A Comparison of Genome-wide and Exome-wide Somatic Mutation Patterns in Tumors

Emily Mulhern[1], Robert Vaughn[1], Mikaela Vega[2], Edward Abel[2], Astrid Cerrato[2], Yashwan Kalainesan[2], Roxanne Hinch[2], Sarangan Ravichandran[3] and Brian T. Luke[3]

[1]Summer Internship Program, Advanced Biomedical Computing Center, National Cancer Institute, Frederick, MD, USA. [2]Werner H. Kirsten Student Internship Program, Advanced Biomedical Computing Center, National Cancer Institute, Frederick, MD, USA. [3]Advanced Biomedical Computing Center, Frederick National Laboratory for Cancer Research, Leidos Biomedical Research Inc., Frederick, MD, USA.

ABSTRACT: It is well known that several mutation and repair processes preferentially act on single-stranded DNA and, combined with selection pressure, suggest that the mutation patterns within exomes should be different from the genome-wide pattern. This study tests this hypothesis by comparing exome-wide and genome-wide mutation patterns in seven different tumor samples. These seven tumor samples were selected because they contain at least 2000 somatic autosomal single base substitutions (SBSs) within exonic regions and allow a direct comparison with the genome-wide mutation patterns. To determine whether they are statistically the same, 1000 Bootstrap samples, without replacement, were created to generate files with the same number of SBSs as the exome-wide results for each tumor sample. The genome-wide, exome-wide, and 1000 Bootstrap mutation samples were each used to build a somatic autosomal mutation matrix (SAMM), which captures the mutation pattern in the context of the central position of a pentanucleotide. The Manhattan distances between the 1000 Bootstrap SAMMs and the genome-wide SAMM are used to determine whether the difference between the exome-wide and genome-wide SAMMs is a result of a smaller number of SBSs, or if there are slight differences in the mutation patterns. One of the exome-wide SAMMs is statistically the same as its corresponding genome-wide pattern; the others show slight differences. To determine whether the mutation patterns are similar, a distance-dependent 6-nearest neighbor classifier was used to predict the tissue of origin of the exome sample when compared to 908 genome-wide mutation patterns from 12 different tissue types. Six of the seven exome patterns are nearest neighbors to their corresponding genome-wide mutation patterns, and all seven exome patterns are correctly classified as to their tissue of origin. Therefore, even though there may well be differences in specific mutations between the genome and exome, exome mutations still contain the overall mutation pattern of the whole genome and the tissue of origin.

KEYWORDS: mutation pattern, genomes, exomes, tissue of origin

CITATION: Mulhern et al. A Comparison of Genome-wide and Exome-wide Somatic Mutation Patterns in Tumors. *Journal of Genomes and Exomes* 2016:5 9–16 doi:10.4137/JGE.S39899.

TYPE: Original Research

RECEIVED: April 11, 2016. **RESUBMITTED:** June 6, 2016. **ACCEPTED FOR PUBLICATION:** September 14, 2016.

ACADEMIC EDITOR: Stephen F. Kingsmore, Editor in Chief

PEER REVIEW: Eight peer reviewers contributed to the peer review report. Reviewers' reports totaled 1924 words, excluding any confidential comments to the academic editor.

FUNDING: This work was supported with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN261200800001E. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: brian.luke@nih.gov

Paper subject to independent expert single-blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

Somatic mutations alter DNA sequences that, in turn, may affect protein expression levels and/or function. This protein dysregulation may lead to diseases including cancer. Whole genome sequencing (or genome-wide sequencing [GWS]) of tumor and blood samples from an individual identifies somatic mutations present in the tumor genome. They represent and interplay between the mutations caused by environmental and biological processes[1–5] and various repair mechanisms.[6,7] Certain mutation processes produce a characteristic mutation pattern, or signature,[8,9] such as C→T and CC→TT mutations from exposure to ultraviolet light.

Deep sequencing has shown that there is heterogeneity in the genome across normal[10–15] and tumor cells[16–25] within the same tissue. Exogenous and endogenous processes cause genetic mutations which accumulate over time, but the exact location of the mutation varies from clone to clone. Competitive advantage of a given clone may cause a temporal expansion, suggesting that a single clone may account for a majority of the cells within a given tumor.[24] Whether or not tumors originate from a single cell,[26] exosomes may initiate tumor formation in normal adjacent cells with different mutation patterns through cellular crosstalk.[27,28] These exosomes may also be excreted and prime other tissues to allow the adhesion of circulating tumor cells, leading to metastasis.[29]

Several different approaches have been used to analyze somatic mutations within a tumor. Some studies examine the genes that are affected, potentially with the goal of identifying the small number of driver mutations from those that are passenger mutations.[22,30,31] A second line of investigation examines the overall pattern of mutations, with the hope of identifying mutational processes responsible for the pattern. Other investigations examine mutation patterns within sub-regions of the genome. This investigation extends earlier work,[32] which examined the overall mutation patterns obtained from GWS.

