Cancer Letters 356 (2015) 479–482

Contents lists available at ScienceDirect

Cancer Letters

journal homepage: www.elsevier.com/locate/canlet

Original Articles

Big genes are big mutagen targets: A connection to cancerous, spherical cells?

Michele L. Parry [a], Michael Ramsamooj [a], George Blanck [a,b,*]

[a] Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL, USA
[b] Immunology Program, Moffitt Cancer Center and Research Institute, Tampa, FL, USA

ARTICLE INFO

Article history:
Received 23 July 2014
Received in revised form 25 August 2014
Accepted 23 September 2014

Keywords:
TCGA
Spherical cells
Flat cells
Tumor suppressor proteins
Oncoproteins

ABSTRACT

We determined the most commonly mutated genes in five cancer genome atlas (TCGA) datasets. Many of these genes were extraordinarily large, as are many cancer fusion gene partners. And many of these genes had cytoskeletal related functions. We further determined that these genes were distributed into high and low frequency mutation groups largely according to overall rate of gene-occurrence in the high and low mutation frequency groups, as was also the case with common metastasis and tumor suppressor genes. Oncoproteins were selectively mutated in the low mutation frequency groups in colon and lung datasets. Thus, genes that have very large coding regions and may impact the cytoskeleton are more commonly mutated than are common metastasis and tumor suppressor genes in both high and low frequency mutation groups. These analyses raise questions related to cell shape: (i) Are cancer cells often spherical because cytoskeletal-related proteins are large mutagen targets? (ii) Is drug-resistance facilitated by relatively common mutant proteins that lead to round cells, with altered cell physiology or reduced surface to volume ratios that could reduce intra-cellular drug concentrations?

Introduction

Despite models of cancer progression that invoke the process of "one mutation → cancer hallmark" [1,2], there are virtually no cases of regulatory oncoproteins or regulatory tumor suppressor proteins that can be exclusively connected with one cancer hallmark or another [3]. We recently noted that classical tumor suppressor proteins, known to regulate the onset of S-phase, are larger than "tumor metastasis suppressor" proteins [3], and that bioinformatics approaches do not indicate any biochemical or molecular, mechanistic distinctions between these two groups of proteins, with the possible exception of a connection between the classical tumor suppressor set and BCL2 [3]. To further address the issue of gene size and mutation susceptibility, we determined the most commonly mutated genes in five cancer genome atlas (TCGA) datasets, and compared these genes with well-studied tumor suppressor and metastasis suppressor proteins in data subsets representing relatively low mutation frequencies.

Materials and methods

Datasets representing exome sequencing and mutation calling were downloaded from the TCGA data portal and further processed using Microsoft Excel files. The Excel files and the detailed processing steps, including macro development, are in the supplementary online material (SOM). We provide an example description here, for obtaining lists of the most frequently mutated genes in the different cancer datasets, but the entire processing approach for this article is detailed as bullet points in the SOM.

To create the files named, *mutfreqdet.cancertype*, representing the accumulated data for Tables 3 and 4, we determined the number of times any one gene has been mutated among the entire collection of samples for a particular cancer type. On the first sheet of an Excel file, we copied column A (HUGO symbol) of the original TCGA dataset (downloaded from the TCGA download portal) and pasted the data into columns A and B on the new sheet, i.e., column A from the original TCGA dataset is duplicated in two adjacent columns in the new sheet. This new sheet is labeled "Gene mutation freq". The duplicate values were removed from column A, and column B still contains column A from the original TCGA dataset. Each column was individually sorted (A–Z). In column C, we entered " = COUNTIF (range of column B, Ax)". Cursor "drag down" was used to extend the series down the entire length of column A. Column C was labeled as frequency of gene mutations among the entire collection of samples. Sheet 2 is created and is labeled as "gene mutation frequency sorted". We then copied column A of Sheet 1 and pasted the information into column A of Sheet 2. We copied column C of Sheet 1 and paste (values only) into column B of Sheet 2. Column B was then sorted (Z–A) with "expanded selection". To create Excel files in SOM labeled as *Parry SOM Table 3,4 ca type*, in Sheet 1 (top 25 most freq gene muts), we copied column A and B entries of the top 25 genes from *mutfreqdet.cancertype* (file above) and pasted these data into a new Excel file. We obtained the number of exons affected for each of the genes in column A from the original TCGA dataset and pasted the number into column C. We obtain the total number of exons of genes in column A from the hg19 version of the human genome database (genome.ucsc.edu) and entered these numbers into column D. The ratio of exons affected to the total number of exons was entered into column E. The amino acid size for all coding regions was obtained from http://www.uniprot.org and entered into column F. Protein function information was placed in column G, H, and I, for each gene by copying the information from columns BB, BC, and BD (GO_Biological_Process, GO_Cellular_Component, GO_Molecular_Function) in the Excel file that represented the original TCGA dataset downloaded from the TCGA data portal.

