Perspective

Topological Data Analysis: A New Method to Identify Genetic Alterations in Cancer

Jie Yu[1], Xinzhong Chang[2]

[1]Foreign Languages College, Tianjin Normal University, [2]Department of Breast Surgery, Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China

Corresponding author: Xinzhong Chang, MD. Department of Breast Surgery, Cancer Institute and Hospital, Tianjin Medical University, Tianjin 300060, China. E-mail: drchangxinzhong@126.com

Received: November 23, 2020; Accepted: November 30, 2020; Published: January 29, 2021

ABSTRACT

Cancer is the largest health problem worldwide. A number of targeted therapies are currently employed for the treatment of different cancers. Determining the molecular mechanisms that are necessary for cancer development and progression is the most critical step in targeted therapies. Currently, many studies have identified a large number of frequently mutated cancer-associated genes using recurrence-based methods. However, only the cancer-associated mutations with a mutation frequency >15% can be identified by these methods. In other words, they cannot be used to identify driver genes that have low mutation frequency but play a major role in tumorigenesis and development. Thus, there is an urgent need for a method for identifying cancer-associated genes that are not based on recurrence. In a study, recently published in Nature Communications, research team led by Prof. Raúl Rabadán from the Columbia University successfully devised a novel topological data analysis approach to identify low-prevalence cancer-associated gene mutations using expression data from multiple cancers.

Key words: Cancer-associated gene, genetic alterations, topological data analysis

Cancer is the second leading cause of death worldwide.[1] It is the largest health problem in China.[2] A number of targeted therapies are currently employed for the treatment of different cancers. Determining the molecular mechanisms that are necessary for cancer development and progression is the most critical step in targeted therapies. It is the basis for all targeted therapy strategies.[3] Some landmark cancer genomics programs such as the International Cancer Genome Consortium[4] and the Cancer Genome Atlas[5,6] are possible to screen and identify targets for cancer targeted therapy by systematically compiling genetic alterations of pan-cancer. Cancers driven by common molecular mechanisms all exhibit recurrently altered genes or signaling pathways. Therefore, many studies have identified a large number of frequently mutated cancer-associated genes using computational methods that seek positive selection signatures.[7,8] They have also shown that the mutation frequencies of most therapeutic gene targets are <10%.[9]

It is very difficult to use recurrence-based methods to identify low prevalence cancer-associated mutations, because these methods require a large sample size to obtain statistical power and modeling the background mutation rates is complicated. The incidence rate of neutral mutations in the same cancer type varies greatly in different patients, mutation types or genomic regions, which limits the effectiveness of the recurrence-based methods. In the current ongoing cross-sectional studies that usually consist

Access this article online

Quick Response Code:

Website: www.apjon.org

DOI: 10.4103/2347-5625.308301

Cite this article as: Yu J, Chang X. Topological Data Analysis: A New Method to Identify Genetic Alterations in Cancer. Asia Pac J Oncol Nurs 2021;8:112-4.

of <1000 patients, only the cancer-associated mutations with a mutation frequency >15% can be identified by recurrence-based methods.[9] Therefore, there is an urgent need for a recurrence-based method that can simulate rare events or a method for identifying cancer-associated genes that are not based on recurrence.

The modeling methods that integrate various types of data from cancer are methods of identifying cancer-associated genes that is not based on mutation rate.[10] When a certain gene mutation is accompanied by consistent changes in some types of data from cancer, such as gene expression level, copy number and/or methylation, these changes can be used to correlate the mutation events with cancer development and progression. At present, there have been several studies to identify cancer-associated mutations by organically combining the cis-effects of the mutated genes (e.g., copy number, gene expression level, and methylation). However, it is difficult to identify cancer-associated mutations based on other genetic alterations (trans-effects) other than the changes in mutated gene, because this procedure needs to provide the information about the relationships between known genes to reduce the incidence of false positives. Several methods, such as CaMoDi,[11] DriverNet,[12] and OncoIMPACT,[13] identify cancer-associated mutations by using trans-effects in gene expression. All of them use expression modules, for example, sets of gene networks, functionally related genes, and co-expressed genes, to limit the expression space dimensionality. These algorithms are prone to false positives, because some effects lead to spurious correlations between expression signatures and mutations, for example, cancers with different expression signatures may have different mutation rates[14] and DNA repair enzymes are more likely to enter the genomic regions where the chromatin is opened resulting in an inverse relationship between gene expression levels and mutation rates.[7] However, there is no method that can fully consider the complexity of expression space currently. In a study recently published in *Nature Communications*, titled "Identification of relevant genetic alterations in cancer using topological data analysis,"[15] research team led by Prof. Raúl Rabadán from the Columbia University tried to address these problems by devising a novel approach to identify cancer-associated gene mutations using expression data from multiple cancers.