Several earlier studies represented the pattern of mutations by storing the mutation within the context of the central

position of a trinucleotide.[2,33] There are 64 possible trinucleotides, but since GTA, for example, is also TAC on the complementary strand, there are only 32 unique trinucleotides. Since the central nucleotide can be mutated to one of the three other nucleotides, the overall mutation frequencies are represented by a 96-element vector. From a collection of M tumor samples, M 96-element vectors are produced. To determine whether there are mutation patterns common to these mutation patterns, several studies have reduced these M 96-element vectors to a set of k 96-element mutation signatures[2,25,26,34–38] using nonnegative matrix factorization (NMF).[39,40] Using matrix notation, NMF can be written as follows:

$$\mathbf{A} \sim \mathbf{W} \times \mathbf{H} \quad (1)$$

$\mathbf{A}$ is a $96 \times M$ matrix that contains the M 96-element mutation frequency vectors stored column-wise, $\mathbf{W}$ is a $96 \times k$ matrix containing the k mutation signatures, and $\mathbf{H}$ is a $k \times M$ matrix containing the k coefficients for each of the M mutation vectors. NMF is unique in that $\mathbf{W}$ and $\mathbf{H}$ only contain nonnegative numbers. In other words, each 96-element frequency vector is approximated by adding some or all of the k mutation signatures, and no frequencies are reduced during the summation. $\mathbf{W}$ and $\mathbf{H}$ are initially constructed using random nonnegative numbers, and a gradient search is used to update their elements. Programs are available to perform this dimensional reduction.[38,39]

While NMF is a useful method to reduce the full set of M mutation vectors to a smaller set of k signatures, there are some limitations. First, the final set of signatures and coefficients represents the local stationary point relative to the randomly generated starting point. Further tests are necessary to ensure that this is a minimum and not a saddle point, and there is no guarantee that the globally optimum set of signatures and coefficients is obtained. Second, the obtained set of signatures and coefficients is not unique. Given any nonsingular, nonnegative matrix $\mathbf{D}$, the matrix multiplication can be rewritten as follows:

$$\mathbf{W} \times \mathbf{D} \times \mathbf{D}^{-1} \times \mathbf{H} \quad (2)$$

In other words, $\mathbf{W} \times \mathbf{D}$ is an equally valid set of k signatures and represents a linear combination of the original signatures. When they are used with the updated coefficient matrix, given by $\mathbf{D}^{-1} \times \mathbf{H}$, they produce the same approximation to the set of M 96-element frequency vectors, $\mathbf{A}$.

As with any dimensional reduction procedures, NMF loses some information. In other words, the product of $\mathbf{W}$ and $\mathbf{D}$ does not exactly reproduce the 96-element mutation frequencies stored in $\mathbf{A}$ (Equation 1). The residual error in each mutation vector may contribute to differences between mutation patterns in different tumors of the same tissue type. The relative contributions of the signatures to different mutation frequencies measure the similarities between mutation patterns within a given tumor type and across tumor types. Finally, there is no one-to-one correspondence between a given signature and a mutation mechanism. A given mechanism may contribute to more than one signature, and there may be extra mutation frequencies within the signature that are not associated with this mechanism. It has been shown that the contributions of two signatures correlate with the time between tumor diagnosis and extraction,[41] which may be due to a single mutation mechanism.

All of the studies described earlier treat a single base substitution (SBS) in the context of occurring at the central position of a trinucleotide. An examination of the mutation pattern generated by aristolochic acid[3] found that A→T (T→A) transversions occurred within the sequence motif A[C|T]AGG, suggesting that the analysis may have to go beyond the trinucleotide. This led to the development of the somatic autosomal mutation matrix (SAMM).[32] This matrix captures the mutation pattern within a pentanucleotide centered on the altered nucleotide by using a sliding trinucleotide, allowing mutations at the first, second, and third positions. It is a 32×12 matrix where each row represents a unique trinucleotide determined by requiring a purine to reside in the central position. The first three columns represent the mutation to an adenine in the first, second, and third positions (denoted as 1.a, 2.a, and 3.a, respectively), followed by thymine (1.t, 2.t, and 3.t), cytosine (1.c, 2.c, and 3.t), and guanine (1.g, 2.g, and 3.g). For example, if aristolochic acid caused the mutation ACAGG→ACTGG, the first of the three trinucleotide mutations would be ACA→ACT. Since the central position contains a pyrimidine, the mutation on the complementary strand would be considered (TGT→AGT) and is denoted as TGT_1.a. The second trinucleotide considered is CAG→CTG (denoted as CAG_2.t), and the third trinucleotide is AGG→TGG (denoted as AGG_1.t). Since a nucleotide cannot mutate into itself, three elements within each row must be zero. By examining all somatic mutations within a tumor sample, a 32×12 count matrix is constructed. Dividing each row by the number of times each trinucleotide and its complement appear in the reference genome produces a frequency matrix. After scaling all frequencies to sum to 1.0, the SAMM is produced.

As the name implies, the SAMM only contains autosomal mutations. The reason for this is that the count of a number of times each trinucleotide appears in all chromosomes depends on the gender of the sample, and in many instances, this information is not available. In addition, Bootstrap sampling[32] showed that the change in the resulting SAMM became reasonably large if the number of SBSs was less than 2000, so there is a requirement that at least 2000 SBSs should be used to generate the SAMM. Therefore, this procedure could not be used to process some of the somatic mutation datasets examined in other studies.[25] The only other caveat is that the pentanucleotide centered on a given SBS is not allowed to have any other mutations. If two mutations are adjacent, or

separated by a single nucleotide, there is no way to determine which mutation occurred first. This makes a definitive determination of what trinucleotide is mutated impossible.

