A Novel Approach to Evaluating Cancer Driver Gene Mutation Densities: Cytoskeleton-related Gene Candidates

TIMOTHY J. FAWCETT[1], MICHELE L. PARRY[2] and GEORGE BLANCK[3]

[1]Department of Chemical and Biomedical Engineering, College of Engineering, Research Computing and
[2]Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL, U.S.A.;
[3]Immunology Program, Moffitt Cancer Center and Research Institute, Tampa, FL, U.S.A.

Abstract. *Background: Oncoprotein genes are over-represented in statically defined, low mutation-frequency fractions of cancer genome atlas (TCGA) datasets, consistent with a higher driver mutation density. Materials and Methods: We developed a "continuously variable fraction" (CVF) approach to defining high and low mutation-frequency groups. Results and Conclusion: Using the CVF approach, an oncoprotein set was shown to be associated with a TCGA, low mutation-frequency group in nine distinct cancer types, versus six, for statically defined sets; and a tumor-suppressor set was over-represented in the low mutation-frequency group in seven cancer types, notably including BRCA. The CVF approach identified single-mutation driver candidates, such as BRAF V600E in the thyroid cancer dataset. The CVF approach allowed investigation of cytoskeletal protein-related coding regions (CPCRs), leading to the conclusion that mutation of CPCRs occurs at a statistically significant, higher density in low mutation-frequency groups. Supporting online material for this article can be found at www.universityseminarassociates.com/Supporting_online_material_for_scholarly_pubs.php*

The cancer genome atlas (1) (http://cancergenome.nih.gov/) and related cancer DNA sequence databases (2) have provided an opportunity for exploiting statistical power to discover commonalities in the genetics, and possibly the advent of cancer. We recently made a number of conclusions (3) by segregating five cancer genome atlas datasets into high and low mutation-frequency groups: (i) coding region

Correspondence to: Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, 12901 Bruce B. Downs Bd. MDC7, Tampa, FL 33612, U.S.A. Tel: +1 813 9749585, Fax: +1 813 9747357, e-mail: gblanck@health.usf.edu

Key Words: TCGA, mutation density, oncoproteins, tumor suppressor proteins, cytoskeletal proteins, mutation frequency.

mutagenesis is largely random, with almost no difference in the occurrence of silent *versus* amino acid changes in the high and low mutation-frequency groups; (ii) the vast majority of coding regions mutations occur in very large coding regions, consistent with the significant stochastic aspect to coding region alterations; (iii) a disproportionate representation of tumor suppressor proteins in the low mutation-frequency group could not be established; (iv) a disproportionate representation of oncoproteins in the low-mutation-frequency groups could be established for only two of the five TCGA datasets. In those two cases, COAD and LUAD, the disproportionate association of oncoproteins with the low mutation-frequency groups relied on the likelihood that oncoproteins often represent degenerate signaling pathways (4) and could, thus, be grouped for increased statistical power, *i.e.* an oncoprotein set, rather than individual oncoproteins, was observed to associate with the COAD and LUAD low mutation-frequency groups.

In the present report, the basic algorithm of dividing the TCGA datasets into high and low mutation-frequency groups has been encoded (scripted), thus allowing the development of a continuously variable definition of "high" or "low" frequency groups *via* the increased computing power. With a variable definition, an over-representation of mutations in any low frequency group (in comparison to the analogous high frequency group) can be detected, rather than relying on such detection in one arbitrarily defined low frequency group. This computational approach has provided for a substantial increase in detection of candidate driver mutations.

Materials and Methods

Mutation Annotation Format (MAF) files from the TCGA database containing only somatic mutations detected in tumor samples were downloaded as tab-delimited files from the TCGA data portal. All 3,158,693 mutations detected within 6,482 tumor samples contained in 26 MAF files spanning 24 TCGA datasets were inserted into

Table I. *Ratio of high mutation-frequency group, oncoprotein and tumor suppressor protein occurrence, to low mutation-frequency group, oncoprotein and tumor suppressor protein occurrence, using approach of Parry et al (3).*

