

Elucidating the Causes of Structural Variants in the Cancer Genome

Riya Abiram

Lynbrook High School, 1280 Johnson Ave, San Jose, California, 95014, 23, USA; riyaabiram@gmail.com
Mentor: Dr. Azhar Khandekar

ABSTRACT: Research on the associations between structural variation and the appearance of endogenous and exogenous factors common in tumorigenesis allows for a better understanding of the causes of structural variants present in tumor genomes. Analysis of genomic variations from 2600 whole genomes compared to various cancer-inducing factors has found a moderate correlation between the development of structural variants and the increase in TP53 mutations and SBS4 mutational signatures in patients with Lung Adenocarcinoma or Squamous Cell Carcinoma.

KEYWORDS: Computational Biology and Bioinformatics, Genomics, Structural Variants, TP53 Mutations, SBS4 mutational signatures.

■ Introduction

Structural variants (SVs), genomic mutations 50 base pairs or larger, are ubiquitous in many tumor genomes. These large-scale alterations can range from small deletions and amplifications to events involving entire chromosome arms or chromosomes. SVs often have a functional impact, as they can delete key tumor suppressor genes, activate key oncogenes, and disrupt the regulatory landscape of the genome.¹

The analysis of whole-genome sequencing (WGS) data from tumors generated by large-scale studies has generated a catalog of SVs and copy number alterations.^{2,3} Despite these developments, the precise causes of SVs have yet to be discovered, mainly due to the complexity of SVs in cancer genomes. Specifically, the association of endogenous (failure of DNA repair pathways) and exogenous processes (tobacco smoking, ultraviolet light) operative in the tumor genome and the amount of genomic instability have largely not been systematically evaluated across cancer types. Here, we utilize two metrics that measure the amount of genomic instability in tumors: SV burden and percent aberrant genome (PGA)³, a measure of what percent of DNA segments have been amplified or deleted. These two factors are compared to mutational signatures known to be related to exogenous factors. Mutational signatures can be formed by both exogenous or endogenous causes and are developed when somatic mutations in cancer genomes that operate during cell lineage generate a characteristic mutation. These mutations can involve base substitutions and rearrangements of the genome.¹ Therefore, by comparing mutational signature attributions caused by a specific exogenous agent with the SV burden, we can determine whether there is a correlation between SVs and that exogenous agent. Analysis of the processes involved in SV burden may be used in developing future cancer prevention strategies and have the potential to act as biomarkers to inform treatment plans of cancer patients at different stages of their disease.⁴

■ Methods

The data used for analysis was generated by the Pan-Cancer Analysis of Whole Genomes. This data contains simple somatic mutations, copy number somatic mutations, and structural somatic variants. PCAWG also contains mutational signature attributions for each sample.⁵ These attributions were then used to assign the mutation as endogenous or exogenous based on their proposed etiologies.

To calculate the number of aberrant segments in the genome, the number of DNA segments in the cancer genome that were not in a heterozygous diploid state (meaning cells that have undergone loss of heterozygosity, a form of allelic imbalance where heterozygous somatic cells become homozygous through losing an allele⁶) were added together. Loss of heterozygosity in DNA segments is a key indicator of the presence of SVs in the genome.⁷ Therefore, a direct correlation between aberrant segments and an exogenous or endogenous factor could indicate an increase in SVs as well. The SV burden was calculated similarly, with the number of each SV in the cancer genome added together to find the total number of segments.

Both the percent aberrant genome and the total SV burden were then compared to the total number of endogenous, single base substitution (SBS) mutations, as well as the exogenous TP53 mutations and SBS4 mutational signature attributions per sample, to determine possible associations. To analyze the contributions of TP53 mutations to the overall SV burden, all samples were grouped based on whether they contained TP53 mutations or had none. Samples with >0 TP53 were considered smokers, and samples with 0 mutations were considered non-smokers. The same criteria were applied to SBS4 mutational signatures.

Figures were adjusted using the Bonferroni correction method to accurately determine correlation significance relative to the other cancers.

■ Results and Discussion

Significant positive correlations ($p \leq 0.05$ and $R > 0.2$) have been found between endogenous SBS mutations (mutations in which one nucleotide base has been replaced with another) and an increased number of aberrant segments and SV burden in the majority of cancers tested (Figures 1,2).

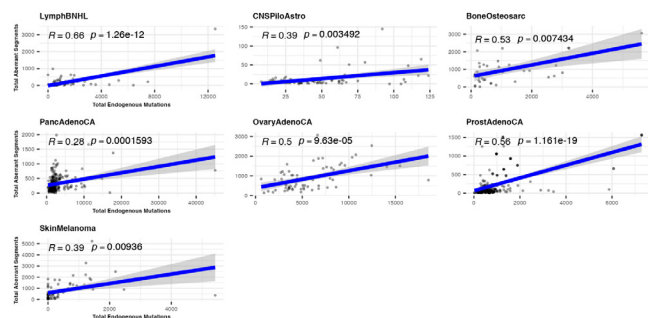


Figure 1: Pearson correlation of total endogenous mutations (x-axis) vs total aberrant segments (y-axis).

