Open Access

OBM Genetics

LIDSEN Publishing Inc.

Research Article

Correlation of Mutational Signatures in Cancer Genes with General Signatures

Junhyeong T. Park [1], Junseong A. Park [1, 2], Igor F. Tsigelny [3, 4, 5, *], Valentina L. Kouznetsova [3, 5]

1. REHS Program, San Diego Supercomputer Center, University of California at San Diego, La Jolla 92093, CA, USA; E-Mails: jtpark2491@gmail.com; japark1995@gmail.com
2. MAP Program, University of California at San Diego, La Jolla 92093, CA, USA
3. San Diego Supercomputer Center, University of California at San Diego, La Jolla 92093, CA, USA; E-Mails: itsigeln@ucsd.edu; vkouznetsova@ucsd.edu
4. Department of Neurosciences, University of California at San Diego, La Jolla 92093, CA, USA
5. BiAna, La Jolla 92038, CA, USA

* **Correspondence:** Igor F. Tsigelny; E-Mail: itsigeln@ucsd.edu

Academic Editor: Lunawati L Bennett

Special Issue: Cancer Genetics and Epigenetics Alterations II

OBM Genetics
2022, volume 6, issue 1
doi:10.21926/obm.genet.2201147

Received: December 07, 2021
Accepted: January 17, 2022
Published: February 13, 2022

Abstract

The occurrence of various mutation patterns, such as changes in the DNA sequence and the loss of some sequences, is called a “mutational signature,” and they represent the molecular fingerprints that exist for the type of mutation occurring in a specific gene. Our study elucidates the correlations of mutational signatures in frequently mutated cancer genes with general mutational signatures previously found for different cancers. We hypothesized that the top twenty most frequently mutated genes (MFMG) of a cancer type would have the highest correlation with the general signatures related to the cancer. The program for our research, SignaGen, was created using MATLAB to take in genomic sequence data and mutation data (consisting of the type, location, and frequency of each mutation) to calculate the mutational signatures of genes. The correlation values for the top twenty MFMG were organized into heatmaps for each cancer type observed. By looking at the heatmaps, we could

see that the MFMG did have relatively higher correlation values with the general signatures that were related to the cancer type. However, there were also cases in which the MFMG had lower correlation values with a related signature than other associated signatures. For example, the MFMG of skin cancer had an average correlation of 7.76% with signature 17, while having an average correlation of 43.02% with signature 11. Not only was the average lower than the average correlation with the other related general signatures, but it was also lower than the average correlation with unrelated signatures. To investigate this inconsistency and verify the significance of these correlation values, we compared the correlation values of the MFMG to the correlation values of randomly selected genes of similar length. Even if the MFMG's correlation with the related signatures is low, our hypothesis would still be supported if they had a higher correlation than the random genes. We took three of the twenty MFMG for each cancer and compared their correlation with the random genes of similar length for each cancer type and found that the MFMG had a higher correlation than the random genes for most cases.

Keywords
Mutational signature; cancer; DNA sequence; SignaGen

1. Introduction

Cancer is a disease caused by the unrestrained growth of tumor cells in the body that are thought to be summoned by mutations in the cells. However, the root of these mutations is still unclear, and numerous scientists continue to look for the answer. What we do know, is that DNA repair mechanisms and carcinogen-induced DNA damage determine the pattern of genomic mutations that are the root cause of cancer [1].

As humans are exposed to various environmental factors and suffer DNA damage from interacting with them, mutations in their DNA accumulate, triggering the growth of tumor cells in the body. These mutations, also called somatic mutations, cause the human body to activate its DNA-repair function to repair the damaged DNA. However, if there is a problem with the DNA-repair function during this time, the mutation will accumulate in the cells, resulting in the development of cancer. The occurrence of various patterns, such as changes in the DNA sequence and loss of some sequences, is called a "mutational signature," and it represents the molecular fingerprints that exist for the type of mutation occurring in a specific carcinoma. In short, when the human body is exposed to a mutagen, a certain mutational signature appears due to the DNA damage mechanism. As a result, various somatic mutations in the human nucleotide sequence occur depending on the specific DNA sequence, DNA replication and RNA transcription process, and epigenetic properties.

Smoking has been known to be the number one risk factor for lung cancer for over 60 years. In order to investigate the mutational effects of smoking on the human genome, a study was conducted by Alexandrov and colleagues to compare the somatic mutations and methylation in smokers and nonsmokers for 17 different cancer types related to smoking [2]. A total of 5243 cancer genome sequences were examined, and it was found that smoking is associated with increased mutation burdens of multiple signatures. These mutational signature patterns could be attributable

to misreplication of DNA damage caused by tobacco carcinogens, indirect activation of DNA editing, and other mechanisms. A single mutational signature was found to be present in all cancer types related to smoking. Results of the study supported the proposition that smoking increased the risk of cancer by increasing the number of specific somatic mutations. In this way, it is possible to infer the cause and mechanism of cancer by studying the traces of the mechanism by which DNA damage occurred. By studying the traces of the mechanism by which the DNA damage occurred, it becomes possible to make inferences about the cause and development of cancer.

In the field of cancer genomics, there have been many studies conducted on the mutational signature, especially on which mutations occur in various carcinomas. Since the publication of the results of large-scale analyses of many mutational signatures across the spectrum of human cancer types [3], the study of mutational signatures in cancer has become an essential field.

Researchers used Genome Sequencing to analyze the genomes of 2700 *C. elegans* that are easy to cultivate and have similar characteristics of genetic information to humans in order to find the genetic elements that determine DNA mutations. Twelve DNA toxic substances were produced in 150 combinations and were then exposed to several small *C. elegans* with defects in various DNA repair functions. Through this process, it was found that the DNA repair function along with the type of DNA damaging substance determines the mutation signature pattern. When the *C. elegans* is exposed to aflatoxin, a carcinogen that causes liver cancer, the base cytosine (C) is replaced with thymine (T), but when exposed to gamma rays, thymine (T) is substituted with adenine (A) or cytosine (C). In addition, when exposed to the same damaging substance, it was confirmed that if the DNA repair function is defective, the occurrence of mutation signatures increased sharply compared to the normal case [1].

p53 is a tumor suppressor protein known to play a role in preventing the accumulation of cells with damaged genomes by causing apoptosis of these cells or stopping the cell cycle [4]. Mutations in the *TP53* gene that encodes the p53 protein hinder the activation of these processes, resulting in the growth and division of tumor cells. Because of its central role in tumor suppression, it is not surprising that it is one of the most frequently mutated genes in cancer, with over 50% of all human cancer types having *TP53* mutations [5]. More than 36 000 *TP53* mutations were found, and about 80% of the p53 mutations were identified as amino-acid substitutions. A study of mutant p53 (mutp53) in mouse models showed that tamoxifen-induced ablation of mutp53 resulted in an increase in survival rate for the mice, as tumor cells underwent apoptosis and regression or stagnation [6, 7]. Another study found that mutations resulting in the inhibition of telomere-binding factor POT1 caused telomere fragility, replication fork stalling, and telomere elongation [8, 9]. The mutations causing the proliferation of cancer cells lacking *POT1* have been found in several human cancers, such as leukemia, glioma, and cutaneous melanoma, producing malignant tumors, and showed potential for new cancer treatment methods such as enzyme therapy for these cancer types.

Results of numerous studies showed that analysis of mutational signatures could suggest which substances cause the specific cancer type to develop and which DNA repair function was impaired that resulted in the specific mutations. It is expected that the analysis of these signatures can also provide clues for developing personalized cancer treatments. The study of mutational signatures in past years has further revealed the principle of determining the type of mutations, and these results mark an important milestone in the development of cancer diagnosis and treatment in the future.

2. Material and Methods

In this study, we developed a program SignaGen that would elucidate the mutational signatures of analyzed genes and their correlation to thirty cancer mutational signatures.

The flowchart in Figure 1 shows the algorithm of the program SignaGen. SignaGen first begins by loading the genomic information of the genes from NCBI (National Center for Biotechnology Information) and the mutation data for each of these genes from COSMIC (Catalogue of Somatic Mutations In Cancer).