TCGA dataset	Ratio of number of total mutated coding regions in high frequency-mutation group to low frequency-mutation group	Ratio of number of oncoprotein coding region mutations in high frequency-mutation group to low frequency-mutation group	*p*-Value for disproportionate representation of oncoprotein coding regions in the low frequency-mutation group	Ratio of number of tumor suppressor coding region mutations in high frequency-mutation group to low frequency-mutation group	*p*-Value for disproportionate representation of tumor suppressor protein coding regions in the low frequency-mutation group
ACC	9.16	8.5	NS	9.33	NS
BLCA	7.76	7	NS	7.18	NS
BRCA	190.28	(infinity)	NA	94	NS
CESC	3.64	1.66	NS	2.05	NS
COAD	53.00	9.09	0.016	26.7	NS
GBM	4.41	3	NS	3.71	NS
HNSC	26.69	13.5	NS	15.57	NS
KIRC	14.01	5	NS	2.6	0.008
KIRP	3.44	0.8	NS	4.25	NS
LAML	12.71	1.11	0.008	(infinity)	NA
LGG	45.55	6.5	NS	25.66	NS
LIHC	680.25	138.28	NS	321.42	NS
LUAD	51.07	7.9	0.006	25	NS
LUSC	6.84	3	NS	5.53	NS
OV	8.82	1.66	NS	3	NS
PRAD	10.58	6	NS	9.5	NS
READ	9.72	4.18	NS	6.69	NS
SKCM	131.67	16.09	0.001	87.66	NS
STAD	79.55	122	NS	34.2	NS
THCA	11.42	1.46	0.0002	(infinity)	NA
UCEC	17.45	3	0.0003	4.06	0.001
UCS	5.08	1.57	NS	3.2	NS

Driver mutations would be expected to be over-represented in the low mutation-frequency groups. Thus, driver mutations would be indicated by a lower ratio than the ratio of overall mutations, for a given dataset. See SOM Table 1 data for details, including references for the list of oncoproteins and tumor suppressor proteins included in the above analyses. "Infinity" refers to absence of a member of the oncoprotein or tumor suppressor protein sets in the low frequency mutation group. NS, Not significant; NA, not applicable.

various tables of a single PostgreSQL 9.3.5 relational database. Queries or database views to select the disease, HUGO symbol of mutated gene, gene type (oncoprotein, tumor suppressor protein, cytoskeletal protein, or none), codon change, start position, and end position of each mutation allow the count of total mutations as well as the count of mutations of a certain gene type, gene, and/or gene with a specific codon change for each dataset. The queries have also been designed to select only distinct mutations, keeping in mind that each dataset (cancer type) within the TCGA collection can be compared to one or more matched normal samples, potentially (and erroneously) giving rise to multiple records for identical mutations within a given tumor sample.

After the total mutation counts have been determined, the individual datasets are sorted by their number of total mutations and separated into high and low mutation-frequency groups containing N samples (TCGA barcodes), with N ranging from 2 (minimum required for statistical testing) to half the total number of samples (maximum without duplicating samples). After the samples are sorted, the ratio of occurrence of particular gene types (oncoprotein, tumor suppressor protein, cytoskeletal protein) as well as each gene with a specific codon change, relative to the total number of mutations is calculated.

However, no mutations were further considered unless the mutation occurred at a minimum of 25-times in the pan-cancer database. The *p*-value from a two-tailed two-sample *t*-test, assuming unequal variances, is used to determine the level of significance at which the mean ratio in the high mutation-frequency group is not equal to the mean ratio in the low mutation-frequency group for any mutation that occurs more than 24 times in the database. Matlab® R2014a was used to query the PostgreSQL database, sort tumor samples, and calculate the reported *p*-values. All code is available in the supporting online material (SOM) (www.universityseminarassociates.com/Supporting_online_material_for_scholarly_pubs.php).

Ethics statement. The corresponding author submitted and received approval for TCGA-use proposal, although all data in this report are publicly available.

Results

Initial work was based on single, arbitrary definitions for high and low mutation-frequency groups (3), reproduced herein, *via* scripting the initial algorithm; and with additional

Table II. *TCGA datasets with oncoprotein set or tumor suppressor set associations.*

