Functionally Enigmatic Genes in Cancer: Are We Looking Under the Lamppost for the Lost Keys?

Alexandra Maertens, Vy Tran, Mikhail Maertens, Andre Kleensang, Thomas Luechtefeld, Thomas Hartung and Channing Paller
Center for Alternatives to Animal Testing, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

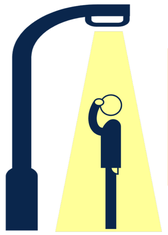
Vy Tran - presenter

I am a 3rd year PhD student in bioinformatics with experiences in cancer biology.
Email: vtran21@jhmi.edu

BACKGROUND

- Despite decades of intense focus, a substantial number of genes implicated in cancer are poorly studied.
- Genes with functionally enigmatic function will likely be missed by any data analysis pipeline, such as enrichment analysis, that depends exclusively on annotations for understanding biological function of transcriptomic studies.

Are we looking under the lamppost for the lost keys?



WE ONLY KNOW THE FUNCTIONS OF A SMALL PORTION OF GENES IN CANCER. THIS CAN AFFECT OUR RESEARCH ON NEW CANCER THERAPIES.

STUDY SUMMARY:

Using large RNA-seq data sets, we showed that a substantial portion of genes statistically associated with cancer biology lack annotations adequate to understanding their role in cancer pathology.

METHODS

DATASETS

RNA-seq data from the Human Pathology Atlas
- 17 different forms of human cancers
- n = 8,000 patients

RNA-seq data from the Human Cancer Atlas
- Prostate adenocarcinoma (PRAD, n = 499 patients)
- Glioma (GBMLGG, n = 1129 patients)
- Colorectal adenocarcinoma (COAD, n = 460 patients)

PUBMED IDs

- PubMed IDs (PMIDS) were identified by querying Entrez with the Entrez GeneID and getting a raw count of PMIDs that mapped to the genes
- Genes with ≤ 50 PMIDS were considered "functionally enigmatic"

ANNOTATIONS

- GO annotation, GO Slim, and Panther pathways were identified using PantherDB
- STRING DB was used for PPI enrichment, visualization of ontologies, and identifying experimental, text-mining, and co-expression relationships amongst proteins

NETWORK ANALYSIS

- We used the WGCNA package and selected the most variant 10,000 genes using median absolute deviance.
- Networks were created based on Topological Overlap Metric based on scale-free topology criterion.

RESULTS

1 Most genes associated clinically with cancer have minimal literature base, and inadequate pathway annotations

- We classified any gene with less than 50 PMIDs as a cut-off "functionally enigmatic" as this likely represents a level at which the literature base is inadequate to fully understand gene function; by this rough metric, the bulk of genes have few articles and the density begins to decrease sharply at 100 (**Figure 1a**).
- Functionally enigmatic genes were less likely to be conserved in eukaryotes and more likely to be primate specific (**Figure 1 b, c**), and were more likely to be unclassified in GO, the narrower and precise annotations in GO Slim, as well as Panther Pathways (**Table 1**).

Figure 1. Density of PubMed IDs (PMIDS) per gene for all prognostic unfavorable genes in various types of cancer from the Human Protein Atlas

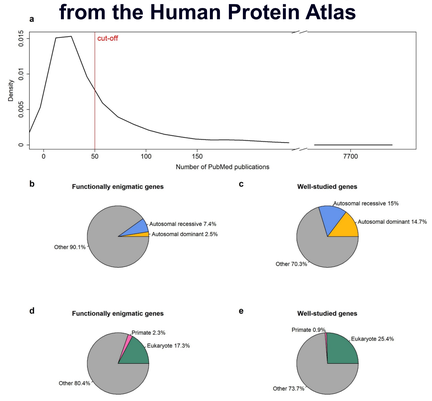


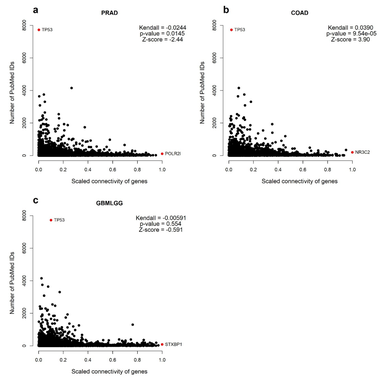
Table 1. Comparison of available annotations for Functionally Enigmatic Genes *vs.* well-studied

	Functionally Enigmatic	Well-studied
GO Unclassified	10.8 %	<0.05 %
GO Slim Unclassified	53%	32%
Panther Pathways Unclassified	93%	70%

2 Genes are not studied in proportion to their importance in network topology or clinical significance

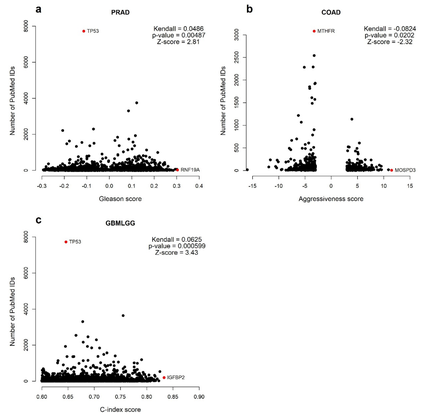
- We expanded our approach to three types of cancer: glioma (**GBMLGG**), colon cancer (**COAD**), and prostate cancer (**PRAD**), using different ways of selecting genes with clinical significance.
- We examined whether scaled connectivity (a metric of whether a gene is acting as a "hub") correlated with depth of literature-base using Kendall rank correlation. Our data suggest that by the metric of scaled connectivity, there is minimal reason to believe that research efforts are focused on the most pertinent genes (**Figure 2 a, b, and c**).