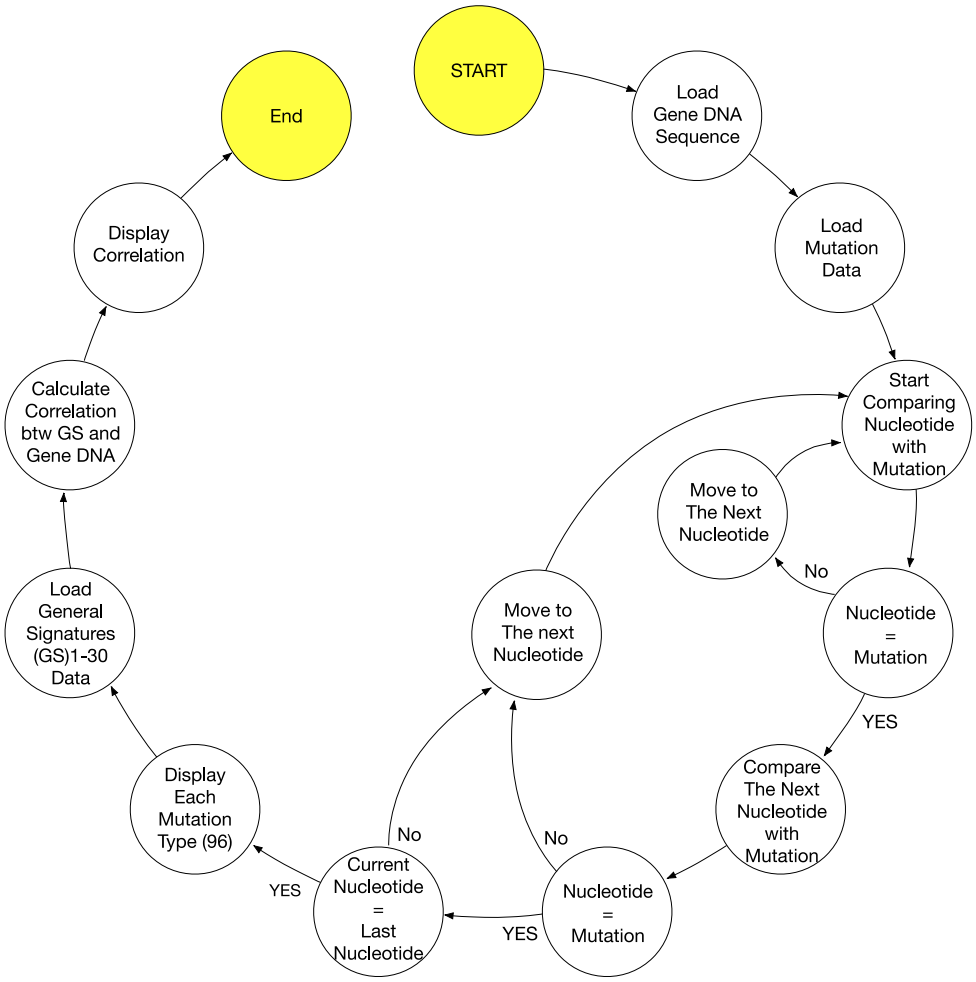

Figure 1 Flowchart of the program SignaGen in calculating mutational signatures and plotting correlation heatmaps.

The mutation data includes all the mutations that are presented in the description of a gene in the tissue samples by COSMIC (see Table S1). In our calculations, we only use the data for single nucleotide polymorphisms and none of the other mutation types. We also assume that every one of these mutations will occur in the sequence coding for the specific gene. Using this data, SignaGen finds the segment of a studied gene where the greatest number of consecutive mutations can possibly occur, to analyze. This process is done by looping through the gene sequence multiple times while using the mutation data to check if the appropriate nucleotide is present at a certain position. For the mutation data for *FAT1* (Table 1), SignaGen will begin by checking for the first mutation: a possible A>T mutation at position 252. If an adenine nucleotide is present at position 252 for the

A>T mutation, SignaGen will continue on to the next mutation: C>T at 270. If there is a cytosine mutation at position 270 for the C>T mutation, it will look at the third mutation, then the fourth, then the fifth, etc. This process continues until it encounters a mutation that is not possible (the appropriate nucleotide that has to be mutated does not appear at the given position). SignaGen will record the number of consecutive mutations and restart the whole process beginning with the first mutation. This time however, it will begin counting from the 2nd nucleotide of the gene sequence, so it will look for the first mutation 270 C>T at position 271. SignaGen will continue to record the number of consecutive possible mutations encountered until the very last nucleotide. Through this procedure, the program finds the optimal segment for its calculations, from the position where it had the greatest number of consecutive possible mutations recorded to the end of the gene sequence (as shown in Figure 1). All subsequent data collection and analysis for the gene are done using the segment obtained through this process.

Table 1 Mutation data for the first ten mutations out of 228 for *FAT1*.

Position (AA)	Mutation (CDS)	Mutation (Amino Acid)	Legacy Mutation ID	Count	Mutation Type
84	c.252A>T	p.K84N	COSM6916946	1	Substitution – Missense
90	c.270C>T	p.L90=	COSM7544359	1	Substitution – coding silent
93	c.277T>C	p.F93L	COSM7547014	1	Substitution – Missense
94	c.281 285delinsCCTTG	p.C94 F95delinsSL	COSM6916955	1	Substitution – Missense
126	c.377dup	p.N126Kfs*3	COSM7609043	1	Insertion – Frameshift
131	c.392C>T	p.A131V	COSM4416272	1	Substitution – Missense
146	c.437del	p.R146Nfs*19	COSM5946185	1	Deletion – Frameshift
173	c.517G>A	p.A173T	COSM1206553	1	Substitution – Missense
179	c.536G>A	p.G179E	COSM9250985	1	Substitution – Missense
224	c.670G>A	p.A224T	COSM3131928	1	Substitution – Missense

SignaGen calculates the frequencies of specific codons from the original gene sequence (Accession: NG), including frequencies of codons from the complementary DNA strand of the top twenty genes that are most frequently mutated for a given cancer type. The representations for these mutational signatures are shown using the six substitution subtypes: C>A, C>G, C>T, T>A, T>C, and T>G. SignaGen references the 2nd column "Mutation (CDS)" of the Table 1 again for the mutation type and nucleotide position. At each of the positions given, if the program encounters one of the six substitution subtypes, it will increment the count of the mutation type, found on the 5th column "Count" of the mutation data, for the mutated codon. If SignaGen encounters a G>A, G>T, G>C, A>G, A>C, or A>T mutation, it reads the complementary strand, converting the trinucleotides into its complementary trinucleotides, checks the number of the mutated nucleotide and adds them to the count of the mutation type of the complementary strand.

Based on its calculations, SignaGen displays the mutational signature according to 96 possible mutations based on the six substitution subtypes and the different combinations of neighboring nucleotides, as shown in Figure 2. The probabilities for the six types of substitutions are displayed as bars of different colors, with the horizontal axis representing the type of mutations and the

vertical axis representing the frequency of a mutation type (total count for a codon divided by the total count for all codons).

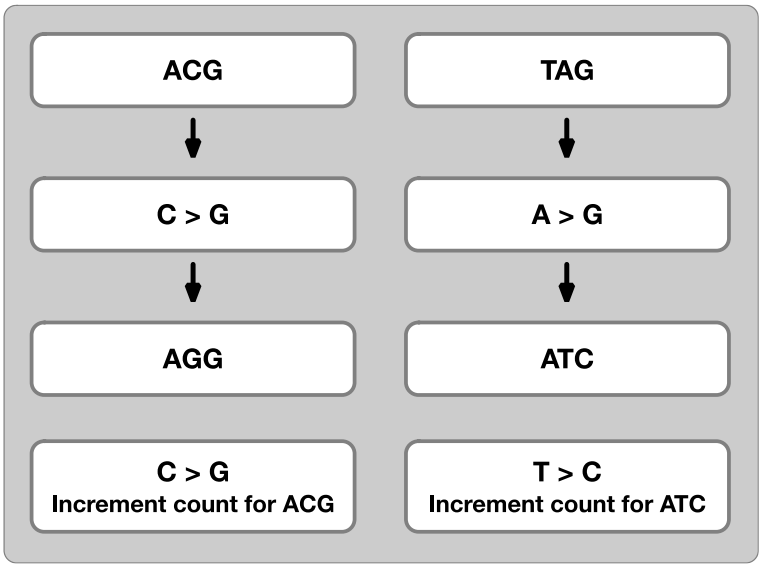

Figure 2 Examples of method used for calculating mutational signature of a gene, the left side showing the result when a mutation is one of the six substitution subtypes and the right showing the result when the mutation is not one of the six. Because C>G is one of the six subtypes, SignaGen will just increment the count of the C>G mutation for the codon ACG. Because A>G is not one of the six subtypes, SignaGen will increment the count of the complementary mutation T>C for the complementary codon ATC.

