Pancancer Status in Firehose

M. Noble

TCGA Pancancer Working Group Telecon

April 5, 2012

Acknowledgements

PI: Lynda Chin, Gaddy Getz

Broad
Michael Noble
Douglas Voet
Gordon Saksena
Rui Jing
Dan DiCara
Raktim Sinha
Hailei Zhang
Pei Lin
Juok Cho
Lee Lichtenstein
Michael Lawrence
Andrey Sivachenko
Steve Schumacher
Aaron McKenna
Kristian Cibulskis
Carrie Sougnez
Petar Stojanov
Lihua Zhou

Belfler-DFCI (MDACC)
Juinhua Zhang
Spring Liu
Sachet Shukla
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




Outline

I. Current snapshot of ingested data

II. Dashboards for PANCANCER dataset in 3/21

III. Sneak peek at analysis reports

Data Dashboard: 2012_03_21

PANCANCER Method
simple aggregation
of all disease types

Need more sophistication?

Tumor	BCR	Clinical	CN	Methylation	mRNA	mRNAseq	miR	miRseq	MAF
BLCA	89	65	35	78	0	0	0	54	28
BRCA	864	844	781	808	529	751	0	781	507
CESC	99	12	36	0	0	0	0	8	36
COADREAD	591	591	564	584	224	83	0	255	224
DLBC	27	0	0	0	0	0	0	0	0
GBM	596	561	537	287	542	0	491	0	276
HNSC	294	255	165	292	0	13	0	89	0
KIRC	502	502	489	500	72	469	0	463	327
KIRP	135	84	43	117	16	14	0	16	0
LAML	202	200	0	192	0	179	0	187	199
LGG	144	140	80	0	27	0	0	30	0
LIHC	84	47	53	0	0	17	0	28	0
LNNH	2	0	0	0	0	0	0	0	0
LUAD	372	273	205	347	32	106	0	95	147
LUSC	290	272	211	282	154	220	0	202	178
OV	592	580	547	551	568	0	564	46	316
PAAD	48	0	14	30	0	0	0	0	0
PRAD	153	0	82	153	0	0	0	63	0
SKCM	253	0	0	240	0	0	0	0	0
STAD	178	166	132	133	0	57	0	123	133
THCA	274	73	85	230	0	0	0	45	0
UCEC	462	424	363	373	54	266	0	359	239
PANCANCER	6251	5089	4422	5197	2218	2175	1055	2844	2610

Data Dashboard: 2012_03_21

PANCANCER Method
simple aggregation
of all disease types

Need more sophistication?

Tumor	BCR	Clinical	CN	Methylation	mRNA	mRNAseq	miR	miRseq	MAF
BLCA	89	65	35	78	0	0	0	54	28
BRCA	864	844	781	808	529	751	0	781	507
CESC	99	12	36	0	0	0	0	8	36
COADREAD	591	591	564	584	224	83	0	255	224
DLBC	27	0	0	0	0	0	0	0	0
GBM	596	561	537	287	542	0	491	0	276
HNSC	294	255	165	292	0	13	0	89	0
KIRC	502	502	489	500	72	469	0	463	327
KIRP	135	84	43	117	16	14	0	16	0
LAML	202	200	0	192	0	179	0	187	199
LGG	144	140	80	0	27	0	0	30	0
LIHC	84	47	53	0	0	17	0	28	0
LNNH	2	0	0	0	0	0	0	0	0
LUAD	372	273	205	347	32	106	0	95	147
LUSC	290	272	211	282	154	220	0	202	178
OV	592	580	547	551	568	0	564	46	316
PAAD	48	0	14	30	0	0	0	0	0
PRAD	153	0	82	153	0	0	0	63	0
SKCM	253	0	0	240	0	0	0	0	0
STAD	178	166	132	133	0	57	0	123	133
THCA	274	73	85	230	0	0	0	45	0
UCEC	462	424	363	373	54	266	0	359	239
PANCANCER	6251	5089	4422	5197	2218	2175	1055	2844	2610