TCGA dataset	Oncoprotein set association	Lowest *p*-value	Fractional point	Association with High or Low frequency-mutation group	Tumor suppressor set association	Lowest *p*-value	Fractional point	Association with High or Low frequency-mutation group
ACC	**Yes**	**0.039**	**0.021**	High	**Yes**	**0.0052**	**0.087**	High
BLCA	**Yes**	**0.009**	**0.069**	High	No	NA	NA	
BRCA	**Yes**	**1.34E-6**	**0.073**	High	**Yes**	**5.86E-7**	**0.074**	Low
CESC	No	NA	NA		No	NA	NA	
COAD	Yes	1.29E-5	0.309	Low	**Yes**	**3.05E-5**	**0.031**	Low
GBM	No	NA	NA		Yes	0.002	0.041	High
HNSC	**Yes**	**0.029**	**0.467**	Low	**Yes**	**0.00002**	**0.257**	Low
KIRC	**Yes**	**0.003**	**0.085**	High	Yes	0.00003	0.497	Low
KIRP	No	NA	NA		**Yes**	**0.034**	**0.06**	High
LAML	Yes	0.0007	0.497	Low	**Yes**	**0.019**	**0.324**	High
LGG	No	NA	NA		No	NA	NA	
LIHC	**Yes**	**0.00004**	**0.48**	Low	**Yes**	**0.00006**	**0.049**	High
LUAD	Yes	7.07E-22	0.434	Low	**Yes**	**0.0007**	**0.327**	Low
LUSC	No	NA	NA		**Yes**	**0.044**	**0.016**	High
OV	Yes	0.037	0.07	High	No	NA	NA	
PRAD	No	NA	NA		**Yes**	**0.0002**	**0.073**	High
READ	No	NA	NA		**Yes**	**0.018**	**0.049**	High
SKCM	Yes	3.38E-12	0.425	Low	**Yes**	**2.57E-8**	**0.035**	High
STAD	**Yes**	**1.04E-5**	**0.032**	High	**Yes**	**0.0003**	**0.482**	Low
THCA	Yes	7.83E-20	0.412	Low	**Yes**	**0.0001**	**0.12**	High
UCEC	Yes	4.14E-8	0.340	Low	Yes	4.96E-8	0.396	Low
UCS	**Yes**	**0.038**	**0.491**	Low	No	0.034	0.052	High

Bold values represent associations detectable with a CVF approach and not detectable with the approach of Parry *et al*. (3).

TCGA datasets (Table I). Results of the scripted version of the algorithm were identical for the previously studied datasets. Several TCGA datasets not previously studied indicated a disproportionate level of oncoprotein coding region mutations (LAML, SKCM, THCA, UCEC) in the low mutation-frequency groups. And for the first time, this approach indicated a disproportionate level of tumor suppressor mutations in previously unstudied datasets (KIRC, UCEC), in the low mutation-frequency groups.

To search for statistical significance of a disproportionate association of oncoprotein and tumor suppressor sets with the low mutation-frequency groups, with the same basic paradigm, but with the CVF strategy, we plotted *p*-values against continuous fractions for definitions of the high and low mutation-frequency groups (Methods; SOM; (www.universityseminarassociates.com/Supporting_online_m aterial_for_scholarly_pubs.php), with results consistent with BRAF as an oncoprotein (Figure 1). We repeated this approach with the previously defined oncoprotein and tumor suppressor sets (3), again keeping in mind extensive signaling pathway degeneracy in cancer (4) (Table II). Results indicated an association of one or the other or both cancer-gene sets with TCGA datasets where the approach using fixed mutation group fractions did not lead to such

detection, illustrating the increased opportunities of the modification represented by the CVF approach. In particular, we detected a significantly increased association of the oncoprotein set, with the low mutation-frequency group, in three additional TCGA datasets: HNSC, LIHC and UCS (Compare Tables I and II). We detected increased association of the tumor suppressor set with the low mutation-frequency group in five additional datasets: BRCA, COAD, HNSC, LUAD and STAD.