Therefore, we have developed a program—SignaGen—that can analyze genomic DNA data, calculate the frequency of specific mutation types, and plot mutational signatures. After calculating the mutational signature of the genes, SignaGen loads in the pattern data of mutational signatures found in human cancers called general signatures provided in the study by Alexandrov and coauthors to calculate the correlation of each of the twenty most frequently mutated genes of specific cancer and the thirty cancer mutational signatures [3].

There are several methods, which could be used to determine the correlation of the gene mutational signature and the cancer mutational signatures, such as Pearson, Kendall, Spearman, etc., but we chose to use cosine similarity because it gives a larger scale for the lower frequency mutations, as it is not affected by a mean value like the Pearson correlation:

$$similarity = cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Because each substitution of the most frequently mutated genes (MFMGs) of a specific cancer type and the thirty cancer mutational signatures will always be a non-negative value, the cosine similarity calculation will always return a value from 0 to 1. Cosine similarity of 1 will indicate that the two signatures are identical, whereas a value of 0 would indicate that they are completely different.

Once the SignaGen finishes calculating the correlation of the top twenty most frequently mutated genes for a cancer type and the thirty general signatures, the result is displayed in a heatmap. This process was done for the top 20 genes for each of 16 studied cancer types.

Using this method, we expected to validate our hypothesis that the signatures of the MFMGs of a specific cancer type will have the highest correlation to the general mutational signatures found to be related to specific cancer. We elucidated the correlation values of the top twenty genes of several cancer types. To determine whether the correlation values were of significance, we compared the correlation values of the top twenty genes to the correlation values of a set of about 440 random genes with the length equivalent to the length of the selected genes. We hypothesized that a comparison of the most frequently mutated genes and the random genes could give us insight into the relation of the gene signature to specific cancer. Finding that the most frequently mutated genes had higher correlation values than the random genes would support our hypothesis; and finding the random genes to have similar correlation values as the most frequently mutated genes may indicate that the signature is relatively weakly correlated with the cancer type.

We can use this program to identify whether the mutation signatures of genes of the analyzed cancer types are correlated with correspondent general mutation signatures. The program SignaGen was developed using MATLAB for better numerical analysis and code expandability.

3. Results

We describe in more detail the calculations of the correlation parameters of a single gene signature and general signatures on example of *FAT1* gene—one of the most frequently mutated in colorectal cancer.

Figure 3 is the result of the matching process of SignaGen, and it shows that based on the mutation data for colorectal cancer, the *FAT1* gene can have a total of 198 mutations and has a maximum match rate of 21.2% (42/198 possible mutations) with the mutation information starting at the 22 089th nucleotide of *FAT1*. SignaGen takes the segment of the sequence, from the 22 089th nucleotide to the end of the sequence, calculates the frequency of each mutation, and displays the mutational signature of the *FAT1* gene (NCBI Reference Sequence: NG_046994.1, Figure 4).

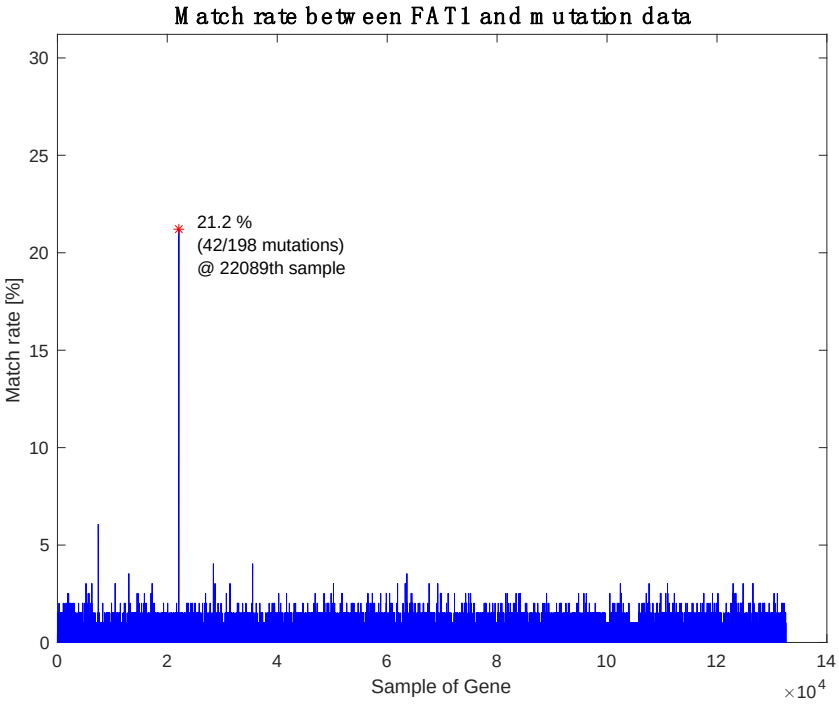

Figure 3 Match rate between *FAT1* and its mutation data for colorectal cancer generated by SignaGen. The greatest number of consecutive mutations was 42 out of the 198 mutations described. The first of these mutation (A>T at position 252) was found at position 22 340.

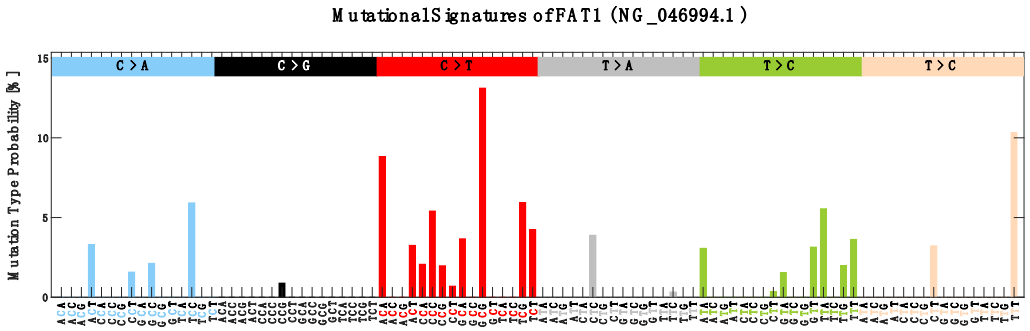

Figure 4 Mutational signature of *FAT1* calculated by SignaGen using its mutation data for colorectal cancer.

If we visually compare this data with each of the thirty general mutual signatures [3], we can see that this comparison result is the most significant similarity to the general Signature 6 shown below in Figure 5.

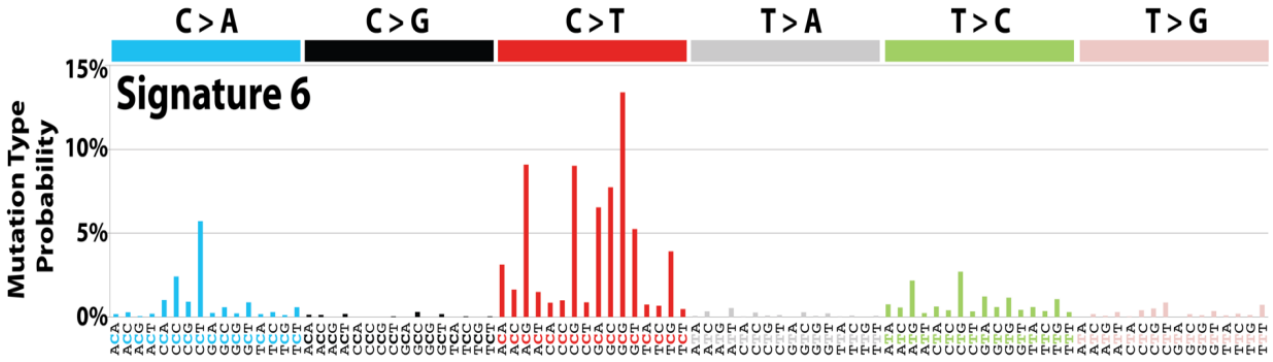

Figure 5 General Signature 6 [3].