+789

In just 2 weeks
since last data run

Data Standardization Run Status

	Pipeline Dataset	Not Available	Available	InProcess	Successful	Unsuccessful
13	Merge_exon__huex_1_0_st_v2__lbl_gov__Level_3__segmented_as_firma__data	0	0	0	0	1
20	Merge_mirnaseq__illuminaga_mirnaseq__bcgsc_ca__Level_3__miR_gene_expression__data	0	0	0	0	1
22	Merge_mirnaseq__illuminahiseq_mirnaseq__bcgsc_ca__Level_3__miR_gene_expression__data	0	0	0	0	1
28	Merge_rnaseq__illuminaga_rnaseq__bcgsc_ca__Level_3__exon_expression__data	0	0	0	0	1
29	Merge_rnaseq__illuminaga_rnaseq__bcgsc_ca__Level_3__gene_expression__data	0	0	0	0	1
30	Merge_rnaseq__illuminaga_rnaseq__bcgsc_ca__Level_3__splice_junction_expression__data	0	0	0	0	1

85% successful

6 failures of 40 datasets :
mostly RNA-Seq, miR-Seq

Issues : hg18 vs hg19?
miRbase v13, v16? (LCD, drops samples)
platform differences

Analyses Run Status

	Pipeline	NotRunnable	Runnable	InProcess	Successful	Unsuccessful
1	Aggregate_Clusters	0	1	0	0	0
2	CopyNumber_GeneBySample	0	0	0	1	0
3	CopyNumber_Gistic2	0	0	0	0	1
4	CopyNumber_Preprocess	0	0	0	1	0
5	Correlate_CopyNumber_vs_miR	0	0	0	1	0
6	Correlate_CopyNumber_vs_mRNA	0	0	0	1	0
7	Correlate_Methylation_vs_mRNA	0	0	0	0	1
8	Methylation_Clustering_CNMF	1	0	0	0	0
9	Methylation_Preprocess	0	0	0	0	1
10	miRseq_Clustering_CNMF	1	0	0	0	0
11	miRseq_Clustering_Consensus	1	0	0	0	0
12	miRseq_Preprocess	1	0	0	0	0
13	miR_Clustering_CNMF	0	0	0	1	0
14	miR_Clustering_Consensus	0	0	0	1	0
15	miR_FindDirectTargets	0	0	0	1	0
16	mRNAseq_Clustering_CNMF	1	0	0	0	0
17	mRNAseq_Clustering_Consensus	1	0	0	0	0
18	mRNAseq_Preprocess	0	0	0	0	1
19	mRNA_Clustering_CNMF	0	0	0	0	1
20	mRNA_Clustering_Consensus	0	0	0	0	1
21	mRNA_Preprocess_Median	0	0	0	1	0
22	Mutation_Assessor	1	0	0	0	0
23	Mutation_Significance	0	0	0	0	1
24	Pathway_FindEnrichedGenes	1	0	0	0	0
25	Pathway_Paradigm_Expression	0	0	0	0	1
26	Pathway_Paradigm_Expression_CopyNumber	0	0	0	0	1
27	Pathway_Paradigm_Lite	0	0	0	1	0
	Total	8	1	0	9	9

This table generated on Thu Apr 5 11:53:45 2012

9 good,
9 bad
9 not run

Issues:

- Size -> memory exhaustion
- lurking configuration assumptions, like 1 tumor/platform
- hg18 vs hg19
- etc

Analysis Overview for PANCANCER

Maintained by TCGA GDAC Team *(Broad Institute/Dana-Farber Cancer Institute/Harvard Medical School)*

Overview

Introduction

Summary

Note: These results are offered to the community as an additional reference point, enabling a wide range of cancer biologists, clinical investigators, and genome and computational scientists to easily incorporate TCGA into the backdrop of ongoing research. While every effort is made to ensure that Firehose input data and algorithms are of the highest possible quality, these analyses have not been reviewed by domain experts.