To determine whether the CVF approach had the resolving power of detecting a disproportionate association of individual mutations and genes with the low mutation-frequency group, we applied the algorithm to every mutation position in the TCGA database. Results with a *p*<0.01 are indicated in Table III; results with a *p*<0.05 are indicated in Table S3 in the SOM (www.universityseminarassociates.com/Supporting_online_ma terial_for_scholarly_pubs.php); and in Excel files in the SOM. As expected, a number of well-studied mutations and genes, such as BRAF, with the V600E amino acid alternation, were readily detected as disproportionately associated with the low mutation-frequency groups, in particular for the SKCM (Figure 1B and Table III) and THCA datasets (Table III). In addition, numerous mutations were specifically associated with the high mutation-frequency mutation groups, which
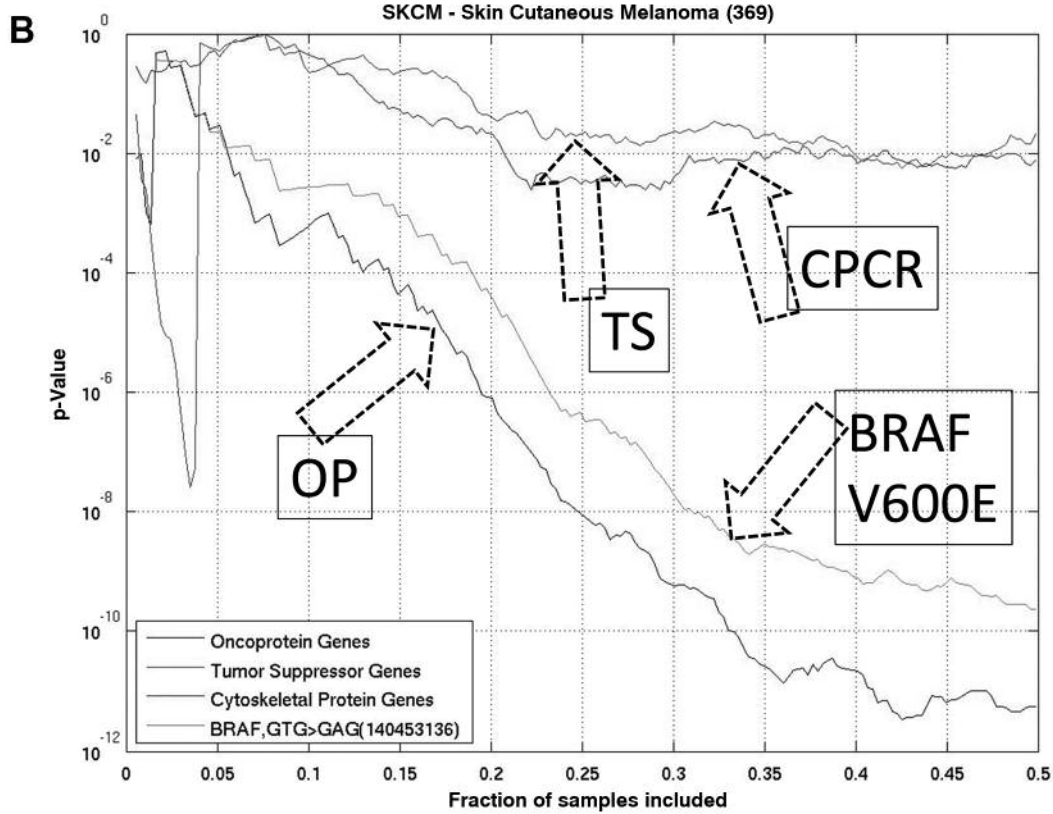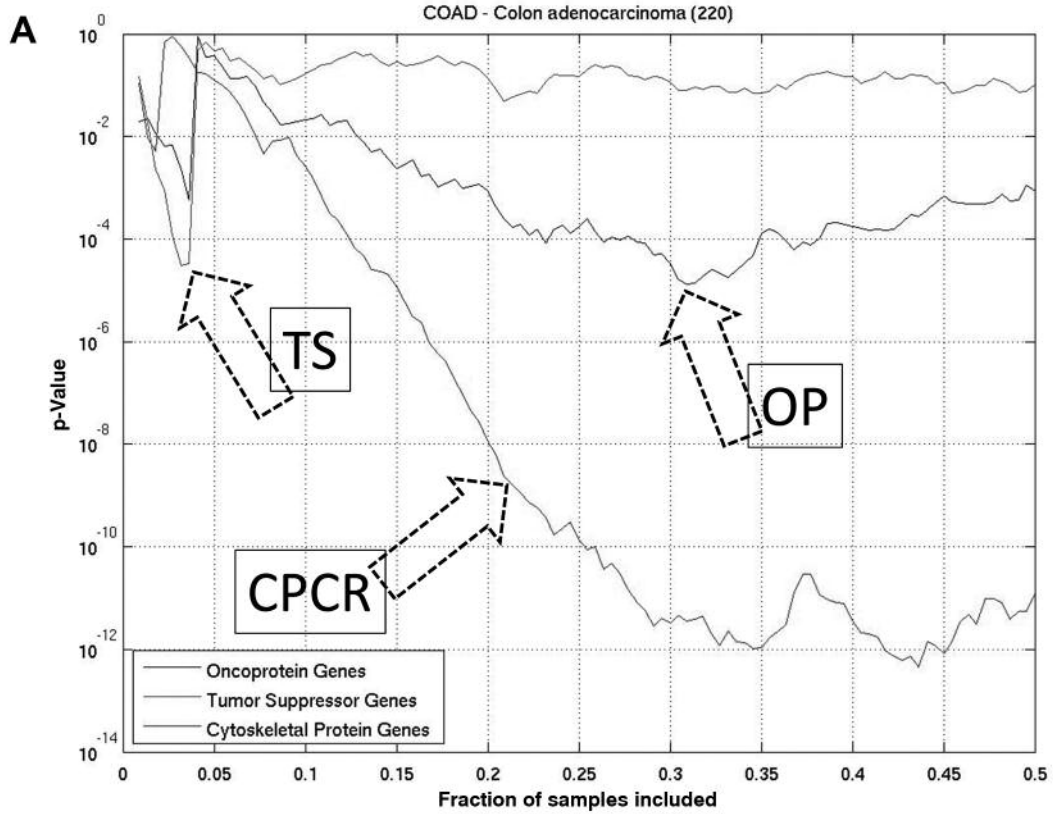
Table III. *Individual mutations disproportionally associated with the high or low mutation-frequency groups with a p-value <0.01.*