An earlier study examined the genome-wide somatic mutations of 909 cancer genomes[32,42] and, using a distance-dependent 6-nearest neighbor (DD-6NN) classification algorithm, showed that an SAMM could be used to determine the tissue of origin of the tumor to a high accuracy. It should be noted that this classification algorithm measured the difference between SAMMs and was therefore different from earlier studies that compared mutation signatures,[2,25,26,34–38] which examined the similarity between mutation patterns. These studies used a relatively small set of signatures to approximately recreate the mutation vectors from a large number of tumors. In contrast, the SAMM investigation[32] generated putative mechanistic template mutation matrices (MTMMs), representing oxidative damage, photo damage, ^{5m}CpG deamination, and mutations caused by the action of the APOBEC family of deaminases. This is not a complete list of mutational processes that act on a genome, and these putative MTMMs accounted for at most 58% of all mutation frequencies and, in several tumors, less than 10% of the total SAMMs. On the other hand, the oxidative damage to MTMM had many of the largest contributions in lung cancers, the ^{5m}CpG deamination MTMM was generally larger in pancreatic cancers, and the APOBEC deamination template was the largest in breast cancers. Again, this showed a partial similarity in the SAMMs among the lung, pancreas, and breast SAMMs, respectively, but a large fraction of the mutation frequencies was unaccounted for. Therefore, the Manhattan distance across all of the mutation frequencies was used in the classifier to determine the tissue of origin.

It is known that exonic mutations must undergo selection pressure and that specific mutational mechanisms may act on single-stranded DNA, including strand bias with transcription-coupled repair.[7] It has also been observed that specific mutations exhibit a strand bias within exonic regions but not within whole genome studies.[34] This is not surprising since transcription occurs only on a single strand of a given gene, and this is not true during the mitotic replication of intergenic regions. Even so, there is still the question of whether the overall mutation patterns are different between GWS and exome-wide sequencing (EWS) investigations of somatic mutation.

Determining whether EWS mutation patterns (EWS-SAMM) are statistically similar to their respective GWS patterns (GWS-SAMM) can be accomplished by comparing the difference between EWS- and GWS-SAMMs against 1000 SAMMs randomly generated from the GWS somatic mutations. To determine whether a EWS-SAMM still maintains a mutational pattern consistent with GWS-SAMMs from the same tissue, a DD-6NN classifier is used to compare each EWS-SAMM against 908 GWS-SAMMs representing 12 different tissue types.[32,42]

Methods

SAMM generation. To be consistent with the earlier investigation,[32] we require that each EWS-SAMM represents at least 2000 SBSs. Examining each of the 909 somatic mutation files used in the previous study, only seven files contained a sufficient number of mutations, such that more than 2000 SBSs resided within exonic regions. These seven samples are listed in Table 1, along with the number of SBSs used to generate the original GWS-SAMM and the number of SBSs present within the EWS-SAMM. They represent two pancreatic tumors, two melanomas, one liver tumor, one myeloid leukemia, and one lung tumor.

The construction of the GWS-SAMMs required dividing the observed number of mutations in each position of a given trinucleotide by the number of times that trinucleotide appeared within the human reference genome.[32] Therefore, construction of the EWS-SAMM first required determining the number of times that each unique trinucleotide appears within exonic regions. With this information, the EWS-SAMMs were generated using the procedure outlined earlier. Inherent in this procedure is the assumption that mutation frequencies are similar across exonic regions. It has been shown that mutation rates vary within exonic regions and are higher in genes with low expression levels and that appear late in DNA replication.[22] The trinucleotide counts within these genes could therefore be given a larger weight. Conversely, mutation

Table 1. Description of the seven tumor samples.

SAMPLE	EXAMINED SBSs#	EXONIC SBSs#	SOURCE	TYPE
DO33091	77987	2194	PACA-AU_ICGC	Pancreatic
DO35442	85732	2803	PACA-CA_ICGC	Pancreatic
DO45299	166215	4237	LIRI-JP_ICGC	Liver
DO49530	354522	4233	LAML-KR_ICGC	Myeloid leukemia
LUAD-5V8LT	287174	3943	Lung_Adeno_Sanger	Lung
ME009	263374	4979	Melanoma_Berger	Melanoma
ME044	140333	2006	Melanoma_Berger	Melanoma

Notes: #Examined SBSs is the number of nonadjacent autosomal single base substitutions within the genome examined earlier.[32] #Exonic SBSs are those that reside within exonic regions. Source is the dataset and source of the sample. Type is the tumor/tissue type.

rates were also observed in exonic regions that do not conform to these rules, suggesting that sample-specific processes may be involved.[22] For example, replication timing may be affected by germline mutations in replication timing trait loci.[43] Therefore, each sample may have to be individually examined to determine local mutation frequencies and use this information to obtain weighted trinucleotide frequencies. This would also apply to the examination of whole genome trinucleotide counts. Since the earlier study of GWS-SAMMs[32] used unweighted counts across the entire genome, a comparable method will be used here for exonic regions.

