Broad GDAC Pipeline
Run Status
February 2011

March 3, 2011

Michael S. Noble
mnoble@broadinstitute.org

Summary of TCGA Tumor Data
Ingested into Broad GDAC Pipeline
02/17/2011 Run

TumorType	Biospecimen	Any_Level_1	Clinical	CNA	Methylation	mRNA	miR	MAF
BRCA	434	186	244	265	186	346	0	0
CESC	8	0	0	0	0	0	0	0
COAD	203	151	160	137	167	155	0	64
GBM	508	448	460	466	288	471	415	199
HNSC	39	0	0	0	0	0	0	0
KIRC	354	39	19	254	219	41	0	0
KIRP	48	39	15	16	36	41	0	0
LAML	202	0	0	0	188	0	0	135
LGG	30	0	0	0	0	0	0	0
LUAD	128	21	11	56	128	33	0	0
LUSC	161	116	42	117	133	134	0	0
OV	576	570	524	519	425	519	566	384
READ	79	52	78	51	69	69	0	13
STAD	82	35	0	81	82	0	0	0
THCA	25	0	0	0	0	0	0	0
UCEC	192	24	103	114	70	0	0	0
Total	3069	1681	1656	2076	1991	1809	981	795

Two new tumor types: CESC, THCA

Combined COAD+READ dataset analyzed & uploaded to DCC

Another run planned next 2-3 days: new colorectal data for AWG

Summary of TCGA Tumor Data
Ingested into Broad GDAC Pipeline
January 14, 2010 Run

Tumor Type	Biospecimen	Any_Level_1	Clinical	CNA	Methylation	mRNA	miR	MAF
BRCA	346	186	244	265	186	280	0	0
COAD	203	151	130	137	167	155	0	64
GBM	508	448	490	466	288	444	415	169
HNSC	39	0	0	0	0	0	0	0
KIRC	355	39	19	254	219	41	0	0
KIRP	48	39	0	16	36	41	0	0
LAML	202	0	0	0	188	0	0	0
LGG	30	0	0	0	0	0	0	0
LUAD	128	21	11	56	128	33	0	0
LUSC	160	116	42	117	133	116	0	0
OV	584	570	532	519	425	519	566	384
READ	79	52	72	51	69	69	0	12
STAD	82	35	0	81	82	0	0	0
UCEC	145	24	0	114	70	0	0	0
Totals	2909	1681	1540	2076	1991	1698	981	629

No time to create diffs plot yet, but high on the list ...

Analyses Summary : 17 Tumor datasets

Tumor Type	# Completed	Percentage
OV	25	100%
GBM	25	100%
All Combined	22	88%
COAD	15	60%
LUSC	15	60%
COADREAD	14	56%
BRCA	12	48%
KIRC	12	48%
LUAD	12	48%
READ	10	40%
KIRP	7	28%
UCEC	4	16%
STAD	3	12%
LAML	2	8%
CESC	0	0%
HNSC	0	0%
THCA	0	0%

New FireHose web services make this status reporting easy
Suite of staus reports should be available online within ~1 week
When new **gdac.broadinstitute.org** website goes live

Results

- Every completed pipeline uploaded to DCC
- Available from **tinyurl.com/tcga-gdac-broad/[TUMOR]/20110217[00]** *
- Circa 841 files: 300 mergers, 541 analysis results (primary + aux)
- As promised, pipeline names have changed to be more:
    - compact
    - consistent
    - clear
    - descriptive (e.g. indicate data type and/or function)

* https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/tcga4yeo/other/gdacs/gdacbroad [TUMOR]/20110217[00]

New Pipeline Nomenclature

Old	New
convertCNLevelIIIData	CopyNumber_Preprocess
MakeReducedSegment	CopyNumber_GeneBySample
Gistic2	CopyNumber_Gistic2
GDAC CNMF_mRNA_clustering	mRNA_Clustering_CNMF
GDAC mRNAConsensusClustering	mRNA_Clustering_Consensus
GDAC median mRNA Expression	mRNA_Preprocess_Median
GDAC CNMF_miRNA_clustering	miR_Clustering_CNMF
GDAC miRNAConsensusClustering	miR_Clustering_Consensus
Find miR Direct Targets	miR_FindDirectTargets

New Pipeline Nomenclature ...