We also calculated the correlations of the *FAT1* individual gene signature with the other general signatures, as shown below in Figure 6. As we expected, we found the mutational signature of *FAT1* to be most correlated to general Signature 6 [3] with a calculated correlation value of 53.6%. It can also be seen from Figure 6 that there is a high correlation of about 50% between *FAT1* and Signatures 1, 5, 9, 14, and 15. We note that according to the analysis demonstrated in [3], colorectal cancer is highly correlated with Signatures 1, 5, 6, and 10 (see Figure 6 and Table S2).

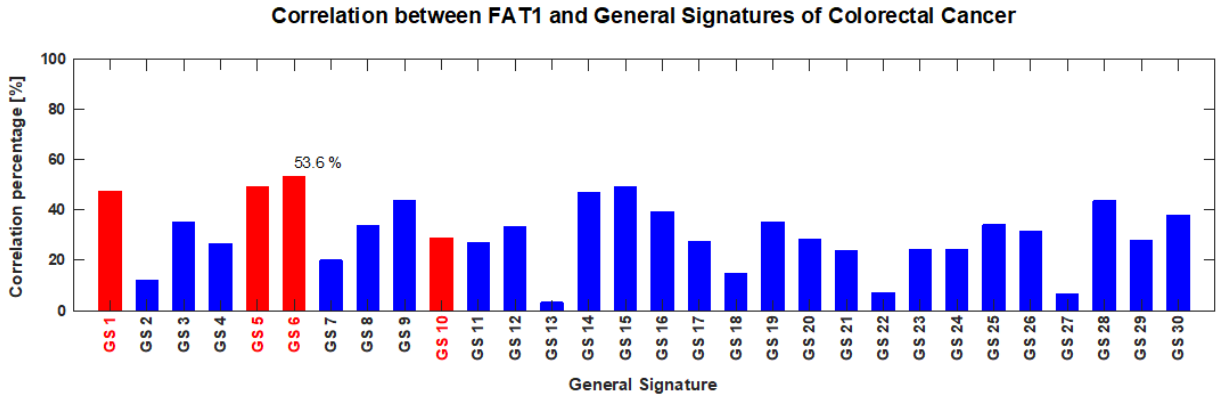

Figure 6 Correlations between FAT1 mutational signature for colorectal cancer and the thirty general signatures calculated by SignaGen. In red are shown the general signatures found to correlate with colorectal cancer (signatures 1, 5, 6, and 10 [3]).

We performed the same simulation that we did for *FAT1*, one of the top twenty genes of colorectal cancer, for the remaining most frequency mutated genes of colorectal cancer from top 20. Using the program SignaGen, we then extracted the correlation between the twenty most frequently mutated genes of colorectal cancer and the general signature and presented the results using a heatmap. The same procedure was done to produce heatmaps for each of the cancer types observed in this study, found in Figure 7A–P.

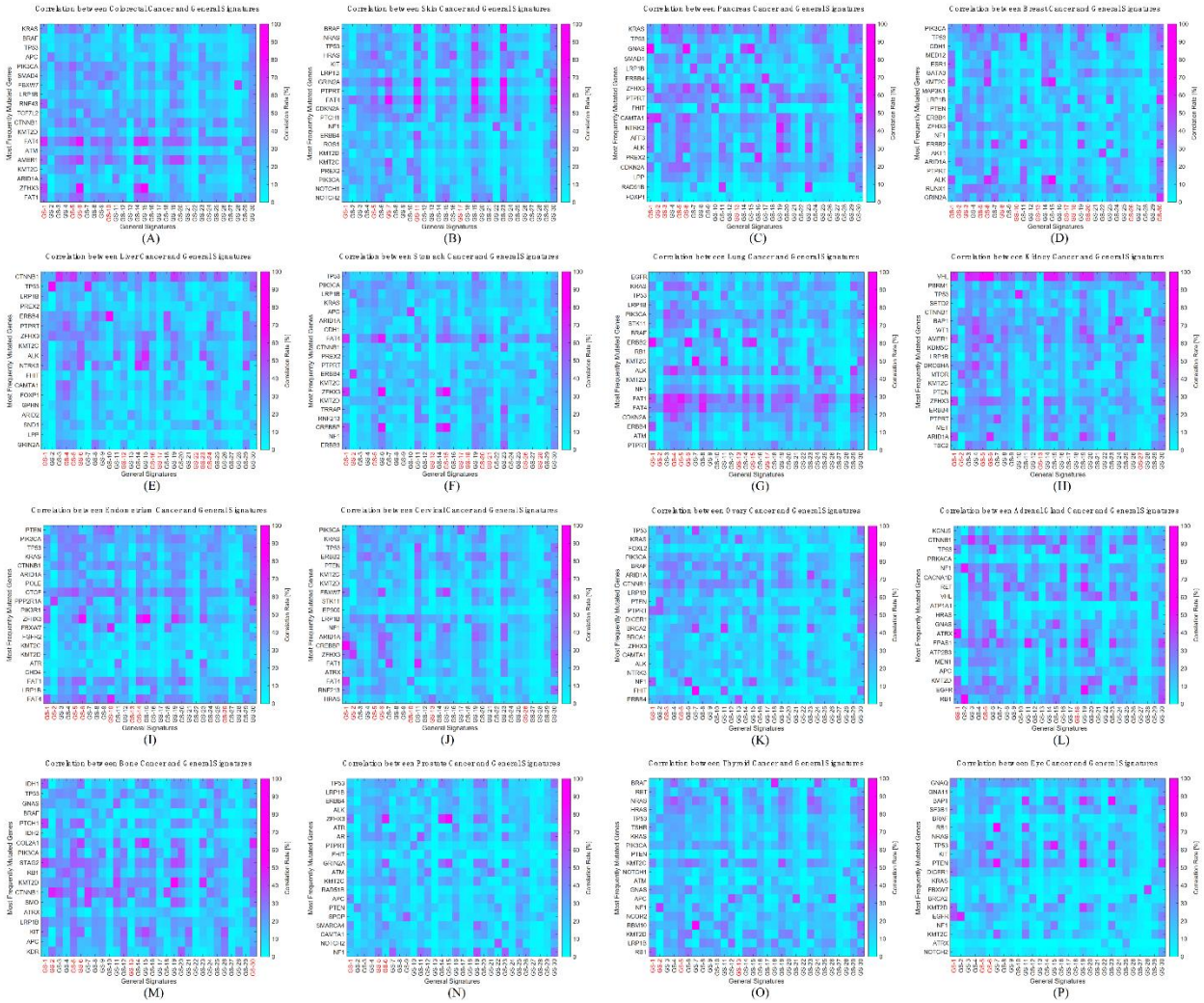

Figure 7 Heatmaps for correlation between the top twenty most frequently mutated genes of various cancers and thirty general mutational signatures. The general signatures that have been found to be related to the specific cancer type for each heatmap is highlighted in magenta, which correlations of three top genes are shown in Table S2. (A) The heatmap for colorectal cancer with general Signatures 1, 5, 6, and 10 highlighted in red on the X axis. (B) Skin cancer; (C) Pancreatic cancer; (D) Breast cancer; (E) Liver cancer; (F) Stomach cancer; (G) Lung cancer; (H) Kidney cancer; (I) Endometrial cancer; (J) Cervix cancer; (K) Ovary cancer; (L) Adrenal gland cancer (neuroblastoma); (M) Bone cancer (osteosarcoma); (N) Prostate cancer; (O) Thyroid cancer; (P) Eye cancer.

3.1 Colorectal Cancer

Figure 7A shows that the genes *FAT4*, *ZFHX3*, *FAT1*, *RNF43*, and *KRAS* have relatively high correlation values with the general signatures that were found to be related to colorectal cancer: Signatures 1, 5, 6, and 10 (Table S2). Some genes such as *FAT4* and *ZFHX3* had high correlation with the related Signatures for several cancer types, indicating that these genes may have more significant roles in colorectal cancer growth than some of the other most frequently mutated genes.