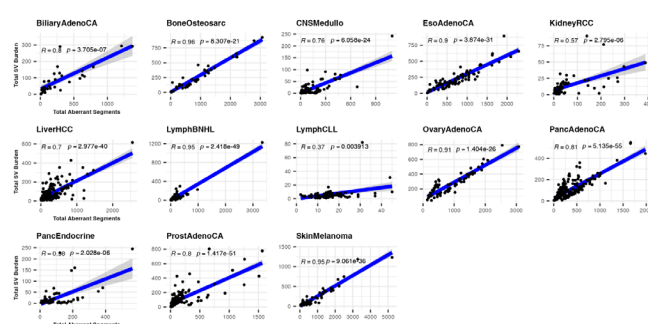


Figure 2: Pearson correlation of total aberrant segments (x-axis) vs total SV burden (y-axis).

The age of the cancer patient also showed no correlation to an increase in SV burden (Figure 3).

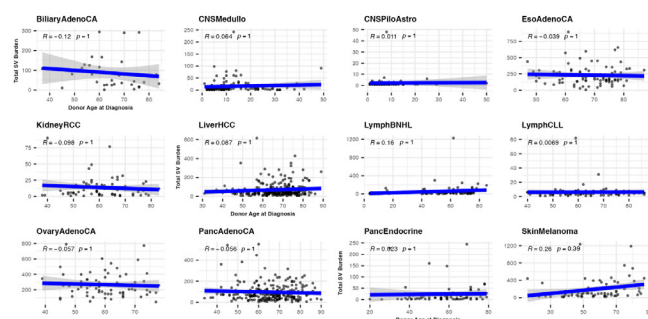


Figure 3: Pearson correlation of donor age at diagnosis (x-axis) vs total SV burden (y-axis).

Lastly, we tested the relationship between mutations in the TP53 gene and genome instability in both Lung Squamous Cell Carcinoma (LUSC) and Lung Adenocarcinoma (LUAD). LUSC samples with TP53 mutations had a higher SV burden than those without TP53 mutations ($p = 4.07 \times 10^{-4}$). (Figures 4,5,6,7).

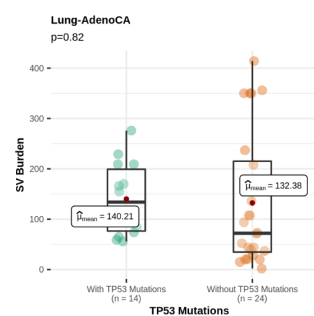


Figure 4a: Mann-Whitney test of TP53 mutations (x-axis) vs. SV burden (y-axis) for Lung Adenocarcinoma samples.

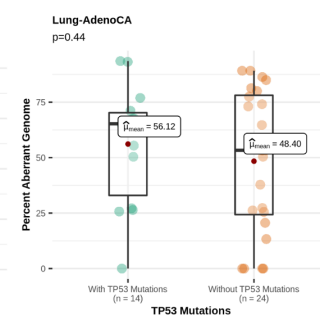


Figure 4b: Mann-Whitney test of TP53 mutations (x-axis) vs. percent aberrant genome (y-axis) for Lung Adenocarcinoma samples.

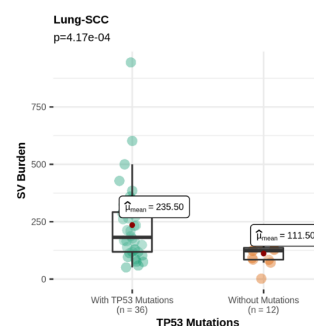


Figure 5a: Mann-Whitney test of TP53 mutations (x-axis) vs. SV burden (y-axis) for Squamous Cell Carcinoma samples.

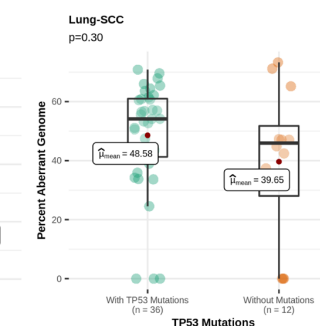


Figure 5b: Mann-Whitney test of TP53 mutations (x-axis) vs. SV burden (y-axis) for Cell Carcinoma samples.

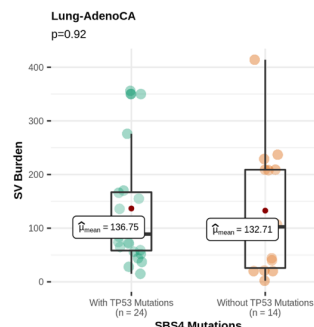


Figure 6a: Mann-Whitney test of SBS4 mutations (x-axis) vs. SV burden (y-axis) for Lung Adenocarcinoma samples.