TCGA dataset	HUGO symbol, codon change and nucleotide number	Amino acid change	High or low	Pubmed cancer entries as of 12/16/2014
HNSC	PARG,GCA>ACA(51093329)	Ala > Thr	High	85
	TP53,CGG>TGG(7577539)	Arg > Trp	Low	7983
KIRC	EEF1B2,CCG>CCA(207025366)	Pro > Pro (repeat of 3)	High	3
LGG	IDH1,CGT>CAT(209113112)	Arg > His	Low	950
LUAD	CHEK2,AAG>GAG(29091840)	Lys > Glu	Low	913
	KRAS,GGT>GAT(25398284)	Gly > Asp	Low	7581
	KRAS,GGT>TGT(25398285)	Gly > Cys	Low	-
	KRAS,GGT>GTT(25398284)	Gly > Val	Low	-
PRAD	RGPD8,CCT>GCT(113127775)	Pro > Ala	High	0
	EEF1B2,AGC>GGC(207025358)	Ser > Gly	High	3
SKCM	BRAF,GTG>GAG(140453136)	Val > Glu	Low	5344
STAD	TRIM48,TAC>CAC(55035844)	Tyr > His	High	0
	RGPD3,AAC>GAC(107049681)	Asn > Asp	High	0
	UBBP4,ATC>ACC(21731270)	Ile > Thr	High	0
	CR1,CGA>TGA(207787753)	Arg > stop	High	NA (overlap term)
	PGM5,ATC>GTC(70993145)	Ile > Val	High	1
THCA	BRAF,GTG>GAG(140453136)	Val > Glu	Low	-
UCEC	PIK3CA,GAG>AAG(178936091)	Glu > Lys	Low	1682
	PIK3CA,CAT>CGT(178952085)	His > Arg	Low	-
	TRIM48,TAC>CAC(55035844)	Tyr > His	High	0
	HSD17B7P2,AAT>AGT(38654432)	Asn > Ser	High	0
	PTEN,CGA>GGA(89692904)	Arg > Gly	Low	6814

Codons, nucleotide number and amino acid change are indicated; shaded HUGO symbols represent particularly novel cancer associations.

←

Figure 1. *Plot of p-value versus fractional definition of the high and low mutation-frequency groups. p-Value is on y-axis; fraction of TCGA dataset on x-axis. The x-axis represents a continuously variable fraction of the TCGA dataset used to simultaneously define the high and low mutation-frequency groups. For example, at a fraction of 0.2, the mutation-frequency group was defined by the 20% of samples representing the fewest mutations; and by the 20% of samples representing the highest number of mutations in the dataset. Then, a p-value was calculated for the significance of the difference between the ratio of the overall number of mutations in the high frequency group to the low frequency group, in comparison for the number of mutations in the high and low frequency groups for the indicated proteins or groups of proteins. This process first indicated over-representation of oncoproteins in low frequency mutation groups in ref. (3). (A) COAD TCGA dataset; TS (tumor suppressor and metastasis suppressor proteins as a group, as defined in ref. (3)); OP (oncoprotein group as defined in ref. (3)); CPCR (cytoskeletal protein related coding regions as defined in SOM (www.universityseminarassociates.com/Supporting_online_material_for_scholarly_pubs.php) for this article, represented by the following HUGO symbols: TTN, APC, MUC16, NEFH, SYNE1, FAT4, SSPO, PLEC, MUC4, NEB, DNAH11, DNAH8, PCDHAC2, PCDHGC5, FLG, NF1, PCLO, SPTA1, PKHD1, DNAH5, RELN, XIRP2, FAT3, PCDH15, ANK2, DST, FBN2, COL11A1, SPTAN1. (B) Same as (A), but for SKCM TCGA dataset; and with indication of the p-value as a function of high/low frequency fractional definition for the BRAF V600E mutation.*

could represent numerous possibilities including lack of a driver status, *i.e.*, an artifact of high level mutagenesis; or a requirement for cooperation with other mutations for driver status (with such cooperation only occurring when there is high enough level of mutagenesis for "two hits").

Other coding regions, less well connected with the study of cancer, have been significantly associated with the low mutation-frequency mutation groups, representing several different cancer datasets, at $p<0.05$ but $p>0.01$: SKCM, MYO5B,GTC>GCC (nucleotide position, 47363917); HNSC AQP7,AGT>AGA(33385614); ACC KRT8,TCC>GCC (53298675); LIHC; KRT8,TCC>GCC(53298675) (Table S3, SOM).