Old	New
GDAC_clinicalDataMergerPipeline_clinical	Clinical_Merge_Tier1
GDAC_clinicalDataPickerPipeline_clinical	Clinical_Pick_Tier1
Paradigm	Pathway_ParadigmLite
GDAC_geneListPathwayEnrichmentPipeline	Pathway_FindEnrichedGenes
MutSig	Mutation_Significance
MutationAssessor	Mutation_Assessor

New Pipeline Nomenclature ...

Old	New
Correlate microRNA Expression with Clinical Data	Correlate_Clinical_vs_miR
Correlate mRNA Expression with Clinical data	Correlate_Clinical_vs_mRNA
Correlate gene mutation status with Clinical data	Correlate_Clinical_vs_Mutation
Correlate miRNA CNMF clustering with Clinical data	Correlate_Clinical_vs_miR_Clusters_CNMF
Correlate miRNA consensus clustering with Clinical data	Correlate_Clinical_vs_miR_Clusters_Consensus
Correlate mRNA CNMF clustering with Clinical data	Correlate_Clinical_vs_mRNA_Clusters_CNMF
Correlate mRNA consensus clustering with Clinical data	Correlate_Clinical_vs_mRNA_Clusters_Consensus
GetCopyNumberExpCor	Correlate_CopyNumber_vs_mRNA
GetCopyNumberExpCorMiRNA	Correlate_CopyNumber_vs_miR
GDAC Correlate Expression with Methylation	Correlate_Methylation_vs_mRNA

Nozzle : new pipeline reporting library

- Nils Gehlenborg, Lihua Zou, et al
- R implementation; others may follow
- Reduces need to write HTML to ~zero : analysts focus on science content
- Being integrated into Firehose infrastructure & GenePattern pipelines
- Inconsistent/static HTML ⟶ Consistent/dynamic, using CSS & JavaScript
- Tags: allow automatic seeding 70-80% of tumor analysis summary report
- Should be visible in results of late March run

Data Volatility : Biggest Impediment to Automation

- Clinical: consider COAD Tier 1 CDEs: in late Feb (19th?) 2011:
    - admin.dayofdccupload
    - changed to: admin.daystodccupload
- DCC gets advance notification, but others downstream do not
- Breaks pipeline automation -- requires seemingly endless manual intervention
- How can DCC validation help maintain downstream automation?
- To whom else can we look to define & enforce standards?

Data Volatility: clinical requests ...

- Can experts in each cancer WG provide us list of clinical variables of interest to that cancer type?

- From which we can then generate the tier 2 clinical data?

- And recommend simple analyses that could be provided by our pipelines?

- Whose outputs can be added as new columns into the tier 1/2 data?

Data Volatility : New-Submission-Format SDRF?

- With BCR data: no indication of which tar files go together
- EXAMPLE: recently we were informed there was new GBM data ...
- Saw one new batch was submitted, and went to work on it ...
- ... next day remaining 14 batches showed up
- A new-submission-format SDRF file would have minimized problems
- Batch 37 was mirrored Feb 24, the others on Feb 25

Data Volatility : Samples in multiple batches

- Ovarian clinical samples appear to be hopping between batches
- Last month: BCR samples from batch 3 began being duplicated in batch 4, for both biospecimen and clinical data (GBM)
- Fixed now, but now similar issue cropped up again in Feb 11 snapshot:
    - TCGA-12-0670 in batches 8 and 10
- A new-submission-format SDRF would make this obvious to detect
- As there would be non-unique keys

How Many Tumors Should We Run Against?

- Presently initiate runs against all tumor types with ANY data
- But as shown in slide 4, and mostly for a lack of samples:
    - >1/3 of our tumor passes are currently < 30% success
        - And several are at 0%
- Should we introduce threshold samples value?
- Move away from vacuous workspaces, dashboard entries, etc
- Towards parsimonious determinism