As shown in Figure 8A and Table S2, *FAT4* had greater correlation than the median of the random genes for Signatures 1 (+30.6%), 5 (+28.3%), 6 (+27.6%), and 10 (+42.0%). *ZFHX3* had greater correlation than the median of the random genes for Signatures 1 (+50.0%), 5 (+6.1%), 6 (+67.7%), and 10 (+11.3%). *FAT1* had greater correlation than the median of the random genes for Signatures 1 (+20.6%), 5 (+17.4%), 6 (+34.0%), and 10 (+16.1%).

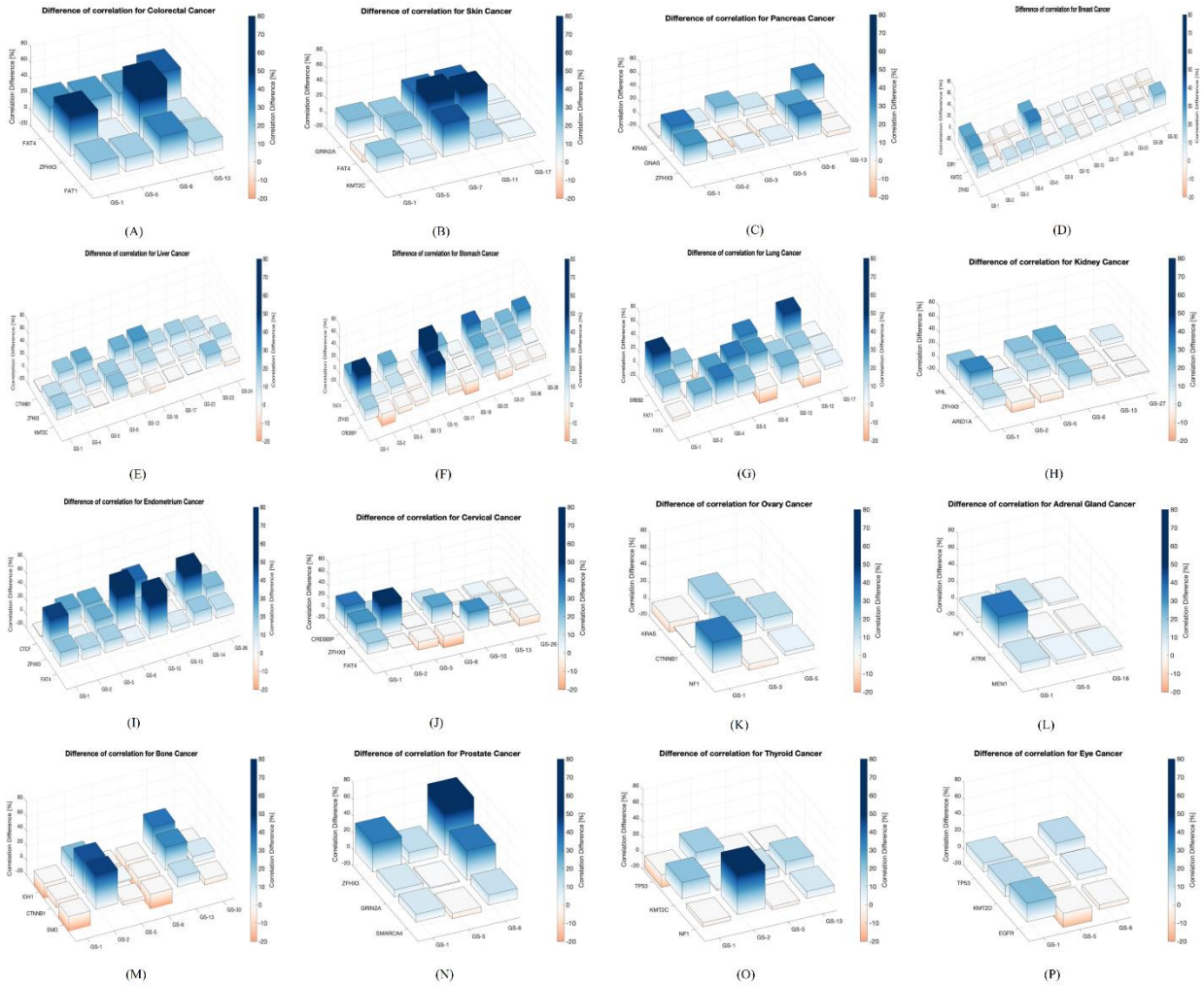

Figure 8 3D bar graphs for difference in correlation of the top twenty most-frequently mutated genes (MFMGs) and random genes of similar length. Positive values indicate a greater top twenty MFMGs correlation, and negative values indicate a greater random gene median value. (A) Colorectal cancer; (B) Skin cancer; (C) Pancreatic cancer; (D) Breast cancer; (E) Liver cancer; (F) Stomach cancer; (G) Lung cancer; (H) Kidney cancer; (I) Endometrial cancer; (J) Cervix cancer; (K) Ovary cancer; (L) Adrenal gland cancer (neuroblastoma); (M) Bone cancer(osteosarcoma); (N) Prostate cancer; (O) Thyroid cancer; (P) Eye cancer.

In order to get an overall look at these correlation values, we also took the average of the correlation values for each cancer type, as shown in Figure 9A and Figure 9B. Colorectal cancer had the greatest average correlation to the general signatures related to the cancer type with a value of

23.71% and the best average correlation difference (6.25%) between its MFMGs' correlation and the random genes' correlation (Figure 9A and Figure 9B).

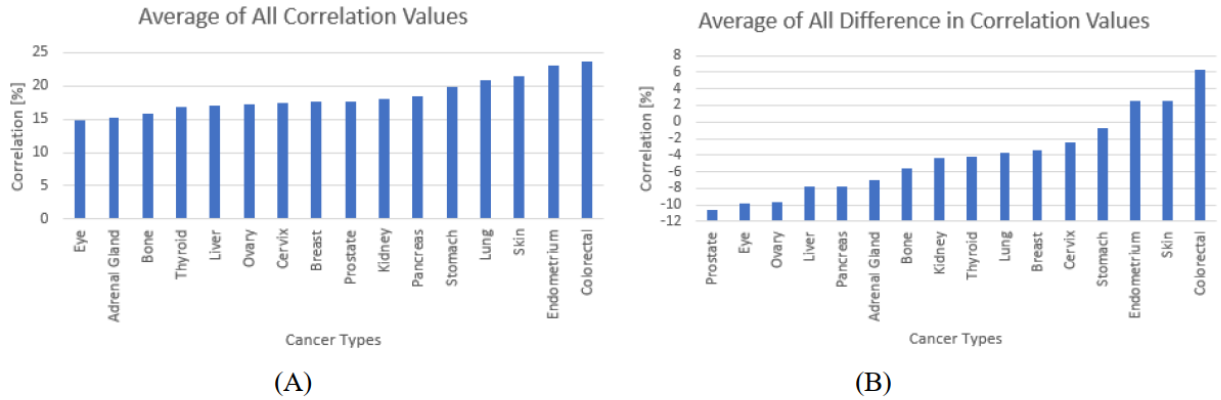

Figure 9 Bar graphs with each bar representing the average of all the correlation values of a cancer type's top twenty most frequently mutated genes (MFMGs) and related general signatures. (A) The average of the correlation values. (B) The average of the difference in correlation values (MFMG–random).

Based on the above method, we performed the same simulations and analyses for the other sixteen cancer types shown below. It was found that the median values of the top twenty most frequently mutated genes in these cancers were mostly higher than the median values of random genes.

3.2 Skin Cancer

For skin cancer, the genes *BRAF*, *TP53*, *HRAS*, *GRIN2A*, *PTPRT*, *FAT4*, *CDKN2A*, *ERBB4*, *KMT2C*, and *NOTCH1* had the highest correlation to the related general Signatures: 1, 5, 7, 11, and 17 (Figure 7B; Table S2). In particular, most of the above mentioned genes were found to have the greatest correlation with general Signatures 7 and 11.