Large coding regions are particularly susceptible to the stochastic process that plays a large role in the mutational process, as reflected by the TCGA datasets (3). The role of gene size in mutational susceptibility is consistent with large gene size being a significant factor in gene-partner inclusion in cancer fusion genes (5, 6). Cytoskeletal protein related coding regions (CPCRs), many of which are among the largest coding regions in the human genome (3), have long been thought to play a role in cancer development and metastasis, but with contradictory conclusions (7-13). Interestingly, a recent report indicated a CPCR mutation associated with breast cancer metastasis to the lymph node (14).

Table IV. *Ratio of high mutation-frequency mutation group CPCR mutations to low mutation-frequency group CPCR mutations, using approach of Parry et al (3).*

TCGA dataset	Ratio of number of total mutations in high frequency mutation group to low frequency-mutation group	Ratio of number of CPCR mutations in high frequency mutation group to low frequency-mutation group	*p*-Value for disproportionate representation of CPCRs in either the high or low frequency-mutation group
ACC	9.16	12.75	NS
BLCA	7.76	4.07	0.002
BRCA	190.28	145	NS
CESC	3.64	3.05	NS
COAD	53.00	19.43	0.0003
GBM	4.41	3.3	NS
HNSC	26.69	36.42	NS
KIRC	14.01	22.5	0.014
KIRP	3.44	3.14	NS
LAML	12.71	12	NS
LGG	45.55	34.45	NS
LIHC	680.25	910.58	NS
LUAD	51.07	62.38	NS
LUSC	6.84	7.00	NS
OV	8.82	11.8	NS
PRAD	10.58	28.66	0.013
READ	9.72	5.70	0.043
SKCM	131.67	176.80	NS
STAD	79.55	48.93	NS
THCA	11.42	21	NS
UCEC	17.45	12.76	NS
UCS	5.08	6.75	NS

NS, Not significant.

To determine whether CPCRs as a class (3) (SOM: Excel file: list of CPCRs) (www.universityseminarassociates.com/Supporting_online_material_for_scholarly_pubs.php) could represent mutated, candidate driver genes, we determined whether there was disproportionate representation of CPCR genes in the low mutation-frequency groups using the approach of Parry *et al*., with a static definition of high and low mutation-frequency groups (Table IV), and with a CVF approach (Figure 1 and Table V). Results indicated a highly significant, disproportionate presence of CPCR mutations in the low mutation-frequency groups for numerous TCGA datasets.

Discussion

The above results indicate that varying the fraction of samples included in defining high and low mutation-frequency groups, dramatically extends the usefulness of a mutation density-based approach to identifying candidate driver mutations, *i.e.* leads to identification of a greater number of candidate driver mutations. All data-mining-based algorithms used for identification of candidate drivers identify just that: candidates. Furthermore, there is the presumption that empirical work is required for verification of the function of the candidate driver mutation *in*

vivo. However, the above CVF approach readily identifies several known oncoproteins, such as IDH1 and BRAF V600E. Thus, other mutations and genes revealed by the CVF approach, hitherto not considered extensively in oncogenesis, are likely to be revealed as functionally relevant with empirical approaches, for example, MYO5B, discussed above (Table S3).

It is likely, at a low level of statistical significance, that background mutagenesis would be identified by the CVF approach. For example, cells having few mutations could have a higher density of mutations due to DNA replication error rates that are more common (repeated) at one position in the genome, in comparison to cells with a very high mutation rate due to greater exposures to mutagens, for example. The high concentration of mutagens could conceivably overwhelm a propensity for a high replication error rate (with a bias for a particular genome position), and thus such background mutations would be detected as statistically "significantly" associated with the low mutation-frequency group. Indeed, Table S3 (SOM) (www.universityseminarassociates.com/Supporting_online_material_for_scholarly_pubs.php) lists several silent mutations, all of which have a *p*-value<0.05 but greater than 0.01.