Producing the graphical displays. An in-house program produced a heatmap for each of the SAMMs by generating input for the imaging program *fly* (http://martin.gleeson.com/fly/). The intensity of the red color is determined after multiplying each of the scaled frequencies by 100; a value above 6.5 is represented in full red, values between 6.5 and 5.5 yield a slightly weaker red, and so on. Scaled values below 0.5 are white. The seven EWS and GWS-SAMMs are compared by calculating the Manhattan distance across all matrix elements for every pair of SAMMs. This is just the sum of the absolute difference in mutation frequencies across all matrix elements. The resulting distance matrix is used to produce an unweighted average linkage dendrogram using the program Multidendrogram.[44]

Comparison of the EWS-SAMM and GWS-SAMM. To describe the method used to determine if each EWS-SAMM was statistically similar to its corresponding GWS-SAMM, the pancreatic tumor data from sample DO33091 (PACA-AU_ICGC)[32] will be used as an example. This sample contains 77,987 SBSs that are used to construct the GWS-SAMM, and of these, 2,194 SBSs reside within exonic regions and are used to construct the EWS-SAMM (Table 1). Using a random number generator, 1000 Bootstrap subsets of the tumor mutations are selected; each containing 2194 SBSs. This selection is done without replacement so that all of the SBSs are unique. Each subset is used to construct a SAMM, denoted Boot-SAMM(I). The Manhattan distance between the EWS-SAMM and the GWS-SAMM is compared to the distances between Boot-SAMM(I) and the GWS-SAMM (I = 1, 2, …, 1000). This set of 1000 Bootstrap distances is used to calculate a mean and standard deviation. From this, a z-score is determined for the distance between the EWS-SAMM and the GWS-SAMM, from which a P-value is determined. This procedure was repeated for all of the samples listed in Table 1.

Inferring the tissue of origin. To determine whether the EWS-SAMM still maintained the mutation pattern representing the tissue of origin, or tumor type, the same DD-6NN classifier used in the previous study[32,42] is employed. In this previous study, 909 tumor samples were examined representing 13 tissue types, but during the classification of tissue of origin, only 904 samples were examined from 11 tissue types. The acute lymphoblastic leukemia and acute myeloid leukemia samples were excluded, since there were only one sample and four samples within these tissue types, respectively. Since one of the myeloid leukemia samples is used in this investigation (DO49530), the tissue of origin investigation includes this tissue type and each EWS-SAMM is compared to 908 GWS-SAMMs representing 12 tissue types.

Results

The results of comparing the distance to the GWS-SAMM between the EWS-SAMM and the subset Boot-SAMMs generated using the same number of randomly selected SBSs are shown in Table 2. For each sample, the second column shows the Manhattan distance between the EWS-SAMM and the corresponding GWS-SAMM. The minimum, maximum, and mean distance between the 1000 Boot-SAMMs and the GWS-SAMM are shown in columns 3–5, respectively. The sixth column lists the standard deviation of the 1000 Bootstrap distances from the mean, and the seventh column shows the z-score for the distance between the EWS and GWS-SAMMs. The final column shows the P-value obtained from the z-score.

In six of the seven cases, the EWS-SAMM is further from the GWS-SAMM than any of the 1000 Boot-SAMMs. In fact, they are between 8.6 and 25.2 standard

Table 2. Comparison of the EWS-SAMM and the 1000 Bootstrap SAMMs to the corresponding GWS-SAMM.

SAMPLE	d(EWS-GWS)	MIN[b(BOOT)]	MAX[b(BOOT)]	MEAN	ST.DEV	Z-SCORE	P-VALUE
DO33091	0.2114	0.0814	0.1624	0.1116	0.0116	8.6229	<0.00001
DO35442	0.2328	0.0634	0.1263	0.0918	0.0101	13.9586	<0.00001
DO45299	0.1812	0.0685	0.1184	0.0913	0.0082	10.9923	<0.00001
DO49530	0.3205	0.0858	0.1373	0.1067	0.0085	25.2181	<0.00001
LUAD-5V8LT	0.1308	0.0978	0.1491	0.1218	0.0084	1.0720	0.1419
ME009	0.1743	0.0541	0.1024	0.0727	0.0071	14.2952	<0.00001
ME044	0.2431	0.0944	0.1808	0.1294	0.0132	8.6299	<0.00001

Notes: For each sample in Table 1: d(EWS-GWS) is the Manhattan distance to the GWS-SAMM for the EWS-SAMM; the minimum (min[b(Boot)]), maximum (max[b(Boot)]), and mean distance 1000 Bootstrap-generated SAMMS (mean), the standard deviation of the 1000 Bootstrap distances (SD), the z-score of the EWS-GWS distance from a one-sample t-test (z-score), and the corresponding P-value.

deviations from the mean separation of the Bootstrap samples. These large z-values suggest that there is less than 0.001% chance representing the same mutation pattern. In contrast, the EWS-SAMM from the lung tumor sample LUAD-5V8LT is closer to the GWS-SAMM of this sample than the furthest of the Bootstrap-generated SAMM. When Boot-SAMM(I) is ordered from the closest to the furthest from the GWS-SAMM, the EWS-SAMM lies between the 855th and 856th Bootstrap samples. The resulting z-score is 1.072, producing in a P-value of 0.142. Therefore, one cannot reject the null hypothesis that the EWS-SAMM and GWS-SAMM are same.