Figure 2. Kendall correlation between scaled connectivity and number of PubMed publications in different cancers



- Additionally, when looking at a correlation between the number of PMIDs and C-index in glioma, Gleason score, or colon cancer aggressiveness, there is again no consistent association between clinical phenotype metric and bibliometric interest (**Figure 3 a, b, c**), although this must be treated with caution as the statistical association of any given gene with a clinical outcome cannot be presumed to be directly equivalent to the magnitude of effect.

Figure 3. Kendall correlation between disease scores and number of PubMed publications for different cancers



3 Functionally enigmatic genes are not evenly distributed across the network

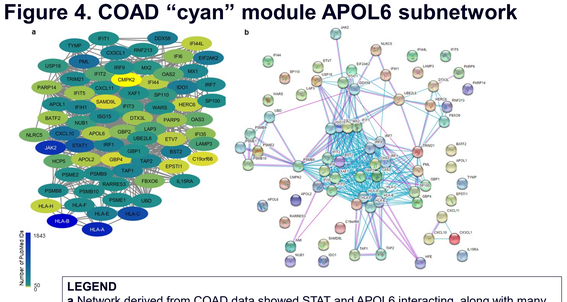
- We clusters genes into modules based on topological similarity, assuming that genes with similar "neighbors" are more likely to have similar function.
- The modules with the highest percentage of functionally enigmatic genes were enriched for terms associated with mRNA splicing, spliceosome, or ncRNA while cell-cycle modules tended to have relatively few functionally enigmatic genes (**Table 2**).
- This suggests that that there are large areas of the "cancer map" - likely representing genes regulated in a coordinated way - where the overwhelming majority of genes have scant attention in the literature.

Table 2. Modules for the PRAD dataset ranked by percentage of functionally enigmatic genes

PRAD Module Color	Total Genes	Functionally Enigmatic (%)	Unmapped in STRING (%)	PPI Enrichment p-value	GO Biological Process Term Description
green	336	80.95	8	< 1.00E-16	mRNA processing
cyan	60	80	4	6.35E-06	translational initiation
magenta	137	77.37	6	< 1.00E-16	RNA splicing
turquoise	3163	74.01	1	< 1.00E-16	intracellular transport
yellow	823	71.81	1	< 1.00E-16	single-organism intracellular transport
brown	974	70.12	1	< 1.00E-16	chromatin modification
salmon	72	69.45	4	4.28E-10	positive regulation of cellular protein metabolic process
blue	1451	67.26	1	< 1.00E-16	cell morphogenesis involved in differentiation
lightcyan	40	65	4	5.25E-14	muscle structure development
grey60	40	62.5	2	< 1.00E-16	defense response to virus
red	285	62.11	0	1.75E-12	response to hormone
black	269	60.45	1	< 1.00E-16	tissue development
midnightblue	57	50.87	3	< 1.00E-16	vasculature development
purple	127	50.39	1%	< 1.00E-16	extracellular matrix organization
pink	234	43.59	2	< 1.00E-16	immune response
greenyellow	103	40.73	2%	< 1.00E-16	response to organic cyclic compound
tan	90	34.44	1%	< 1.00E-16	cell cycle

- To get a sense of the kind of biology likely missed, we have highlighted a handful of the genes that were functionally enigmatic. Within the COAD data set network, APOL6 had the highest absolute ranking for aggressiveness, yet it has relatively few PMIDS. Based on the annotations the interacting genes, APOL6 appears to be associated with the STAT1 pathway (**Figure 4a**).
- However, APOL6 was not shown as connected to the STAT1 pathway in the STRING annotation database (**Figure 4b**).
- Indeed, we were able to verify that APOL6 expression was strongly correlated with STAT1 in colon cancer (Spearman's rank correlation 0.67, *p*-value 7.99e-54) and all other cancers with expression data sets greater than 400 patients after adjusting for tumor purity.

Figure 4. COAD "cyan" module APOL6 subnetwork



LEGEND
a Network derived from COAD data showed STAT and APOL6 interacting, along with many other genes with relatively few PMIDS (indicated as node color).
b PPI network from STRING at medium confidence level.
purple edges: experimental data
blue edges: databases

CONCLUSION

- A substantial number of genes implicated in cancer are relatively poorly studied and will likely be missed by any data analysis pipeline.
- There is no indication that the amount of research - indicated by number of publications - is correlated to any objective metric of gene significance.
- Poorly studied genes are more likely to be primate-specific and less likely to have a Mendelian inheritance pattern, tend to cluster in some biological processes and not others.