It is also likely that a more comprehensive application of statistical tools would enhance the yield from what is

Table V. *TCGA datasets with mutated, cytoskeletal protein-related coding regions (CPCRs) associated with the high or low mutation-frequency groups.*

TCGA dataset	CPCR set association (with the high or low frequency group)?	Lowest *p*-value	Fractional point
ACC	**Yes**	**0.037**	**0.043**
BLCA	Yes	0.0001	0.5
BRCA	No	NA	NA
CESC	No	NA	NA
COAD	Yes	4.43E-13	0.436
GBM	No	NA	NA
HNSC	No	NA	NA
KIRC	Yes	0.0002	0.055
KIRP	No	NA	NA
LAML	**Yes**	**0.0008**	**0.045**
LGG	**Yes**	**0.001**	**0.01**
LIHC	**Yes**	**0.001**	**0.287**
LUAD	**Yes**	**1.19E-5**	**0.251**
LUSC	**Yes**	**0.015**	**0.011**
OV	**Yes**	**0.0002**	**0.077**
PRAD	Yes	6.02E-5	0.5
READ	Yes	0.014	0.222
SKCM	**Yes**	**0.002**	**0.284**
STAD	**Yes**	**0.025**	**0.238**
THCA	**Yes**	**0.0005**	**0.049**
UCEC	**Yes**	**0.007**	**0.324**
UCS	No	NA	NA

Bold values represent associations detectable with a CVF, for definition of high and low frequency-mutation groups, that were not detectable by the Table IV approach of Parry *et al*. (3).

fundamentally a mutation density based approach to identifying cancer driver genes. However, the clear and dramatic identification of positive controls, such as BRAF V600E in the SKCM (melanoma) dataset, indicate that the current statistical analysis is productive.

CPCRs are extensively mutated in many cancers (3), but understanding the role of these mutations is complicated by the large sizes of the CPCR coding regions. Large coding region sizes make routine experimental approaches, for example, DNA transfections, difficult, and thus there is less groundwork available to justify more extensive approaches, such as generating mice with relatively sophisticated genetic engineering features. In addition, past research regarding the cytoskeleton in tumorigenesis has been contradictory (7, 12, 13, 15-17). Thus, the CVF approach offers a third avenue of investigation, for the study of CPCRs that circumvents the impracticality of conventional experimental approaches. In particular, several of the mutations indicated as significantly associated with the low mutation-frequency group at the $p<0.05$ level, *e.g.*, MYO5B and KRT8, are related to the cytoskeleton and cell shape functions.

The above CPCR results raise the question of whether cell-shape changes that accompany a disorganized cytoskeleton are due to the relatively common mutation of very large, genetically vulnerable CPCRs? The above results also raise the question of whether the common spheroid shape of cancer-drug resistant cells (18-25) is traceable to the same genetic vulnerability, leading to cells with reduced surface area to volume ratios and thus cells with lower intracellular drug concentrations?

Interestingly, as noted by inspection of Table IV, three of the TCGA data sets (BLCA, COAD, READ) demonstrate an association of CPCR mutations with the low mutation-frequency groups; and two of the TCGA datasets (KIRC, PRAD) indicate that the CPCRs are associated with the high mutation frequency groups. This latter result raises the question of whether the tumorigenic effects of CPCR mutations in certain cancer types occur only with a high mutation burden, where there is the greater possibility that a CPCR mutation will cooperate with mutations of other groups of proteins, such as conventional oncoproteins or tumor suppressor proteins?

Supporting Online Material

www.universityseminarassociates.com/Supporting_online_material_for_scholarly_pubs.php

Supporting online material can also be obtained by emailing Authors.

References

1   Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N and Sander C: Emerging landscape of oncogenic signatures across human cancers. Nat Genet *45*: 1127-1133, 2013.

2   Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR and Futreal PA: COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res *39*: D945-950, 2011.

3   Parry ML, Ramsamooj M and Blanck G: Big genes are big mutagen targets: A connection to cancerous, spherical cells? Cancer Lett *356*: 479-482, 2015.

4   Long K, Abuelenen T, Pava L, Bastille M and Blanck G: Size matters: sequential mutations in tumorigenesis may reflect the stochastic effect of mutagen target sizes. Genes Cancer *2*: 927-931, 2011.

5   Narsing S, Jelsovsky Z, Mbah A and Blanck G: Genes that contribute to cancer fusion genes are large and evolutionarily conserved. Cancer Genet Cytogenet *191*: 78-84, 2009.

6   Pava LM, Morton DT, Chen R and Blanck G: Unifying the genomics-based classes of cancer fusion gene partners: large cancer fusion genes are evolutionarily conserved. Cancer Genomics Proteomics *9*: 389-395, 2012.

7   Verderame M, Alcorta D, Egnor M, Smith K and Pollack R: Cytoskeletal F-actin patterns quantitated with fluorescein isothiocyanate-phalloidin in normal and transformed cells. Proc Natl Acad Sci USA *77*: 6624-6628, 1980.

8   Verschueren H, Van der Taelen I, Dewit J, De Braekeleer J and De Baetselier P: Metastatic competence of BW5147 T-lymphoma cell lines is correlated with *in vitro* invasiveness, motility and F-actin content. J Leukoc Biol *55*: 552-556, 1994.

