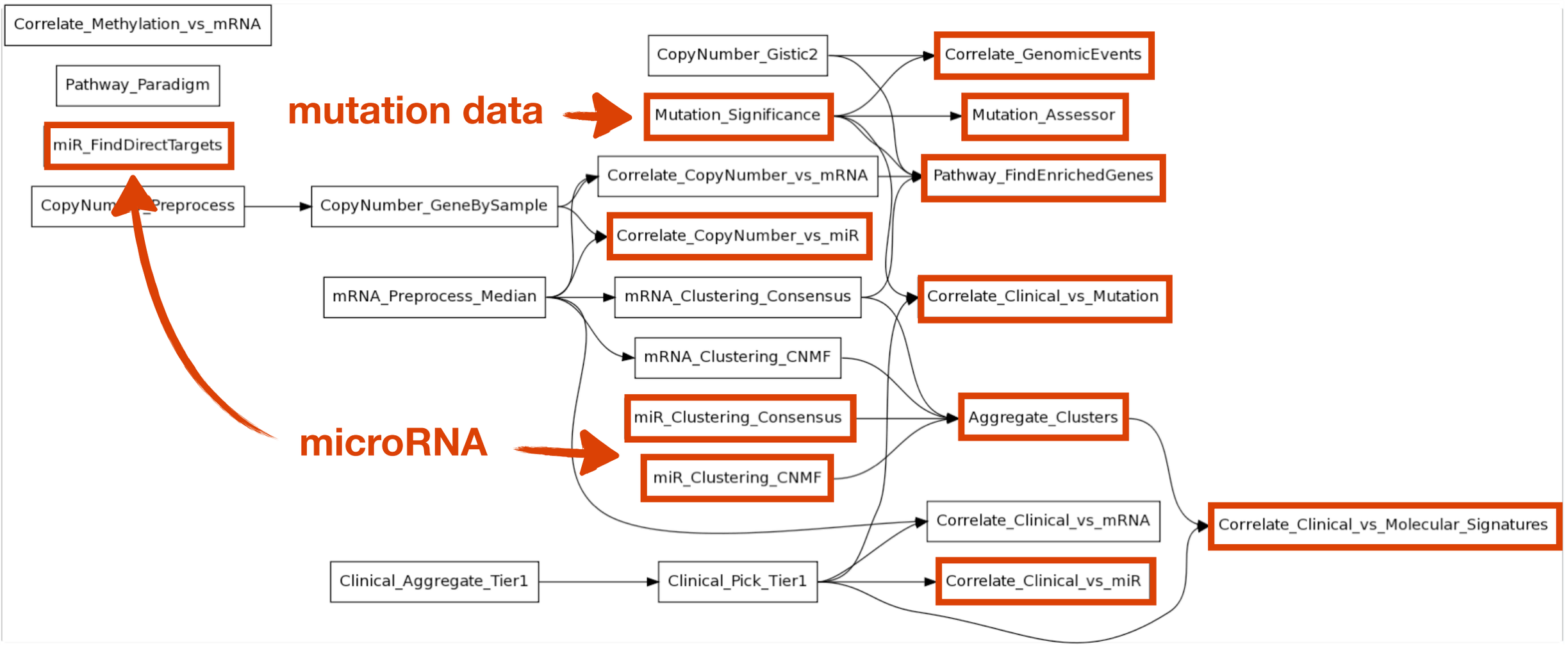
Reports: Issues

microRNA and mutation data

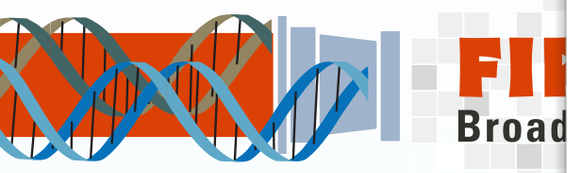
TumorType	Biospecimen	Any_Level_1	Clinical	CNA	Methylation	mRNA	miR	MAF
BLCA	26	12	9	9	0	0	0	0
BRCA	647	390	353	375	186	434	0	0
CESC	23	8	5	8	0	0	0	0
COAD	245	151	207	182	167	155	0	88
COADREAD	338	203	285	253	236	224	0	139
GBM	508	476	465	466	288	506	415	199
HNSC	59	59	0	57	0	0	0	0
KIRC	460	347	192	345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

FIREHOSE
Broad GDAC

Reports: Issues



50% of all analysis modules affected



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%

This File Generated on : Wed Apr 27 14:57:41 EDT 2011

50% of all analysis modules affected

FIREHOSE
Broad GDAC

Reports: Questions

When should we share these reports?

With whom should we share these reports?

How do we address patient privacy issues?