The results from the DD-6NN classification of the tissue of origin are presented in Table 3. The third and fourth columns of this table give the sample identification and tissue type of the nearest neighbor GWS-SAMM. For six of the seven EWS-SAMMs, the nearest neighbor is the corresponding GWS-SAMM. In the seventh case, a melanoma sample (ME044) is slightly closer to the GWS-SAMM of another melanoma sample, ME049 (0.203), than to its own GWS-SAMM (0.243).

As in the previous studies,[32,42] the classification of the tissue of origin from the DD-6NN classifier was performed in two ways. The first requires that the membership in a tissue type be at least 0.5 for a definitive classification, and the second uses a maximum likelihood approach where the EWS-SAMM is assigned to the tissue type with the largest membership. These classification results are presented in the fifth and sixth columns of Table 3, respectively. Using the first criterion, five of the seven EWS-SAMMs are assigned to the correct tissue of origin. One of the pancreatic cancer samples (DO35442) and the myeloid leukemia sample (DO49530) received a classification of *Undetermined*, since none of the membership probabilities exceeded 0.50. This is not unexpected in the latter case since the set of 908 GWS-SAMMs contained only four samples of this tissue type, requiring that some of the nearest neighbors be from other tissue types. When a maximum likelihood criterion is used, all seven EWS-SAMMs are assigned to the correct tissue of origin.

A comparison of the membership probabilities in each tissue type for the seven EWS-SAMMs and the six corresponding GWS-SAMMs[32,42] is presented in Supplementary Table 1. Again, it should be noted that the myeloid leukemia sample was not examined in the previous study due to the small number of samples. In three of the six remaining samples, the EWS-SAMM had a larger membership in the correct tissue type than the corresponding GWS-SAMM (DO33091, DO45299, and ME009). For the lung tumor sample (LUAD-5V8LT), the EWS-SAMM membership in the lung tissue type (0.936) is virtually same as the GWS-SAMM (0.938). In the other two samples, DO35442 and ME044, the GWS-SAMM membership in the correct tissue type is larger than the EWS-SAMM.

Dulak et al[45] examined the exome and whole genome mutation patterns of esophageal adenocarcinomas and identified one individual where a specific mutation pattern was significantly different between the genome and exome. In particular, AAN→ACN transversions were significantly higher across the entire genome than within exonic regions, as measured in mutations/Mb. Some of this difference represents true decreases in the mutations in exonic regions, and some is due to the fact that these trinucleotides are present in exons less than expected by chance. For example, the AAT/ATT trinucleotide is present 133,375,539 times in the reference genome. Since 2.56% of all nucleotides reside within exonic regions, one might expect that this trinucleotide to be present 3,414,505 times within exons. In reality, AAT/TAA is only present 2,244,199 times within exonic regions, or only 65.7% of the expected number. For AAA/TTT, AAC/GTT, and AAG/CTT, the exonic counts are 66.9%, 87.2%, and 96.7% of the expected number, respectively. Therefore, some of the observed decrease in mutations/Mb may simply be due to a decrease in the number of these trinucleotides within exonic regions.

Table 3. Results of the distance-dependent 6-nearest neighbor classification of the tissue of origin when each exome-wide SAMM is compared against 908 genome-wide SAMMs representing 12 different tissue types, as well as the identity of the nearest neighbor (NN) genome-wide SAMM.

SAMPLE	TISSUE TYPE	NN_SAMPLE	NN_TISSUE-TYPE	PREDICT	EXP_MAX
DO33091	Pancreatic	DO33091	Pancreatic	Pancreatic	Pancreatic
DO35442	Pancreatic	DO35442	Pancreatic	Undetermined	Pancreatic
DO45299	Liver	DO45299	Liver	Liver	Liver
DO49530	Myeloid Leukemia	DO49530	Myeloid Leukemia	Undetermined	Myeloid Leukemia
LUAD-5V8LT	Lung	LUAD-5V8LT	Lung	Lung	Lung
ME009	Melanoma	ME009	Melanoma	Melanoma	Melanoma
ME044	Melanoma	ME049	Melanoma	Melanoma	Melanoma

Notes: Tissue type is the known tissue/tumor type of the sample. NN_Sample is the nearest neighbor GWS-SAMM. NN_Tissue-Type is the tissue/tumor type of this nearest neighbor GWS-SAMM. Predict is the predicted tissue/tumor type requiring a membership of at least 50% in the classifier. Exp_Max is the predicted tissue/tumor type with the largest membership value.

To test this hypothesis, the previous investigation of genome-wide somatic mutations[32] examined 16 esophageal adenocarcinomas with at least 2000 somatic autosomal SBSs (ESAD-UK from ICGC). Of these, 12 contained at least one A→C mutation for each of the four AAN trinucleotides. Supplementary Table 2 gives the binomial probability of observing the reported number of mutations or less for each trinucleotide in each sample using both the genome-wide probability of a trinucleotide residing within an exonic region (0.0256) and the probability based on the ratio of the number of counts of each trinucleotide in exonic regions to the entire genome. The former will give results similar to those reported in Dulak et al,[45] and the latter employs the procedure used in calculating a SAMM.[32] In four cases, changing the probability of observation caused the AAT→ACT mutation to vary from significant ($\alpha = 0.05$) to not significant. For DO10850, this mutation is not significantly reduced within exonic regions using either probability, and for DO10852, only the AAG→ACG is significantly lower within exonic regions using either probability.