9   Carlier MF, Pernier J and Avvaru BS: Control of actin filament dynamics at barbed ends by WH2 domains: from capping to permissive and processive assembly. Cytoskeleton (Hoboken) *70*: 540-549, 2013.

10 Meddens MB, Rieger B, Figdor CG, Cambi A and van den Dries K: Automated podosome identification and characterization in fluorescence microscopy images. Microsc Microanal *19*: 180-189, 2013.

11 Boylan MO, Athanassiou M, Houle B, Wang Y and Zarbl H: Activation of tumor suppressor genes in nontumorigenic revertants of the HeLa cervical carcinoma cell line. Cell Growth Differ *7*: 725-735, 1996.

12 Xu W, Mezencev R, Kim B, Wang L, McDonald J and Sulchek T: Cell stiffness is a biomarker of the metastatic potential of ovarian cancer cells. PLoS One *7*: e46609, 2012.

13 Kim TY, Vigil D, Der CJ and Juliano RL: Role of DLC-1, a tumor suppressor protein with RhoGAP activity, in regulation of the cytoskeleton and cell motility. Cancer Metastasis Rev *28*: 77-83, 2009.

14 Blighe K, Kenny L, Patel N, Guttery DS, Page K, Gronau JH, Golshani C, Stebbing J, Coombes RC and Shaw JA: Whole genome sequence analysis suggests intratumoral heterogeneity in dissemination of breast cancer to lymph nodes. PLoS One *9*: e115346, 2014.

15 Pokorna E, Jordan PW, O'Neill CH, Zicha D, Gilbert CS and Vesely P: Actin cytoskeleton and motility in rat sarcoma cell populations with different metastatic potential. Cell Motil Cytoskeleton *28*: 25-33, 1994.

16 Zachary JM, Cleveland G, Kwock L, Lawrence T, Weissman RM, Nabell L, Fried FA, Staab EV, Risinger MA and Lin S: Actin filament organization of the Dunning R3327 rat prostatic adenocarcinoma system: correlation with metastatic potential. Cancer Res *46*: 926-932, 1986.

17 Kopelovich L, Conlon S and Pollack R: Defective organization of actin in cultured skin fibroblasts from patients with inherited adenocarcinoma. Proc Natl Acad Sci USA *74*: 3019-3022, 1977.

18 Liao J, Qian F, Tchabo N, Mhawech-Fauceglia P, Beck A, Qian Z, Wang X, Huss WJ, Lele SB, Morrison CD and Odunsi K: Ovarian cancer spheroid cells with stem cell-like properties contribute to tumor generation, metastasis and chemotherapy resistance through hypoxia-resistant metabolism. PLoS One *9*: e84941, 2014.

19 Yan YR, Xie Q, Li F, Zhang Y, Ma JW, Xie SM, Li HY and Zhong XY: Epithelial-to-mesenchymal transition is involved in BCNU resistance in human glioma cells. Neuropathology *34*: 128-134, 2014.

20 Wang H, Zhang G, Zhang H, Zhang F, Zhou B, Ning F, Wang HS, Cai SH and Du J: Acquisition of epithelial-mesenchymal transition phenotype and cancer stem cell-like properties in cisplatin-resistant lung cancer cells through AKT/beta-catenin/ Snail signaling pathway. Eur J Pharmacol *723*: 156-166, 2014.

21 Lito P, Rosen N and Solit DB: Tumor adaptation and resistance to RAF inhibitors. Nat Med *19*: 1401-1409, 2013.

22 Yue P, Zhang X, Paladino D, Sengupta B, Ahmad S, Holloway RW, Ingersoll SB and Turkson J: Hyperactive EGF receptor, Jaks and Stat3 signaling promote enhanced colony-forming ability, motility and migration of cisplatin-resistant ovarian cancer cells. Oncogene *31*: 2309-2322, 2012.

23 Sims JN, Graham B, Pacurari M, Leggett SS, Tchounwou PB and Ndebele K: Di-ethylhexylphthalate (DEHP) modulates cell invasion, migration and anchorage independent growth through targeting S100P in LN-229 glioblastoma cells. Int J Environ Res Public Health *11*: 5006-5019, 2014.

24 Yeon SE, No da Y, Lee SH, Nam SW, Oh IH, Lee J and Kuh HJ: Application of concave microwells to pancreatic tumor spheroids enabling anticancer drug evaluation in a clinically relevant drug resistance model. PLoS One *8*: e73345, 2013.

25 Arai K, Sakamoto R, Kubota D and Kondo T: Proteomic approach toward molecular backgrounds of drug resistance of osteosarcoma cells in spheroid culture system. Proteomics *13*: 2351-2360, 2013.