Figure 7B and Table S2 show the calculation result for skin cancer using the three genes with the highest correlation values, *GRIN2A*, *FAT4*, and *KMT2C*, and random genes. As shown, the correlation of the top three genes of skin cancer was higher than that of the random genes for the most of the signatures besides Signature 17. In case of the Signature 17, the median value of the correlation with the top three genes as well as random genes was below 10% correlation. Through this, it can be seen that, unlike other dominant general signatures, Signature 17 is relatively weakly correlated with skin cancer.

The general signatures shown in the difference plots (Figure 2B) are the signatures that have been found to be related to the cancer type. *GRIN2A* had greater correlation than the median of the random genes for Signatures 1 (+21.9%), 5 (+18.7%), 7 (+38.7%), 11 (+40.2%), and 17 (+5.2%). *FAT4* had greater correlation than the median of the random genes for Signatures 5 (+21.6%), 7 (+61.0%), 11 (+54.7%), and 17 (+6.1%). Note that random genes had better correlation values for Signature 1. *KMT2C* had greater correlation than the median of the random genes for Signatures 1

(+19.4%), 5 (+7.3%), 7 (+43.5%), 11 (+14.1%), and 17 (+4.4%). Skin cancer had an average correlation of 21.57% and an average correlation difference of 2.54% (Figure 9A and Figure 9B; Table S2).

3.3 Pancreatic Cancer

For pancreatic cancer, there was no NCBI Reference Sequence (NG) data for *ZNF521* and *EBF1*, so we excluded both cases from the simulation. The genes *KRAS*, *GNAS*, *ZFHX3*, *PTPRT*, *CAMTA1*, *ALK*, and *PREX2* had the highest correlation to the related general Signatures: 1, 2, 3, 5, 6, and 13 (Figure 7C; Table S2).

As shown in Figure 8C and Table S2, *KRAS* had a greater correlation than the median of the random genes for Signatures 2 (+6.5%), 3 (+22.6%), 5 (+9.9%), and 13 (+33.7%). Random genes had better correlation values for Signatures 1 and 6. *GNAS* had a greater correlation than the median of the random genes for Signatures 1 (+37.1%) and 6 (+27.1%). Random genes had better correlation values for Signatures 2, 3, 5, and 13. *ZFHX3* had greater correlation than the median of the random genes for Signatures 1 (+29.7%), 2 (+4.4%), 3 (+4.5%), 5 (+6.8%), and 6 (+33.4%). Random genes had better correlation values for Signature 13. Pancreatic cancer had an average correlation of 18.40% and the greatest average correlation difference of −7.76% (Figure 9A and Figure 9B).

3.4 Breast Cancer

For breast cancer, the genes *PIK3CA*, *ESR1*, *KMT2C*, *ERBB4*, *ARID1A*, *ZFHX3* and *ALK* had the highest correlation to the related general Signatures: 1, 2, 3, 5, 6, 8, 10, 13, 17, 18, 20, 26, and 30 (Figure 7D; Table S2).

As shown in Figure 8D and Table S2, *ESR1* had greater correlation than the median of the random genes for Signatures 1 (+26.2%), 6 (+25.3%), 10 (+1.7%), 13 (+1.8%), and 18 (+2.9). Random genes had better correlation values for Signatures 2, 3, 5, 8, 17, 20, 26 and 30. *KMT2C* had greater correlation than the median of the random genes for Signatures 1 (+28.8%), 6 (+33.4%), 10 (+2.6%), 18 (+4.2%). Random genes had better correlation values for Signatures 2, 3, 5, 8, 13, 17, 20, 26, and 30. *ZFHX3* had greater correlation than the median of the random genes for Signatures 1 (+19.9%), 2 (+2.3%), 3 (+10.7%), 5 (+13.6%), 6 (+12.0%), 8 (+3.5%), 10 (+8.6%), 17 (+5.0%), 18 (+1.1%), 20 (+7.6%), 26 (+1.5%), and 30 (+21.8%). Random genes had better correlation values for Signature 13. Breast cancer had an average correlation of 17.62% and an average correlation difference of −3.45% (Figure 9A and Figure 9B).

3.5 Liver Cancer

For liver cancer, there was no NG data for *ZNF521* and *MAML2*, so we excluded both cases from the simulation. The genes *CTNNB1*, *ZFHX3*, *KMT2C*, *ALK*, *NTRK3*, and *CAMTA1* had the highest correlation to the related general Signatures: 1, 4, 5, 6, 12, 16, 17, 22, 23, and 24 (Figure 7E; Table S2).

As shown in Figure 8E and Table S2, *CTNNB1* had greater correlation than the median of the random genes for Signatures 4 (+19.4%), 5 (+24.8%), 12 (+24.1%), 16 (+29.0%), 17 (+13.7%), 22 (+18.2%), 23 (+14.1%), and 24 (+3.9%). Random genes had better correlation values for Signatures 1 and 6. *ZFHX3* had greater correlation than the median of the random genes for Signatures 1 (+13.3%), 4 (+7.9%), 5 (+12.7%), 6 (+18.2%), 12 (+7.5%), 16 (+16.1%), 17 (+3.0%), 22 (+7.4%), 23

(+7.1%), and 24 (+15.0%). *KMT2C* had greater correlation than the median of the random genes for Signatures 1 (+17.7%), 4 (+5.7%), 6 (+20.4%), 22 (+0.7%), and 23 (+20.5%). Random genes had better correlation values for Signatures 5, 12, 16, 17, and 24. Liver cancer had an average correlation of 17.00% and an average correlation difference of −7.86% (Figure 9A and Figure 9B).

3.6 Stomach Cancer

For stomach cancer, the genes *LRP1B*, *FAT4*, *ERBB4*, *ZFHX3*, and *CREBBP* had the highest correlation to the related general Signatures: 1, 2, 5, 13, 15, 17, 18, 20, 21, 26, and 28 (Figure 7F; Table S2).

As shown in Figure 8F and Table S2, *FAT4* had greater correlation than the median of the random genes for Signatures 1 (+29.1%), 5 (+25.0%), 15 (+10.1%), 17 (+24.2%), 18 (+1.6%), 20 (+39.3%), 21 (+17.1%), 26 (+22.8%), and 28 (+32.5%). Random genes had better correlation values for Signatures 2 and 13. *ZFHX3* had greater correlation than the median of the random genes for Signatures 1 (+55.5%), 2 (+6.8%), 5 (+11.7%), 15 (+68.5%), 17 (+2.8%), 18 (+0.4%), 20 (+25.3%), 21 (+14.4%), 26 (+19.7%), and 28 (+1.1%). Random genes had better correlation values for Signature 13. *CREBBP* had greater correlation than the median of the random genes for Signatures 1 (+16.6%), 15 (+48.3%), and 20 (+17.2%). Random genes had better correlation values for Signatures 2, 5, 13, 17, 18, 21, 26, and 28 (Table S2). Stomach cancer had an average correlation of 19.82% and an average correlation difference of −0.67% (Figure 9A and Figure 9B).

3.7 Lung Cancer

For lung cancer, there was no NG data for *KEAP1*, so we excluded the case from the simulation for lungcancer. The genes *KRAS*, *PIK3CA*, *ERBB2*, *ALK*, *FAT1*, *FAT4*, *CDKN2A*, and *ERBB4* had the highest correlation to the related general Signatures: 1, 2, 4, 5, 6, 13, 15, and 17 (Figure 7G; Table S2).

As shown in Figure 8G and Table S2, *ERBB2* had greater correlation than the median of the random genes for Signatures 1 (+48.8%), 2 (+22.2%), 6 (+34.4%), 13 (+0.1%), 15 (+46.8%), and 17 (+1.3%). Random genes had better correlation values for Signatures 4 and 5. *FAT1* had greater correlation than the median of the random genes for Signatures 1 (+23.4%), 4 (+27.3%), 5 (+35.8%), 6 (+29.9%), 13 (+18.0%), 15 (+19.2%), and 17 (+5.8%). Random genes had better correlation values for Signature 2. *FAT4* had greater correlation than the median of the random genes for Signatures 2 (+24.8%), 4 (+35.4%), 5 (+21.8%), 13 (+24.2%), and 17 (+5.2%). Random genes had better correlation values for Signatures 1, 6, and 15. Lung cancer had an average correlation of 20.82% and an average correlation difference of –3.67% (Figure 9A and Figure 9B).