* Corresponding author. Tel.: +813 974 9585; fax: 813-974-7357.
E-mail address: gblanck@health.usf.edu (G. Blanck).

http://dx.doi.org/10.1016/j.canlet.2014.09.044
0304-3835/

Table 1
Ratio of silent mutations to AA changes.

TCGA dataset	Mutation frequency group	Average number of AA changes among 25 samples	Ratio of silent mutations to AA changes
Bladder	High	711	0.266
	Low	92	0.247
Colon	High	2683	0.262
	Low	51	.224
Glioblastoma	High	132	.257
	Low	30	.221
Lung	High	2072	.201
	Low	33	.187
Stomach	High	4234	.213
	Low	46	.209

See Excel files labeled "Parry SOM hi freq cancer type" OR "Parry SOM lo freq cancer type" in the SOM; access Excel file sheets, e.g., "silent vs. non-silent".

T-test results were entered into Sheet 2 of this file as follows: column A: top 25 AA numbers from Sheet 1; column B: AA numbers for the metastasis suppressor and classical tumor suppressor coding regions; and column E: Excel based function "TTEST" as indicated in the file. In column C, we provided AA numbers for the largest 500 human coding regions for comparison, but these values are not discussed further in the text.

Results

To distinguish and characterize low frequency mutation subsets, from among the five TCGA datasets, we determined the number of genes mutated (in coding exons) among the TCGA samples with the top 25 number of mutations and the number of genes mutated among the samples representing the 25 lowest number of mutations. The average number of mutated genes in each group, along with the ratio of silent mutations to amino acid (AA) substitutions in each group, is shown in Table 1 (supplementary online material, SOM). While the ratio of silent mutations to AA altering mutations shows a slight trend favoring AA substitutions in the low frequency set, the ratios of the high and low frequency sets are almost identical. Presumably the slight trend favoring AA substitutions in the low frequency set represents the selection for AA substitutions that drive cancer, but the subsets indicated lack the statistical power to verify this possibility.

We next determined the ratios of the total number of mutated genes, the number of mutated metastasis suppressor and tumor suppressor genes, from among a common set of metastasis and classical tumor suppressor genes [3], and the number of mutated oncoproteins in the high and low mutation frequency groups (Table 2; SOM), which again indicated a slight trend favoring metastasis and tumor suppressor proteins in the low frequency group. Note that in every dataset, the ratio of mutated, metastasis and tumor suppressor proteins in the high and low frequency sets is just slightly lower than the ratio of the total numbers of mutated genes. However, statistical tests do not indicate a significant difference for this dataset.

Table 2
Ratios of high to low frequency mutation groups, for average number of mutated genes in the various gene sets.

TCGA dataset	All mutated genes	Metastasis and tumor suppressor set	Oncoproteins	Oncoprotein *p*-value
Bladder	7.7	7.5	6.5	ND
Colon	53.0	26.4	**9.1**	(**$p < 0.02$**)
Glioblastoma	4.4	3.1	3.8	ND
Lung	62.9	30.0	**8.0**	(**$p < 0.006$**)
Stomach	91.3	56.1	144.0	ND

See Excel files labeled "Parry SOM hi freq cancer type" OR "Parry SOM lo freq cancer type" in the SOM; access Excel file sheets, e.g., "ts and op".
Bold indicates significantly distinct over-representation of oncoproteins in the low frequency mutation group.