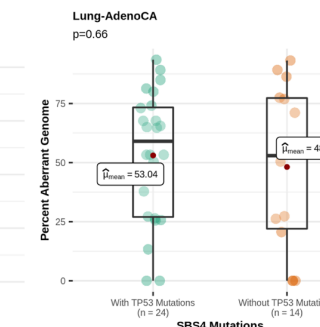


Figure 6b: Mann-Whitney test of SBS4 mutations (x-axis) vs. percent aberrant genome (y-axis) for Lung Adenocarcinoma samples.

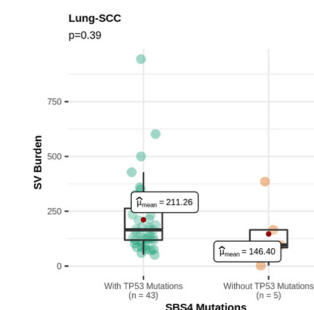


Figure 7a: Mann-Whitney test of SBS4 mutations (x-axis) vs. SV burden (y-axis) for Squamous Cell Carcinoma samples.

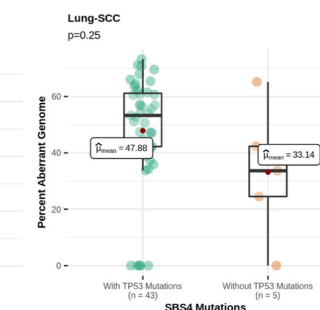


Figure 7b: Mann-Whitney test of SBS4 mutations (x-axis) vs. percent aberrant genome (y-axis) for Squamous Cell Carcinoma samples.

■ Discussion

TP53 mutations, caused largely by smoking, are present in around 50 percent of lung cancers, and SBS4 mutations, caused by alcohol usage and long-term smoking, are present in around 20 percent of all lung cancers. Despite their frequent appearances in these cancers, there was no correlation found between the TP53 and SBS4 mutations and SV. Conversely, endogenous factors correlate more strongly with SV burden than exogenous factors. Endogenous SBS mutations showed strong correlations to the total SV burden in almost every cancer tested. In contrast, exogenous mutations were only shown to be weakly correlated to two cancers (LUSC and LUAD) that are known to be primarily caused by external causes (alcohol use and smoking). Finally, age does not play a major role in the total SV burden, as shown by the weak correlation between the two variables. This indicates that genome instability is not due to the errors from the natural decay of our cells over time, unlike single base substitution signatures 1 and 5, but rather from a specific trigger such as the failure of a repair pathway.

Many samples were disregarded when looking over possible correlations due to the sample not containing specific information necessary for analysis. Conclusions largely fit into similar research currently being conducted by showing a strong link between genomic mutations due to endogenous processes and genome instability.⁸

■ Conclusion

A strong correlation between endogenous mutations and genome instability has been determined based on statistical analysis tests showing the correlation between various endogenous cancer-causing factors and an increase in SV burden. Results also show common causes of tumorigenesis, namely age and exogenous factors such as TP53 mutations and SBS4 mutations, that are not associated with genome instability, indicating that exogenous mutations do not play as large of a role in genome instability. Determining associations between these factors elucidates the causes of genome instability and, therefore, the causes of cancer development.

■ References

1. Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, *et al.* 2020. "The Repertoire of Mutational Signatures in Human Cancer." *Nature* 578 (7793): 94–101. c
2. Li, Yilong, Nicola D. Roberts, Jeremiah A. Wala, Ofer Shapira, Steven E. Schumacher, Kiran Kumar, Ekta Khurana, *et al.* 2020. "Patterns of Somatic Structural Variation in Human Cancer Genomes." *Nature* 578 (7793): 112–21.
3. Mahmoud, Medhat, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. 2019. "Structural Variant Calling: The Long and the Short of It." *Genome Biology* 20 (1): 246.
4. Belzen, Ianthe A. E. M. van, Alexander Schönhuth, Patrick Kemmeren, and Jayne Y. Hehir-Kwa. 2021. "Structural Variant Detection in Cancer Genomes: Computational Challenges and Perspectives for Precision Oncology." *NPJ Precision Oncology* 5 (1): 15.
5. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. "Pan-Cancer Analysis of Whole Genomes." *Nature* 578 (7793): 82–93.
6. Dutra, Amalia, *et al.* 2023. "LOH (Loss of Heterozygosity)." *National Human Genome Research Institute*.

onal Human Genome Research Institute.

7. Schuy, Jakob, Christopher M. Grochowski, Claudia M. B. Carvalho, and Anna Lindstrand. 2022. "Complex Genomic Rearrangements: An Underestimated Cause of Rare Diseases." *Trends in Genetics* 38 (11): 1134–46.
8. Boyling, Alexandra, Gonzalo Perez-Siles, and Marina L. Kennerston. 2022. "Structural Variation at a Disease Mutation Hotspot: Strategies to Investigate Gene Regulation and the 3D Genome." *Frontiers in Genetics* 13 (March): 842860.

■ Author

Riya Abiram is a 12th grader student attending Lynbrook High School in San Jose, California. Riya is extremely passionate about biology and hopes to pursue a career in the medical field in the future.