Results

- *Clustering Analyses*
  - **Clustering of miR expression: consensus NMF**
    View Report | We filtered the data to *150* most variable miRs. Consensus NMF clustering of *564* samples and *150* miRs identified *3* subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.
  - **Clustering of miR expression: consensus hierarchical**
    View Report | We filtered the data to *150* most variable miRs. Consensus average linkage hierarchical clustering of *564* samples and *150* miRs identified *3* subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.
- *Correlation Analyses*
  - **Correlations between copy number and mRNA expression**
    View Report | The correlation coefficients in 10, 20, 30, 40, 50, 60, 70, 80, 90 percentiles are -0.00404, 0.04782, 0.104, 0.17178, 0.2493, 0.3319, 0.40534, 0.47608, 0.55744, respectively.
  - **Correlations between copy number and miR expression**
    View Report | The correlation coefficients in 10, 20, 30, 40, 50, 60, 70, 80, 90 percentiles are -0.0365, -0.0124, 0.0053, 0.0249, 0.0542, 0.0968, 0.1851, 0.2803, 0.4009, respectively.
- *Other Analyses*
  - **Identification of putative miR direct targets**
    View Report | This pipeline use a relevance network approach to infer putative miR:mRNA regulatory connections. All miR:mRNA pairs that have correlations < *-0.3* and have predicted interactions in three sequence prediction databases (Miranda, Pictar, Targetscan) define the final network.

Methods & Data

Input

- *Firehose Directory* = `/xchip/cga1/tcga_gdac_firehose_output`
- *Run Prefix* = `analyses__2012_03_21`
- *Summary Report Date = Fri Mar 30 18:05:54 2012*

Sneak Peek
Current Run

http://gdac.broadinstitute.org/runs/tmp/PANCANCER_2012_03_21/

PANCANCER: Correlations between copy number and mRNA expression

Maintained by John Zhang *(Dana-Farber Cancer Institute)*

Overview

Introduction

Summary

The correlation coefficients in 10, 20, 30, 40, 50, 60, 70, 80, 90 percentiles are -0.00404, 0.04782, 0.104, 0.17178, 0.2493, 0.3319, 0.40534, 0.47608, 0.55744, respectively.

Results

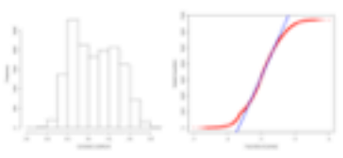
Correlation results

Number of genes and samples used for the calculation are shown in Table 1. Figure 1 shows the distribution of calculated correlation coefficients and quantile-quantile plot of the calculated correlation coefficients against a normal distribution. Table 2 shows the top 20 features ordered by the value of correlation coefficients.

Table 1. Counts of mRNA and number of samples in copy number and expression data sets and common to both

Category	Copy number	Expression	Common
Sample	4422	1645	1500
Genes	29390	17815	15551

Figure 1. Summary figures. Left: histogram showing the distribution of the calculated correlations across samples for all Genes. Right: QQ plot of the calculated correlations across samples. The QQ plot is used to plot the quantiles of the calculated correlation coefficients against that derived from a normal distribution. Points deviating from the blue line indicate deviation from normality.



GET FULL TABLE

Table 2. Top 20 features (defined by the feature column) ranked by correlation coefficients

feature	r	p-value	q-value	chrom	start	end	geneid
LSM1	0.8567	0	0	8	38140014	38153183	27257
CLNS1A	0.8302	0	0	11	77004847	77026495	1207
BRF2	0.8176	0	0	8	37820558	37826569	55290
WHSC1L1	0.8175	0	0	8	38251717	38358947	54904
ASH2L	0.8147	0	0	8	38082223	38116216	9070
RPS6KB1	0.8106	0	0	17	55325225	55382569	6198

http://gdac.broadinstitute.org/runs/tmp/PANCANCER_2012_03_21/