Table 3
Very large coding regions, encoding proteins related to the cytoskeleton, among the most frequently mutated genes in the five TCGA datasets.

TCGA dataset	Average coding region size for top 25 most frequently mutated genes	*p*-value for average size of tumor and metastasis suppressor set vs. top 25 most frequently mutated genes	Number of cytoskeletal-related proteins among top 25 mutated proteins
Bladder	6296	<.001	10
Colon	5567	<.005	11
Glioblastoma	5508	<.007	10
Lung	5403	<.004	11
Stomach	6973	<.001	9

See Materials and Methods and specific SOM files indicated therein.

The ratio of oncoprotein mutations shows some variability, but in the case of the colon and lung datasets, there is a significant selection for oncoprotein mutations in the low frequency group. Thus, the overall conclusion from the data of Tables 1 and 2 is that both frequency groups have largely similar mutation profiles with regard to AA (vs. silent) mutation selectivity and selectivity for the overall set of tumor (and metastasis) suppressor proteins.

The above assessments establish a low frequency mutation group that according to at least two parameters, ratio of silent mutations to AA substitutions and ratio of metastasis and tumor suppressor proteins, does not have mutation-selectivity processes that differ significantly from a high frequency mutation group.

We next determined the 25 most commonly mutated genes in the entire collection of samples in the five datasets. In all cases, the 25 most frequently mutated genes had average coding region sizes significantly larger than the average coding region sizes of the common tumor suppressor and metastasis suppressor coding regions (Table 3, SOM).

We then determined the distribution of the most commonly mutated genes among the high and low frequency mutation groups. Overall, the distribution mirrored the distribution of the overall collection of mutated genes and the distribution of metastasis and suppressor genes (Table 4; Fig. 1; SOM). Thus, most of the genes in

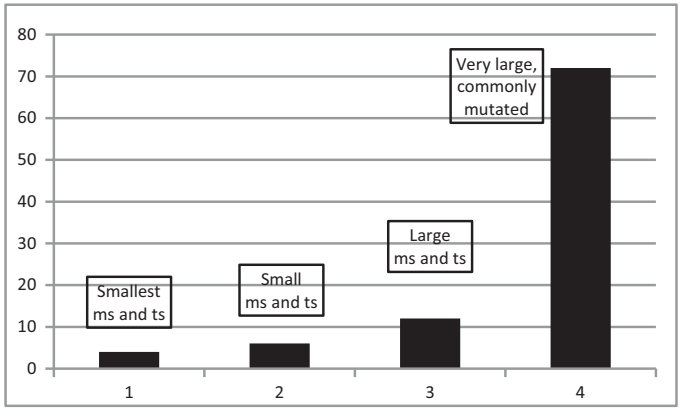

Fig. 1. Frequencies of large and small genes in the low frequency groups. Occurrences of any of 10 "smallest" metastasis (ms) and classical tumor suppressor (ts) genes, as defined by coding region size (number of AA) (Excel file, "Parry SOM tu sup, met sup, oncop lists" in the SOM); or any of 10 "small" metastasis and classical tumor suppressor genes, representing the next smallest group, after the "smallest" group (SOM); or any of the 10 largest metastasis and classical tumor suppressor genes (SOM); or any of 10 largest of the top 25 most commonly mutated genes (Excel file "Parry SOM, largest 10 mutated genes", in the SOM) → IN THE LOW mutation frequency groups ($p < 0.03$, for the very large, commonly mutated genes versus any other category or ANY grouping of the other categories). In the case below, about 70 very large genes are mutated among 125 samples in the overall, low frequency group collections for the five cancer datasets, first defined in Table 4.