Discussion

The heatmaps for the EWS- and GWS-SAMMs are shown in Supplementary Figure 1 for each of the samples. The frequencies for each of the top 10 mutations are also included below each heatmap. The two pancreatic cancer samples (DO33091 and DO35442) show strong signals representing ^{5m}CpG deamination.[32] This mechanism is consistent with the eight highest mutation frequencies in both the EWS- and GWS-SAMMs. They both also show a pattern of C→T transitions within the motif NUC (any nucleotide followed by a purine and cytosine). This motif accounts for the last two mutations shown in Supplementary Figure 1 for both samples. Similar patterns are present in the EWS- and GWS-SAMM heatmaps of the liver sample, DO45299. In contrast to the pancreatic samples, the EWS- and GWS-SAMMs for DO45299 also show a pattern consistent with NAC→NGC and TAN→TGN. TAC→TGC, which is the common mutation to both motifs, has one of the top 10 frequencies in both the EWS- and GWS-SAMMs. The heatmaps for the leukemia sample, DO49530, contain many of the same mutations with large frequencies as the other heatmaps. They differ from the others in the mutations with smaller frequencies. For example, both the EWS and GWS heatmaps contain mutations of the form NAT→NAC and NAT→NGT. The EWS and GWS heatmaps for the lung sample, LUAD-5V8LT, are extremely similar to each other and very different from the others. The first observation is the lack of a strong ^{5m}CpG deamination signal. In addition, all of the top 10 mutation frequencies correspond to G→T (C→A) transversions. Mutations at a guanine are consisted with oxidative damage caused by cigarette smoke.[46] The EWS and GWS heatmaps for the two melanoma samples are also very different from those of the other samples. There is a strong similarity between the EWS and GWS heatmaps for each sample, but the heatmaps for ME009 are considerably different from those for ME044.

To elucidate the similarities and differences of the EWS- and GWS-SAMMs further, the dendrogram produced from an unweighted average linkage clustering is shown in Supplementary Figure 2. As a point of reference, since the SAMMs have frequencies scaled to sum to 1.0, the maximum theoretical distance between any pair of SAMMs is 2.0. Starting from the left of the dendrogram, each pancreatic sample first clusters the EWS-SAMM with its GWS-SAMM, and then they cluster together. As described earlier, the liver sample has mutation frequencies similar to the pancreatic samples, so the EWS-SAMM and GWS-SAMM first cluster together and then join the cluster with the pancreatic samples. This occurs at an average distance slightly above 0.4. The leukemia sample shows the largest difference between the EWS- and GWS-SAMMs (Table 2), but they still cluster with each other before joining the cluster containing the pancreatic and liver SAMMs, at a distance above 0.6. The EWS- and GWS-SAMMs for each melanoma sample are similar, but the two samples show different patterns from each other. They form a melanoma cluster at an average distance of about 0.7, which does not merge with the above samples until the average distance threshold exceeds 1.2. As stated earlier, the EWS- and GWS-SAMMs for the lung sample are very similar and have the smallest EWS-GWS distance (Table 2). They are also very different from the other SAMMs and only cluster with them at an average distance above 1.4.

Though many of the SAMMs from different tissues have similar patterns in the mutations with the largest frequencies, they would not contribute to the Manhattan distance used to determine the tissue of origin. It is where the SAMMs disagree that determines this distance. All seven EWS-SAMs are predicted to belong to the correct tissue of origin, and for six of them, the nearest neighbor was the corresponding GWS-SAMM (Table 3).

A comparison of the membership probabilities in each tissue type for the seven EWS-SAMMs and the six corresponding GWS-SAMMs[32,42] is presented in Supplementary Table 1. Again, it should be noted that the myeloid leukemia sample was not examined in the previous study due to the small number of samples. For the six other samples, the EWS-SAMM had a larger membership in the correct tissue type than the corresponding GWS-SAMM for three of the samples (DO33091, DO45299, and ME009). For the lung tumor sample (LUAD-5V8LT), the EWS-SAMM membership in the lung tissue type (0.936) is virtually same as the GWS-SAMM (0.938). In the other two samples, DO35442 and ME044, the GWS-SAMM membership in the correct tissue type was larger than the EWS-SAMM.

As shown in Supplementary Table 2, differences in mutations between exons and the entire genome, when measured

as mutations per megabase, may simply be due to differences in the frequency of specific trinucleotides in different genomic regions. This is not to suggest that AAN→CAN transversions are unimportant in esophageal adenocarcinomas. When the 16 ESAD-UK somatic mutation files are represented as SAMMs, the AAG→ACG transversion had the highest frequency in six tumor samples and was in the top 10 for 12 of the 16 tumors (Supplementary Table 3). The AAC→ACC mutation was in the top 20 mutation frequencies for 11 of the 16 samples, but the AAA→ACA and AAT→ACT mutation frequencies were not in the top 20 for any sample.

If mutation counts per megabase are used,[25,45,47] the results are different (Supplementary Table 4). AAG→ACG has the highest mutation count in 12 of the 16 tumor samples; AAA→ACA, AAT→ACT, and AAC→ACC mutation counts are in the top 20 for 11, 7, and 10 of the 16 samples, respectively. This suggests that using counts per megabase exaggerates the importance of these specific mutations and that mutation frequencies using observed trinucleotide counts are a better measure.

