Gene 669 (2018) 91–98

Contents lists available at ScienceDirect

Gene

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

Research paper

Germline cytoskeletal and extra-cellular matrix-related single nucleotide variations associated with distinct cancer survival rates

Shayan Falasiri[a], Tasnif Rahman[a], Yaping N. Tu[a], Timothy J. Fawcett[b], George Blanck[a,*]

[a] Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL, United States
[b] Department of Chemical and Biomedical Engineering, College of Engineering, Research Computing, University of South Florida, Tampa, FL, United States

ARTICLE INFO

Keywords:
Germline cytoskeleton gene SNVs
Germline ECM gene SNVs
MUC4
TCGA
Germline SNVs and survival rates

ABSTRACT

Background: 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.
Methods: 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.
Results: 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.
Conclusions: 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. The above results contribute to the establishment of what is potentially a new class of inherited cancer-facilitating genes, namely dominant negative tumor suppressor proteins.

1. Introduction

Mutagenesis has a significant, random component, which leads to large genes and coding regions being more susceptible to genetic change and genetic damage. For example, most cancer fusion proteins are derived from genes that have unusually large introns, making the generation of a fusion gene, based on randomly positioned DNA breakage sites, more likely, than in the case of genes with small introns. And when the origin of the fusion gene is small, the cancer is very rare, such as in Ewings sarcoma (Narsing et al., 2009; Pava et al., 2012). Likewise, genes that contribute to cancer development, and have mutations late in the process of development of a fully metastatic cancer, are relatively small (Long et al., 2011), befitting the reduced likelihood of these small genes incurring a mutation. In fact, a class of relatively small tumor suppressor genes is routinely mutated late in cancer development, and because of their late mutation occurrence, have been considered as metastasis suppressor genes. However, the biochemical functions of such metastasis suppressor genes are not notably distinguishable from the biochemical functions of classical tumor suppressor genes; and the later stage mutation occurrence in these genes likely reflect only their small sizes (Long et al., 2011).

A simple search of the cancer genome atlas (TCGA) for the most commonly mutated coding regions includes a very large number of cytoskeleton and extra-cellular matrix (CECMP) coding regions (Parry and Blanck, 2015; Parry and Blanck, 2016), along with such genes as p53, BRAF and isocitrate dehydrogenase, the latter three examples having been commonly associated with cancer development using many experimental tissue culture and other approaches. The CECMP coding regions, particularly the CEMCP coding regions that are frequently mutated in cancer, represent about 8.5% of the human exome, and thus their mutations are not unexpected, given the randomness of DNA damage. The CECMPs have also been associated with cancer

Abbreviations: CECMP, cytoskeletal and extra-cellular matrix protein; DFS, disease-free survival; ECM, extra-cellular matrix; GDC, genome data commons; OS, overall survival; SNP, single nucleotide polymorphism; SNV, single nucleotide variation; SPSS, statistical package for the social sciences; TCGA, the cancer genome atlas; VCF, variant calling file; VEP, variant effect predictor; WXS, whole exome sequence

[*] Corresponding author at: 12901 Bruce B. Downs Bd., Tampa, FL 33612, United States.
E-mail address: gblanck@health.usf.edu (G. Blanck).

https://doi.org/10.1016/j.gene.2018.05.037
Received 3 April 2018; Accepted 13 May 2018
Available online 17 May 2018

development, but in far less certain manners. For example, a disorganized cytoskeleton has been associated with an extreme tumorigenic phenotype (Verderame et al., 1980; Kopelovich et al., 1977; Pollack et al., 1968; Vogel et al., 1973; Steinberg et al., 1979; Shin et al., 1975), and cytoskeletal proteins have been indicated as candidate, cancer driver proteins when mutated (Fawcett et al., 2015). Yet, at least certain components of the cytoskeleton are required for migration and presumably metastasis (Pokorna et al., 1994; Zachary et al., 1986; Xu et al., 2012; Gilardi et al., 2016; Pina-Medina et al., 2016; Narumiya et al., 2009; Kim et al., 2009). It is possible a resolution to this contradiction is in the offing with a focus on particular coding regions, whereby certain CECMP coding regions may represent cancer drivers and others are “protected” from mutation in cancer development (Segarra et al., 2017).

Again reflecting the random nature of DNA damage, many inherited tumor suppressor mutations occur in large genes, such as BRCA1 and RB1. Thus, keeping in mind the above indications of a role for the cytoskeleton and ECM (Naba et al., 2014) in cancer development, we considered the possibility that germline CECMP coding region mutations (or single nucleotide polymorphisms) would reflect survival rate distinctions within certain TCGA datasets. The results below indicate this is indeed the case, with MUC4 representing the most striking example, particularly due to its previous, extensive consideration as a factor in cancer development via somatic mutations (Ogata et al., 1992; Balague et al., 1994; Gautam et al., 2017; King et al., 2017).