3.8 Kidney Cancer

For kidney cancer, the genes *VHL*, *ZFHX3*, and *ARID1A* had the highest correlation to the related general Signatures: 1, 2, 5, 6, 13 and 27 (Figure 7H; Table S2).

As shown in Figure 8H and Table S2, *VHL* had greater correlation than the median of the random genes for Signatures 1 (+19.6%), 5 (+20.6%), 6 (+26.1%), and 27 (+7.5%). Random genes had better correlation values for Signature 2 and 13. *ZFHX3* had greater correlation than the median of the random genes for Signatures 1 (+33.1%), 2 (+5.1%), 5 (+15.3%), and 6 (+23.0%). Random genes had

better correlation values for Signatures 13 and 27. *ARID1A* had a greater correlation than the median of the random genes for Signatures 1 (+14.6%,) and 6 (+19.6%). Random genes had better correlation values for Signatures 2, 5, 13, and 27. Kidney cancer had an average correlation of 17.98% and an average correlation difference of −4.26% (Figure 9A, B).

3.9 Endometrial Cancer

For endometrial cancer, the genes *PTEN*, *PIK3CA*, *TP53*, *KRAS*, *ARID1A*, *CTCF*, *PPP2R1A*, *PIK3R1*, *ZFHX3*, *FBXW7*, *FAT1*, and *FAT4* had the highest correlation to the related general Signatures: 1, 2, 5, 6, 10, 13, 14, and 26 (Figure 7I; Table S2).

As shown in Figure 8I and Table S2, *CTCF* had greater correlation than the median of the random genes for Signatures 2 (+26.8%), 5 (+26.0%), 10 (+40.0%), 13 (+10.0%), 14 (+13.5%), and 26 (+8.7%). Random genes had better correlation values for Signatures 1 and 6. *ZFHX3* had greater correlation than the median of the random genes for Signatures 1 (+46.9%), 2 (+3.1%), 5 (+24.6%), 6 (+64.7%), 10 (+18.4%), 14 (+59.0%), and 26 (+21.9%). Random genes had better correlation values for Signature 13. *FAT4* had greater correlation than the median of the random genes for Signatures 1 (+20.9%), 2 (+14.7%), 5 (+16.5%), 6 (+4.7%), 10 (+65.7%), 13 (+7.9%), 14 (+18.4%), and 26 (+12.8%). Endometrial cancer had an average correlation of 23.09% and an average correlation difference of 2.47% (Figure 9A and Figure 9B).

3.10 Cervix Cancer

For cervical cancer, the genes *FBXW7*, *ARID1A*, *CREBBP*, *ZFHX3*, *FAT4*, and *HRAS* had the highest correlation to the related general Signatures: 1, 2, 5, 6, 10, 13, and 26 (Figure 7J; Table S2).

As shown in Figure 8J and Table S2, *CREBBP* had greater correlation than the median of the random genes for Signatures 1 (+37.0%), 2 (+14.9%), 6 (+11.2%), and 13 (+4.1%). Random genes had better correlation values for Signatures 5, 10, and 26. *ZFHX3* had greater correlation than the median of the random genes for Signatures 1 (+29.3%), 2 (+56.0%), 5 (+0.9%), 6 (+27.1%), 10 (+1.0%), 13 (+6.6%), and 26 (+2.5%). *FAT4* had greater correlation than the median of the random genes for Signatures 1 (+17.8%), 2 (+0.1%), and 10 (+30.7%). Random genes had better correlation values for Signatures 5, 6, 13, and 26. Cervical cancer had an average correlation of 17.40% and an average correlation difference of −2.48% (Figure 9A and Figure 9B).

3.11 Ovarian Cancer

For ovary cancer, the genes *KRAS*, *PIK3CA*, *BRAF*, *CTNNB1*, *PTPRT*, and *NF1* had the highest correlation to the related general Signatures: 1, 3, and 5 (Figure 7K; Table S2).

As shown in Figure 8K and Table S2, *KRAS* had a greater correlation than the median of the random genes for Signatures 3 (+18.0%). Random genes had better correlation values for Signatures 1 and 5. *CTNNB1* had a greater correlation than the median of the random genes for Signatures 1 (+0.7%), 3 (+19.4%), and 5 (+14.7%). *NF1* had a greater correlation than the median of the random genes for Signatures 1 (+36.1%) and 5 (+4.9%). Random genes had better correlation values for Signature 3. Ovarian cancer had an average correlation of 17.24% and an average correlation difference of −9.63% (Figure 9A and Figure 9B).

3.12 Adrenal Gland Cancer (Neuroblastoma)

For adrenal gland cancer, there was no NG data for *DAXX*, so we excluded the case from the simulation. The genes *CTNNB1* and *ATRX* had the highest correlation to the related general Signatures: 1, 5, and 18 (Figure 7L; Table S2).

As shown in Figure 8L and Table S2, *NF1* had a greater correlation than the median of the random genes for Signatures 1 (+5.7%), 5 (+11.0%), and 18 (+2.6%). *ATRX* had a greater correlation than the median of the random genes for Signatures 1 (+36.7%) and 5 (+0.7%). Random genes had better correlation values for Signature 18. *MEN1* had a greater correlation than the median of the random genes for Signatures 1 (+10.7%), 5 (+4.6%), and 18 (+3.7%). Adrenal gland cancer had an average correlation of 15.26% and an average correlation difference of −6.93% (Figure 9A and Figure 9B).

3.13 Bone Cancer (Osteosarcoma)

For bone cancer, there was no NG data for *H3F3B*, so we excluded the case from the simulation. The genes *IDH1*, *GNAS*, *PTCH1*, *COL2A1*, *STAG2*, *CTNNB1*, and *SMO* had the highest correlation to the related general Signatures: 1, 2, 5, 6, 13, and 30 (Figure 7M; Table S2).

As shown in Figure 8M and Table S2, *COL2A1* had more significant correlation than the median of the random genes for Signatures 1 (+9.9%), 5 (+4.1%), 6 (+15.3%), and 30 (+17.7%). Random genes had better correlation values for Signatures 2 and 13. *STAG2* had a greater correlation than the median of the random genes for Signatures 1 (+19.9%), 2 (+5.9%), 5 (+0.6%), and 6 (+16.4%). Random genes had better correlation values for Signatures 13 and 30. *CTNNB1* had a greater correlation than the median of the random genes for Signatures 2 (+43.1%), 5 (+1.6%), 13 (+31.4%), and 30 (+8.9%). Random genes had better correlation values for Signatures 1 and 6. Bone cancer had an average correlation of 15.94% and an average correlation difference of −5.63% (Figure 9A and Figure 9B).

3.14 Prostate Cancer

For prostate cancer, the genes *ZFHX3*, *GRIN2A*, *SMARCA4*, and *NF1* had the highest correlation to the related general Signatures: 1, 5, and 6 (Figure 7N; Table S2).

As shown in Figure 8N and Table S2, *ZFHX3* had a greater correlation than the median of the random genes for Signatures 1 (+34.6%), 5 (+13.2%), and 6 (+61.1%). *GRIN2A* had a greater correlation than the median of the random genes for Signatures 1 (+8.6%), 5 (+1.2%), and 6 (+35.3%). *SMARCA4* had a greater correlation than the median of the random genes for Signatures 1 (+9.3%) and 6 (+12.2%). Random genes had better correlation values for Signature 5. Prostate cancer had an average correlation of 17.71% and an average correlation difference of −10.65% (Figure 9A and Figure 9B).

3.15 Thyroid Cancer

For thyroid cancer, the genes *NF1* and *RB1* had the highest correlation to the related general Signatures: 1, 2, 5, and 13 (Figure 7O; Table S2).

As shown in Figure 8O and Table S2, *TP53* had greater correlation than the median of the random genes for Signature 2 (+17.9%). *KMT2C* had a greater correlation than the median of the random genes for Signatures 1 (+20.5%), 2 (+2.2%), 5 (+11.3%), and 13 (+16.0%). *NF1* had greater correlation

than the median of the random genes for Signatures 2 (+54.2%) and 13 (+8.9%). Random genes had better correlation values for Signatures 1 and 5. Thyroid cancer had the lowest average correlation of 16.79% and an average correlation difference of −9.63% (Figure 9A and Figure 9B).

