Germline Cytoskeletal and ECM Single Nucleotide Variations (SNVs) are Associated with Distinct Survival Rates: The Prominence of MUC4 Germline SNVs

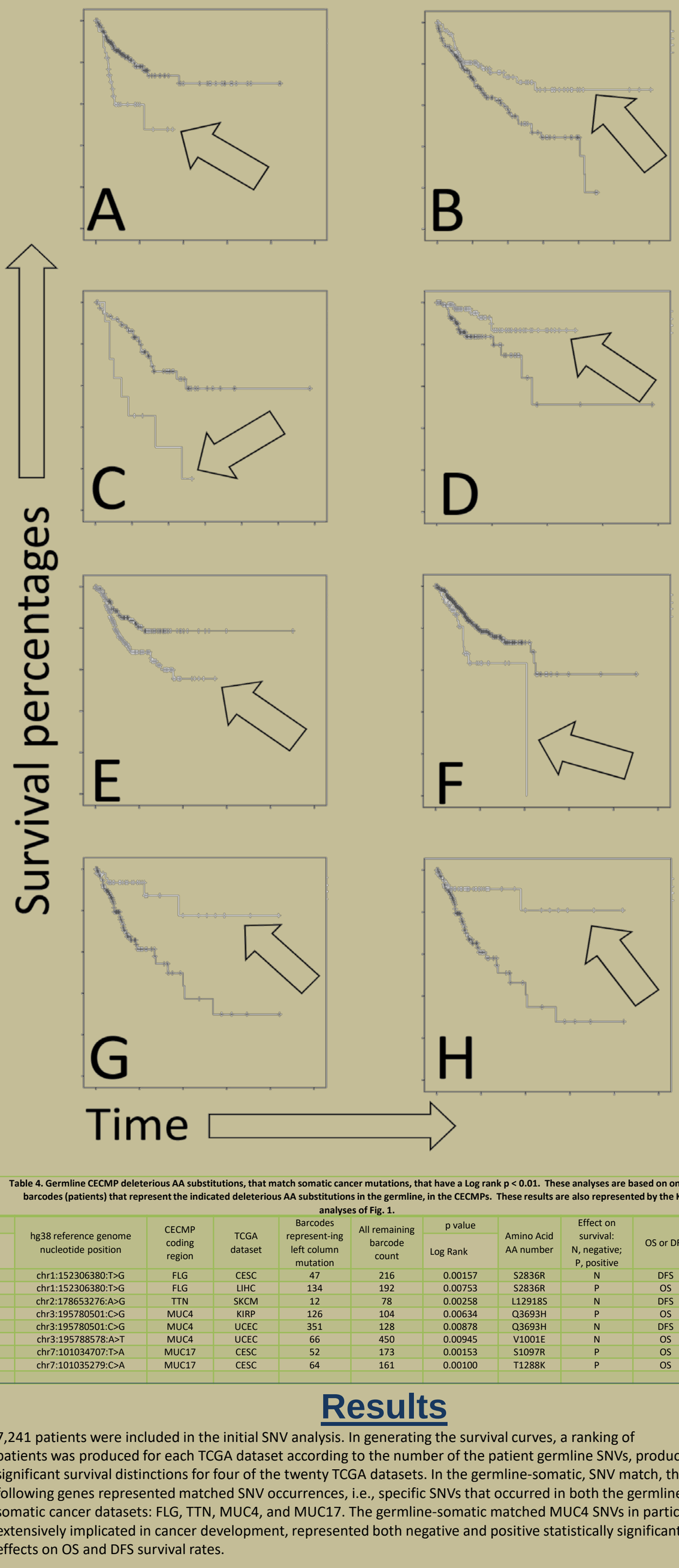
Shayan Falasiri, George Blanck

Department of Molecular Medicine, Morsani College of Medicine

RESEARCH SUPPORTED BY: This research was supported by a scholarly award from the Scholarly Concentrations Program and Research, Innovation & Scholarly Endeavors at USF Health, Morsani College of Medicine