2. Methods

2.1. Overview

Tumor and matching blood sample exome (WXS) slices were obtained from the genome data commons (GDC) via dbGaP approved project #6300 (Grossman et al., 2016). All matching nucleotides, in the matching cancer and blood files were retained, and all nucleotides representing the hg38 reference genome and single nucleotide polymorphism databases (as detailed below) were removed. The remaining nucleotides for each barcode (patient) were termed, “germline SNVs (single nucleotide variations)” for the purpose of this study (supporting online material (SOM), Tables S1, S2).

2.2. CERBERUS girdle

This program is run in terminal to prepare and run the subsidiary scripts in parallel, by creating the data architecture for subsidiary programs (APOLLO, CHARON, CERBERUS, HERACLES, COMPILER). Although the primary purpose for this study was to identify germline SNVs as defined in the above overview, this program will retain both blood and cancer somatic mutations.

2.3. APOLLO

APOLLO reads the metadata for a TCGA cancer dataset (e.g., TCGA-BLCA). Using UNIX Bash commands and their associated options, the program extracts the UUID of the file and corresponding file name for the exome (WXS) BAM files. Some inaccuracies in the data collection can be resolved by recovering the proper file names from the metadata. APOLLO then determines whether there is a blood and primary tumor WXS BAM file pair. If no such pair is found for a barcode, all corresponding files for that particular barcode are then excluded from an APOLLO generated manifest used by CHARON (below). At the end of its run, APOLLO reports successful creation of a manifest (KIRC APOLLO manifest example, Table S3) along with the number of all available blood-primary tumor pairs for that TCGA cancer set.

2.4. CHARON

CHARON is the downloader arm of the CERBERUS Girdle. It pulls the manifest created by APOLLO for the requested TCGA set. Running CHARON via the CERBERUS Girdle prompts several diagnostic checks. The CERBERUS Girdle attempts to find and reference any internal memory corresponding to the requested TCGA set, which avoids the overwriting of successfully analyzed data and conserves usage of bandwidth and storage. After the user designates the preferred number of downloaders, the CERBERUS Girdle terminal splits the requested TCGA manifest into the specified number of parallel downloaders and inserts the specified chromosome regions. All parallel instances of CHARON attempt to then BAM slice all barcodes (as described above) with tumor and blood BAM file pairs.

2.5. CERBERUS

CERBERUS is the core variant calling component of the CERBERUS Girdle pipeline (Tables S4, S5). The CERBERUS Girdle runs several diagnostic checks similar to what occurred prior to the use of CHARON. First, the CERBERUS Girdle attempts to validate the BAM sliced files. BAM sliced files that are corrupted during transfer will likely also contain corrupted Headers or EOFs (end of file lines) (GDC's application programming interface). Therefore, the CERBERUS Girdle uses SAMtools (v1.3.1) (Li et al., 2009) and its quick-check function to identify all newly downloaded BAM files with corrupted Headers and EOF's. The files that fail this test are recorded into a failed file registry. Then, all files in the designated TCGA set download directory are written into a new manifest for CERBERUS, excluding all files deposited into the failed file registry. Second, the CERBERUS Girdle then attempts to reference any internal memory corresponding to the requested TCGA set. As in the case of CHARON and indicated above, this is a redundancy to ensure that no successfully analyzed files are overwritten. With the creation of a final working manifest for CERBERUS, CERBERUS Girdle will create the data file directories needed for the output of the pipeline. The CERBERUS Girdle will then execute the CERBERUS program.

The manifest imported into CERBERUS is split into single barcodes which are set into a loop. Each barcode is associated with a blood and a tumor WXS BAM file. Minimally filtered variants, with respect to the hg38 reference genome are then generated from these files, using SAMtools's mpileup (with a mapping quality threshold of 10) and using BCFtools's (v1.2) call functions. The minimally variant files are saved as compressed VCF files. Upon successful analysis of each individual barcode, CERBERUS writes the corresponding barcode and file names to a registry and internal memory of successfully run data. The blood and tumor VCF files representing each barcode are then intersected, using Tabix's (v0.2.6) (Li, 2011) index and BCFtools's isec functions, i.e., to find germline SNVs (with respect to hg38) represented by each barcode. Simultaneously, variants that are private to only the tumor VCF file of each patient are classified and saved as the somatic tumor mutations for that patient; likewise, the variants that are private to only the blood VCF file of each patient are classified and saved as the somatic blood mutations for that patient. As it might have been possible that already discovered SNP's may have been present in the germline variant files, two SNP databases were referenced (1000 Genomes: 1000G Phase 1 SNPs High Confidence and NIH: db147), and any SNP's matching in both exact position and allele change were filtered out of all germline data for all patients. However, as indicated in the Results below, several SNVs, representing tri-allelic variants, in turn representing specific nucleotide positions in db147, were recovered by the above process. This occurred because the process was designed to remove db147 SNVs that matched both position and identity of only a single minor allele nucleotide.