This present study examined the overall mutation patterns within exons and the whole genome. A similar methodology could be used to examine mutation patterns within other regions of the genome such as introns, but care must be used. Mutation rates are known to vary within specific regions of the genome,[47] such as with chromatin organization.[48] For example, there is a higher probability of a CpG island residing upstream of a gene[49] and these regions are known to resist methylation. Therefore, the frequency of ^{5m}CpG deamination is expected to be lower in promoter regions than in other areas of the genome.[47] In addition, mutation frequencies correlate with DNA replication timing and transcription rate,[22] but these should be consistent within a specific tumor type across individuals. The only caution would be to ensure that the genomic region being considered is not too restrictive, such that there are fewer than 2000 SBSs. Each analysis would require determining the counts for each of the trinucleotides in the selected regions, so that the frequency of observation is determined relative to the trinucleotide counts, not per megabase.

The fact that these EWS-SAMMs contain a similar mutation pattern as the corresponding GWS-SAMMs may be counterintuitive, given that different mutational processes may be operating and exome mutation would be under a higher selection pressure. This finding suggests that insight into the mutational processes at work within a particular tumor type can be obtained from less-expensive EWS studies. This also justifies an earlier study that obtained more than 20 mutational signatures from 7042 cancers that contained both GWS and EWS results,[25] though that study used mutations per megabase, which may unbalance the results. Future investigations will use both GWS and EWS mutation studies to determine if there are tissue-specific mutational patterns present within different tumor types.

Conclusion

This study shows that for six of the seven samples, the EWS-SAMM is statistically different from the corresponding GWS-SAMM, while for the seventh sample (LUAD-5V8LT), the difference is not statistically significant ($P = 0.142$). Requiring a membership of 0.50 or more with a DD-6NN classifier correctly identified the tissue of origin for five of the seven EWS-SAMMs, and when a maximum likelihood criterion was used, the correct tissue of origin was identified in all seven samples. Therefore, although an EWS-SAMM may not be statistically same as its corresponding GWS-SAMM, it still contains the mutational patterns representing the tissue of origin. Therefore, different mutations and repair mechanisms acting on single-stranded or double-stranded DNA do not cause a strong enough change in the overall mutation patterns in exonic regions, or over the entire genome, to hide the underlying pattern of the tissue of origin.

Acknowledgments

The content of this publication does not necessarily reflect the views of policies of the Department of Health and Human Services, or does mention of trade names, commercial products, or organizations imply endorsement by the US government.

Author Contributions

Conceived and designed the study: BTL. Analyzed the data: EM, RV, MV, EA, AC, YK, and RH. Wrote the first draft of the article: BTL. Agreed with the study results and conclusions: SR. All authors reviewed and approved the final article.

Supplementary Materials

Supplementary table 1. A comparison of the tissue of origin classification of (**a**) the exome-wide SAMM and (**b**) the corresponding genome-wide SAMM.

Supplementary table 2. Probability that the A>C mutations in AAN is observed by chance given two different probabilities of observation.

Supplementary table 3. Rank of the A>C mutation frequencies in AAN when determined by the counts of the trinucleotide within exons.

Supplementary table 4. Rank of the A>C mutation counts in AAN, determined as mutations per megabase.

Supplementary figure 1. Heatmaps of the EWS-SAMMs and GWS-SAMMs for each of the seven samples considered and the top 10 mutation frequencies. The rows correspond to the 32 unique trinucleotides containing a purine in the center position, and the columns represent the specific mutations. The intensity of the red color indicates the magnitude of the specific mutation frequency.

Supplementary figure 2. Dendrogram obtained from an unweighted average linkage clustering of the seven EWS-SAMMs and GWS-SAMMs. A Manhattan distance between the SAMMs was used, which is the sum of the absolute difference in mutation frequencies across all elements of the mutation matrices.