Abstract

Human mutagenesis has a large stochastic component. Thus, large coding regions, especially cytoskeletal and extra-cellular matrix protein (CECMP) coding regions are particularly vulnerable to mutations. Recent results have verified a high level of somatic mutations in the CECMP coding regions in the cancer genome atlas (TCGA), and a relatively common occurrence of germline, deleterious mutations in the TCGA breast cancer dataset. The objective of this study was to determine the correlations of CECMP coding region, germline nucleotide variations with both overall survival (OS) and disease-free survival (DFS). TCGA, tumor and blood variant calling files (VCFs) were intersected to identify germline SNVs. SNVs were then annotated to determine potential consequences for amino acid (AA) residue biochemistry. Germline SNVs were matched against somatic tumor SNVs (i.e., tumor mutations) over twenty TCGA datasets to identify 23 germline-somatic matched, deleterious AA substitutions in coding regions for FLG, TTN, MUC4, and MUC17. The germline-somatic matched SNVs, in particular for MUC4, extensively implicated in cancer development, represented highly, statistically significant effects on OS and DFS survival rates.



Methods

This was a retrospective analysis of CECMP coding region, systemic, single nucleotide variation (SNV) data obtained (via NIH approved project number 6300) from cancer genome atlas (TCGA) data available using the Genomic Data Commons (GDC) web tool. The primary endpoint was the recurrence or progression of the cancer versus cases where the patients were disease-free. Patients for each associated cancer were identified, and SNV calling was performed on each patient's tumor and blood genome files using SAMtools (v1.3) and hg38 as the reference genome. The tumor and blood variant calling files (VCFs) were then intersected using BCFtools to find only SNVs present in both (patient-matched) blood and tumor samples to identify germline and somatic SNVs. Next, two single nucleotide polymorphism (SNP) databases (1000 Genomes and NIH db147) were used to filter out common, known DNA sequence variations in the population at large. SNVs were then annotated by ENSEMBL's Variant Effect Predictor (VEP v. 88) to determine potential consequences for amino acid (AA) residue biochemistry. Germline SNVs were matched against somatic tumor SNVs (i.e., tumor mutations) over the twenty TCGA datasets as a whole to identify 23 germline-somatic matched, deleterious AA substitutions, verified with GDC somatic workflows. Survival curves were generated with the Kaplan-Meier method using clinical data available from cBioPortal.org, a source of publicly available, processed data for the TCGA project. Additional endpoints included overall survival.

Table 1. Germline CECMP coding region SNV rates representing the most statistically significant survival distinctions.

All SNVs or only deleterious AA substitutions	TCGA dataset	Number of barcodes representing the top 25% SNV rate	Number of barcodes representing the bottom 25% SNV rate	Log rank p-value	Effect on survival: N, negative; P, positive	Overall (OS) or disease-free survival (DFS)
All mutations	HNSC	13	157	0.0294	N	DFS
All mutations	KIRP	45	80	0.0186	P	OS
All mutations	SKCM	9	43	0.0311	N	OS
Deleterious AA substitutions	STAD	13	122	0.0109	N	OS

Table 2. Germline CECMP, deleterious AA substitutions that match somatic cancer mutations and represent the most statistically significant overall survival (OS) distinctions. P-value maximum for inclusion in this set of results < 0.05, except for chr3:195788578:A>T. This latter exception is due to the inclusion of chr3:195788578:A>T in Table 6, where there is analysis of survival rates relevant to barcodes in this Table 4 but having only germline SNVs and statistical significance of a p-value < 0.01, which is the case for chr3:195788578:A>T. Parentheses indicate the number of barcodes (patients) with the AA substitution only in the germline. Note: This table includes certain cases where there were no somatic mutations in the TCGA dataset indicated in a given row. This occurs because the germline-somatic SNV matches have been established using the entire TCGA database. Thus, in some cases, there are only germline SNVs representing (i.e., within) a specific cancer dataset. The left column SNPdb refers to > 1% minor allele frequency.