3.16 Eye Cancer

For eye cancer, the genes *GNAG*, *TP53*, and *PTEN* had the highest correlation to the related general Signatures: 1, 5, and 6 (Figure 7P; Table S2).

As shown in Figure 8P and Table S2, *TP53* had a greater correlation than the median of the random genes for Signatures 1 (+10.6%) and 6 (+14.2%). Random genes had better correlation values for Signature 5. *KMT2D* had a greater correlation than the median of the random genes for Signatures 1 (+13.9%) and 6 (+7.0%). Random genes had better correlation values for Signature 5. *EGFR* had a greater correlation than the median of the random genes for Signature 1 (+22.1%). Random genes had better correlation values for Signatures 5 and 6. Eye cancer had an average correlation of 14.79% and the lowest average correlation difference of −9.9% (Figure 9A and Figure 9B).

4. Discussion

The main goal of this study was to elucidate the correlation between the mutational signatures of the genes most-frequently mutated in specific cancer and general mutational signatures of this cancer. To do this, we created a program—SignaGen—that calculates the mutational signatures of these genes given the genes' sample DNA sequence and mutation datasets, and calculates their correlation with 30 general signatures of cancer. SignaGen creates a heatmap of the correlation results between the mutational signatures of the most frequently mutated genes (MFMGs) in a specific cancer and the general signatures for this type of cancer. It also has a feature that displays a 3D model for a comparison between the correlation values for this cancer's MFMGs and the correlation values of random genes of similar lengths.

We hypothesized that the top 20 MFMGs of a specific cancer type would have the highest correlation values with the general signatures related to that cancer type. To explore this hypothesis, we analyzed the correlation values between the mutational signatures of the 20 MFMGs of a cancer type, assuming that all of them exist, and the 30 general signatures. Through this analysis, we observed that some of the top 20 MFMGs for a specific cancer had much higher correlation values than the other genes included in top 20 genes of the same cancer type. For example, *FAT4* and *GRIN2A* had much higher correlation values than many of the other top 20 MFMGs for skin cancer (Figure 7B). We could also observe that the top 20 genes had noticeably higher correlation values with some of the related general signatures than other related general signatures (e.g., genes' correlation with Signature 11 vs genes' correlation with Signature 17 in (Figure 7B). In some cases, the correlation between the mutational signatures of the genes with the related signatures were lower than the correlation between the genes and general signatures that were not found to be related to the cancer type (e.g., correlation results with Signature 17 as opposed to results with Signatures 19, 23, and 30 in Figure 7B). We expected the top 20 MFMGs to have the highest correlation with the related signatures, but instead found that some of the related signatures had relatively low correlation values with some or all the genes. Signature 17 is a good example of this discrepancy for skin cancer. Its low correlation with the entire top 20 MFMGs may

signify that it is not as related to skin cancer the same way as Signatures 1, 5, 7, and 11, assuming that our hypothesis is true. High correlation values with signatures that weren't found to be related to skin cancer such as Signatures 19, 23, and 30 in Figure 7B also brings up the questions of whether or not these signatures actually are related to skin cancer, or if the high correlation values for these unrelated signatures are a result of different cancer types sharing the same mutations for the gene. It can also be caused of possible correlation of the intron parts of the genes.

Although the top 20 MFMGs had low correlation with some of the related signatures, our hypothesis would still be supported if the top 20 MFMGs had greater correlation values than random genes. Therefore, we used a random gene set to compare the correlation values of three of the top twenty genes with the highest correlation values to random genes with lengths similar to that of the three selected genes. By doing this analysis, we were able to observe whether the top 20 MFMGs had the higher correlation values or not. In a majority of the cases, the three MFMGs selected for each cancer type did have greater correlation values than the median of the random genes they were compared to. Even for Signature 17 for skin cancer, which we originally thought had too low of a correlation value with the top twenty genes, had a greater correlation with the mutational signature of the three genes with the highest correlation values. However, there was also a small number of cases in which the median of the random genes' correlation were higher than that of the three genes as well.

The analysis of our results mostly supports our hypothesis that the MFMGs of a particular cancer have a higher correlation with the general signatures related to these cancer types, although there were some outliers that did not match the results of past studies. In addition, it was a good opportunity to indirectly confirm the validity of the results of this research through our comparison with random genes.

The program—SignaGen, which we developed, could be a useful tool for future studies. By finding which of the genes have a high correlation with a general signature, we might be able to determine more clear relations of the mutated genes to the type of cancer. SignaGen also could be used to determine what pattern of mutations in a gene will cause it to be related to a certain cancer type. Given a developing cancer patient's DNA sequence data, we can predict which cancer type they may develop and distribute the appropriate treatment to prevent it from developing at an early stage.

5. Conclusions

Our study elucidates the correlations of mutational signatures in frequently mutated cancer genes with general mutational signatures previously found for different cancers. We hypothesized that the top twenty most frequently mutated genes (MFMG) of a cancer type would have the highest correlation with the general signatures related to the cancer. We developed a program SignaGen, that takes in genomic sequence data and potential mutation data (consisting of the type, location, and frequency of each missense mutation) to calculate the mutational signatures of genes. Our results show that the MFMG did have relatively higher correlation values with the general signatures that were related to the specified cancer type. However, there were also cases in which the MFMG had lower correlation values with a related signature than other associated signatures. For example, the MFMG of skin cancer had an average correlation of 7.76% with signature 17, while having an average correlation of 43.02% with signature 11. To investigate this inconsistency and

verify the significance of these correlation values, we compared the correlation values of the MFMG to the correlation values of randomly selected genes of similar length. Even if the MFMG's correlation with the related signatures is low, our hypothesis would still be supported if they had a higher correlation than the random genes. We took three of the twenty MFMG for each cancer, compared their correlations with the random genes of similar length for each cancer type, and found that the MFMG had a higher correlation than the random genes for most cases.

Acronyms and Abbreviations

A, adenine; C, cytosine; CDS, CoDing Sequence; COSMIC, Catalogue of Somatic Mutations In Cancer; G, guanine; MFMG, most frequently mutated gene; NG, NCBI Reference Sequence; T, thymine.

Additional Materials

The following additional materials are uploaded at the page of this paper.

1. Table S1: Gene names and lengths used for this study.
2. Table S2: Correlations of genes' mutation signatures with general signatures of specific cancers.

Author Contributions

IFT and JTP contributed to conception and study design, original manuscript preparation and final draft reviewing and editing. JTP and VLK contributed to original manuscript preparation and final draft reviewing and editing. JTP and JAP prepared and analysed the data. All contributed to the final editing and preparation of the manuscript.

Competing Interests

The authors have declared that no competing interests exist.

References

1. Volkova NV, Meier B, González-Huici V, Bertolini S, Gonzalez S, Vöhringer H, et al. Mutational signatures are jointly shaped by DNA damage and repair. Nat Commun. 2020; 11: 2169.
2. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, et al. Mutational signatures associated with tobacco smoking in human cancer. Science. 2016; 354: 618-622.
3. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500: 415-421.
4. Toufektchan E, Toledo F. The guardian of the genome revisited: p53 downregulates genes required for telomere maintenance, DNA repair, and centromere structure. Cancers. 2018; 10: 135.
5. Zhang Y, Cao L, Nguyen D, Lu H. TP53 mutations in epithelial ovarian cancer. Transl Cancer Res. 2016; 5: 650-663.

6. Alexandrova EM, Yallowitz AR, Li D, Xu S, Schulz R, Proia DA, et al. Improving survival by exploiting tumour dependence on stabilized mutant p53 for treatment. Nature. 2015; 523: 352-356.
7. Moll UM. Improving survival by exploiting tumor dependence on stabilized mutant p53 in mouse models. [abstract]. Proceedings of the 107th Annual Meeting of the American Association for Cancer Research; 2016 April 16th-20th; New Orleans, LA, USA. Philadelphia: American Association for Cancer Research.
8. Pinzaru AM, Hom RA, Beal A, Phillips AF, Ni E, Cardozo T, et al. Telomere replication stress induced by POT1 inactivation accelerates tumorigenesis. Cell Rep. 2016; 15: 2170-2184.
9. Sfeir A, Denchi EL. Stressed telomeres without POT1 enhance tumorigenesis. Oncotarget. 2016; 7: 46833-46834.