REFERENCES

1. Neeley WL, Delaney JC, Henderson PT, Essigmann JM. In vivo bypass efficiencies and mutational signatures of the guanine oxidation products 2-aminoimidazolone and 5-guanidino-4-nitroimidazole. *J Biol Chem*. 2004; 279(42):43568–43573.
2. Nik-Zainal S, Alexandrov LB, Wedge DC, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*. 2012;149(5):979–993.
3. Poon SL, Pang ST, McPherson JR, et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci Transl Med*. 2013; 5(197):197ra101.
4. Taylor BJ, Nik-Zainal S, Wu YL, et al. DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *Elife*. 2013;2:e00534.
5. Nik-Zainal S, Wedge DC, Alexandrov LB, et al. Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat Genet*. 2014;46(5): 487–491.
6. Wilson DM III, Bohr VA. The mechanics of base excision repair, and its relationship to aging and disease. *DNA Repair*. 2007;6(4):544–559.
7. Shuck SC, Short EA, Turchi JJ. Eukaryotic nucleotide excision repair: from understanding mechanisms to influencing biology. *Cell Res*. 2008;18(1):64–72.
8. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet*. 2014;15(9):585–598.
9. Nik-Zainal S, Kucab JE, Morganella S, et al. The genome as a record of environmental exposure. *Mutagenesis*. 2015;30(6):763–770.
10. Martincorena I, Jones PH, Campbell PJ. Constrained positive selection on cancer mutations in normal skin. *Proc Natl Acad Sci U S A*. 2016;113(9):E1128–E1129.
11. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. *Science*. 2015;349(6255):1483–1489.
12. Martincorena I, Roshan A, Gerstung M, et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*. 2015;348(6237):880–886.
13. Alcolea MP, Greulich P, Wabik A, Frede J, Simons BD, Jones PH. Differentiation imbalance in single oesophageal progenitor cells causes clonal immortalization and field change. *Nat Cell Biol*. 2014;16(6):615–622.
14. Klein AM, Simons BD. Universal patterns of stem cell fate in cycling adult tissues. *Development*. 2011;138(15):3103–3111.
15. Simons BD. Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis. *Proc Natl Acad Sci U S A*. 2016;113(1):128–133.
16. Matsumoto T, Shimizu T, Takai A, Marusawa H. Exploring the mechanisms of gastrointestinal cancer development using deep sequencing analysis. *Cancers (Basel)*. 2015;7(2):1037–1051.
17. Hong MK, Macintyre G, Wedge DC, et al. Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nat Commun*. 2015;6:6605.
18. Van Loo P, Voet T. Single cell analysis of cancer genomes. *Curr Opin Genet Dev*. 2014;24:82–91.
19. Stirling PC, Shen Y, Corbett R, Jones SJ, Hieter P. Genome destabilizing mutator alleles drive specific mutational trajectories in *Saccharomyces cerevisiae*. *Genetics*. 2014;196(2):403–412.
20. Burrell RA, Swanton C. Tumour heterogeneity and the evolution of polyclonal drug resistance. *Mol Oncol*. 2014;8(6):1095–1111.
21. Burrell RA, Swanton C. The evolution of the unstable cancer genome. *Curr Opin Genet Dev*. 2014;24:61–67.
22. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013;499(7457):214–218.
23. Setlur SR, Lee C. Tumor archaeology reveals that mutations love company. *Cell*. 2012;149(5):959–961.
24. Nik-Zainal S, Van Loo P, Wedge DC, et al. The life history of 21 breast cancers. *Cell*. 2012;149(5):994–1007.
25. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415–421.
26. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev*. 2014;24:52–60.
27. Azmi AS, Bao B, Sarkar FH. Exosomes in cancer development, metastasis, and drug resistance: a comprehensive review. *Cancer Metastasis Rev*. 2013;32(3–4): 623–642.
28. Zhang X, Yuan X, Shi H, Wu L, Qian H, Xu W. Exosomes in cancer: small particle, big player. *J Hematol Oncol*. 2015;8:83.
29. Costa-Silva B, Aiello NM, Ocean AJ, et al. Pancreatic cancer exosomes initiate pre-metastatic niche formation in the liver. *Nat Cell Biol*. 2015;17(6):816–826.
30. Stephens PJ, Tarpey PS, Davies H, et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature*. 2012;486(7403):400–404.
31. Waddell N, Pajic M, Patch AM, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*. 2015;518(7540):495–501.
32. Temiz NA, Donohue DE, Bacolla A, et al. The somatic autosomal mutation matrix in cancer genomes. *Hum Genet*. 2015;134(8):851–864.
33. Burns MB, Temiz NA, Harris RS. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat Genet*. 2013;45(9):977–983.
34. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*. 2013;3(1):246–259.
35. Jia P, Pao W, Zhao Z. Patterns and processes of somatic mutations in nine major cancers. *BMC Med Genomics*. 2014;7:11.
36. Nik-Zainal S. Insights into cancer biology through next-generation sequencing. *Clin Med*. 2014;14(suppl 6):s71–s77.
37. Alexandrov LB. Understanding the origins of human cancer. *Science*. 2015; 350(6265):1175.
38. Gehring JS, Fischer B, Lawrence M, Huber W. Somatic Signatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics*. 2015;31(22): 3673–3675.
39. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*. 2004;101(12): 4164–4169.
40. Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ. Algorithms and applications for approximate nonnegative matrix factorization. *Comput Stat Data Anal*. 2007;52:155–173.
41. Alexandrov LB, Jones PH, Wedge DC, et al. Clock-like mutational processes in human somatic cells. *Nat Genet*. 2015;47(12):1402–1407.
42. Temiz NA, Donohue DE, Bacolla A, et al. Erratum to: the somatic autosomal mutation matrix in cancer genomes. *Hum Genet*. 2015;134(8):865–867.
43. Shendure J, Akey JM. The origins, determinants, and consequences of human mutations. *Science*. 2015;349(6255):1478–1483.
44. Fernandez A, Gomez S. Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *J Classif*. 2008;25:43–65.
45. Dulak AM, Stojanov P, Peng S, et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat Genet*. 2013;45(5):478–486.
46. Lee W, Jiang Z, Liu J, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature*. 2010;465(7297):473–477.
47. Fischer A, Illingworth CJ, Campbell PJ, Mustonen V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol*. 2013;14(4):R39.
48. Schuster-Bockler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*. 2012;488(7412):504–507.
49. Vavouri T, Lehner B. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. *Genome Biol*. 2012;13(11):R110.