hg38 reference genome nucleotide position (SNPdb, 150 or 147?)	CECMP coding region	TCGA dataset	Barcodes representing left column SNV	All remaining barcode count	Log Rank p-value	Amino Acid	Effect on survival: N, negative; P, positive
chr1:152304195:C>A (150, 147)	FLG	LGG	50 (50)	471 (471)	0.0262	R3564L	P
chr1:152306380:T>G (150, 147)	FLG	LIHC	143 (134)	183 (192)	0.0107	S2836R	P
chr3:195779038:C>T (150, 147)	MUC4	KIRC	13 (7)	78 (84)	0.0249	S4181N	N
chr3:195779374:C>A (novel SNV)	MUC4	KIRP	64 (50)	166 (180)	0.0261	S4069I	P
chr3:195780501:C>G (150)	MUC4	COAD	200 (197)	129 (132)	0.0303	Q3693H	N
chr3:195780501:C>G (150)	MUC4	KIRP	126 (126)	104 (104)	0.0063	Q3693H	P
chr3:195780508:G>C (novel SNV)	MUC4	COAD	145 (140)	184 (189)	0.0386	T3691R	P
chr3:195780508:G>C (novel SNV)	MUC4	KIRP	92 (92)	138 (138)	0.0446	T3691R	P
chr3:195780535:G>A (150, 147)	MUC4	COAD	167 (152)	162 (177)	0.0151	P3682L	P
chr3:195780535:G>A (150, 147)	MUC4	KIRP	106 (93)	124 (137)	0.0107	P3682L	P
chr3:195785341:G>T (150, 147)	MUC4	KIRP	89 (85)	141 (145)	0.0476	P2080H	P
chr3:195788578:A>T (150, 147)	MUC4	UCEC	84 (66)	432 (450)	0.0797	V1001E	N
chr7:101034097:C>A (150, 147)	MUC17	CESC	51 (38)	174 (187)	0.0197	T894K	P
chr7:101034097:C>A (150, 147)	MUC17	STAD	25 (13)	360 (372)	0.0412	T894K	N
chr7:101034707:T>A (150, 147)	MUC17	CESC	52 (52)	173 (173)	0.0015	S1097R	P
chr7:101034707:T>A (150, 147)	MUC17	HNSC	65 (65)	204 (204)	0.0382	S1097R	P
chr7:101035279:C>A (150, 147)	MUC17	CESC	66 (64)	159 (161)	0.0020	T1288K	P
chr7:101038078:C>G (150, 147)	MUC17	GBM	340 (340)	56 (56)	0.0166	P2221R	N
chr7:101038078:C>G (150, 147)	MUC17	SKCM	86 (86)	19 (19)	0.0388	P2221R	P

Table 3. Germline CECMP, deleterious AA substitutions that match somatic cancer mutations and represent the most statistically significant disease-free survival (DFS) distinctions. P-value maximum for inclusion in this set of results is < 0.05. Parentheses indicate the number of barcodes (patients) with the somatic mutation matched, SNV only in the germline. Note: This table includes certain cases where there were no somatic mutations in the TCGA dataset indicated in a given row, per the explanation provided in Table 4 and Results text. The left column SNPdb refers to > 1% minor allele frequency.

hg38 reference genome nucleotide position (SNPdb, 150 or 147?)	CECMP coding region	TCGA dataset	Barcodes representing left column mutation	All remaining barcode count	p value: Log rank	Amino Acid	Effect on survival: N, negative; P, positive
chr1:152306380:T>G (150, 147)	FLG	CESC	55 (47)	208 (216)	0.00471	S2836R	N
chr1:152306380:T>G (150, 147)	FLG	KIRP	38 (29)	180 (189)	0.0286	S2836R	P
chr2:178653276:A>G (150, 147)	TTN	SKCM	12 (12)	78 (78)	0.0026	L12918S	N
chr3:195779374:C>A (novel SNV)	MUC4	KIRC	14 (10)	72 (76)	0.0014	S4069I	N
chr3:195780043:G>A (150, 147)	MUC4	KIRC	10 (8)	76 (78)	0.0209	P3846L	N
chr3:195780501:C>G (150)	MUC4	KIRC	46 (45)	40 (41)	0.0309	Q3693H	N
chr3:195780501:C>G (150)	MUC4	UCEC	361 (351)	118 (128)	0.0165	Q3693H	N
chr3:195780508:G>C (novel SNV)	MUC4	KIRC	32 (30)	54 (56)	0.0251	T3691R	N
chr3:195784525:G>A (novel SNV)	MUC4	KIRP	29 (25)	189 (193)	0.0083	P2352L	P
chr3:195785341:G>T (150, 147)	MUC4	UCEC	228 (197)	251 (282)	0.0398	P2080H	N
chr7:101034097:C>A (150, 147)	MUC17	STAD	21 (11)	289 (299)	0.0420	T894K	N
chr7:101034707:T>A (150, 147)	MUC17	LGG	114 (114)	373 (373)	0.0495	S1097R	P
chr7:101035279:C>A (150, 147)	MUC17	LGG	114 (104)	373 (383)	0.0485	T1288K	P

Table 4. Germline CECMP deleterious AA substitutions, that match somatic cancer mutations, that have a Log rank p < 0.01. These analyses are based on only the barcodes (patients) that represent the indicated deleterious AA substitutions in the germline, in the CECMPs. These results are also represented by the KM analyses of Fig. 1.

hg38 reference genome nucleotide position	CECMP coding region	TCGA dataset	Barcodes represent-ing left column mutation	All remaining barcode count	p value: Log Rank	Amino Acid AA number	Effect on survival: N, negative; P, positive	OS or DFS
chr1:152306380:T>G	FLG	CESC	47	216	0.00157	S2836R	N	DFS
chr1:152306380:T>G	FLG	LIHC	134	192	0.00753	S2836R	P	OS
chr2:178653276:A>G	TTN	SKCM	12	78	0.00258	L12918S	N	DFS
chr3:195780501:C>G	MUC4	KIRP	126	104	0.00634	Q3693H	P	OS
chr3:195780501:C>G	MUC4	UCEC	351	128	0.00878	Q3693H	N	DFS
chr3:195788578:A>T	MUC4	UCEC	66	450	0.00945	V1001E	N	OS
chr7:101034707:T>A	MUC17	CESC	52	173	0.00153	S1097R	P	OS
chr7:101035279:C>A	MUC17	CESC	64	161	0.00100	T1288K	P	OS

Results

7,241 patients were included in the initial SNV analysis. In generating the survival curves, a ranking of patients was produced for each TCGA dataset according to the number of the patient germline SNVs, producing significant survival distinctions for four of the twenty TCGA datasets. In the germline-somatic, SNV match, the following genes represented matched SNV occurrences, i.e., specific SNVs that occurred in both the germline and somatic cancer datasets: FLG, TTN, MUC4, and MUC17. The germline-somatic matched MUC4 SNVs in particular, extensively implicated in cancer development, represented both negative and positive statistically significant effects on OS and DFS survival rates.

Discussion

As expected, given the large genomic space covered by a collection of CECMP coding regions that commonly incur somatic mutations in cancer, we identified numerous germline SNVs in these coding regions. Overall, this study indicates numerous CECMP germline SNVs, represent biochemically dramatic, and likely structurally deleterious AA substitutions, that have apparent survival rate impacts, warranting further investigation about the potential involvement of germline CECMP SNVs in the pathogenesis of these cancers. In short, the systemic, CECMP coding region SNVs may fall into a broad category of germline SNVs, exemplified by BRCA1 mutations, that may facilitate or hamper cancer development.