In this study, the author used topological data analysis to reconstruct the structure of the gene expression space and consider the spurious effects when evaluating the significance of a gene with mutations. In other words, they described the expression profile of a cancer as a point in a high-dimensional expression space by mathematical methods, where the mRNA level of each gene is a dimension, and the number of expressed genes is the dimension of the space. In this space, cancers with similar expression profiles appear as the points lie close to each other. The set of all possible tumors of a cancer type spans a subspace of the expression space. Thus, sampling a limited set of points from the subspace is equivalent to measure individual tumor expression profiles in a cross-sectional study. The authors applied this topological data analysis method to the mutation and expression data of 4476 patients from 12 cancer types, and then identified 95 mutated cancer-associated genes. Among these mutated genes, 38 genes are low-prevalence genes (genes with an average prevalence of ≤5% in the same cancer cohort) that have never been reported in previous studies. Finally, the authors selected ADAM metallopeptidase with thrombospondin type 1 motif 12A (*ADAMTS12*) for further study, because *ADAMTS12* has a low-prevalence inactivating mutation in lung adenocarcinoma. The results show that the lung cancer susceptibility of *ADAMTS12-/-*mice increased fivefold, which confirmed that *ADAMTS12* is an important tumor suppressor gene.

At present, almost all studies about cancer genomic analyses focus on finding out driver genes that are related to cancer initiation and progression by looking for gene mutations with high mutation frequency (recurrence-based methods). Indeed, this strategy is very effective for identifying commonly mutated genes of cancers. However, it cannot be used to identify driver genes that have low mutation frequency but play a major role in tumorigenesis and development. Based on the assumption that mutations in genes are often accompanied by consistent global expression patterns in cancers, Prof. Raúl Rabadán *et al*. proposed a complementary method to recurrence-based approaches. Using this method, the author identified multiple cancer-associated candidate genes that cannot be identified by other methods. One such example is the truncating mutations in the PEST domain of notch receptor 2 (*Notch2*) in breast cancer.[16] These mutation events are very rare and it is impossible to be identified by recurrence-based methods. However, these mutations are always accompanied by global changes in the expression profile of breast cancer. Although these mutations only affect a small proposition of breast cancer patients, they can be the therapeutic targets for Notch signaling pathway inhibitors to treat these patients. Moreover, by using topological data analysis, the author also identified a number of candidate cancer-associated mutations that have not been reported or less studied, such as *ADAMTS12* inactivating mutations occurring in lung cancer. They further characterized the tumor suppressive functions of *ADAMTS12* in *ADAMTS12* knockout mouse models. Overall, this study indicates that data integration

through topological techniques can improve the ability to identify cancer-associated genes with low mutation frequencies.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics. CA Cancer J Clini 2020;70:7-30.
2. Feng RM, Zong YN, Cao SM, Xu RH. Current cancer situation in china: Good or bad news from the 2018 global cancer statistics? Cancer Commun (Lond) 2019;39:22.
3. Yan L, Zhang W. Precision medicine becomes reality-tumor type-agnostic therapy. Cancer Commun (Lond) 2018;38:6.
4. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. Cancer genome landscapes. Science 2013;339:1546-58.
5. Garraway LA, Lander ES. Lessons from the cancer genome. Cell 2013;153:17-37.
6. Shao F, Yang X, Wang W, Wang J, Guo W, Feng X, *et al.* Associations of PGK1 promoter hypomethylation and PGK1-mediated PDHK1 phosphorylation with cancer stage and prognosis: A TCGA pan-cancer analysis. Cancer Commun (Lond) 2019;39:54.
7. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214-8.
8. Wang J, Xi J, Zhang H, Li J, Xia Y, Xi R, *et al.* Somatic mutations in renal cell carcinomas from chinese patients revealed by targeted gene panel sequencing and their associations with prognosis and PD-L1 expression. Cancer Commun (Lond) 2019;39:37.
9. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, *et al*. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014;505:495-501.
10. Cheng F, Zhao J, Zhao Z. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. Brief Bioinform 2016;17:642-56.
11. Manolakos A, Ochoa I, Venkat K, Goldsmith AJ, Gevaert O. CaMoDi: A new method for cancer module discovery. BMC Genomics 2014;15:S8.
12. Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, *et al.* DriverNet: Uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol 2012;13:R124.
13. Bertrand D, Chng KR, Sherbaf FG, Kiesel A, Chia BK, Sia YY, *et al.* Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. Nucleic Acids Res 2015;43:e44.
14. Giacomini CP, Leung SY, Chen X, Yuen ST, Kim YH, Bair E, *et al.* A gene expression signature of genetic instability in colon cancer. Cancer Res 2005;65:9200-5.
15. Rabadán R, Mohamedi Y, Rubin U, Chu T, Alghalith AN, Elliott O, *et al.* Identification of relevant genetic alterations in cancer using topological data analysis. Nat Commun 2020;11:3808.
16. Wang K, Zhang Q, Li D, Ching K, Zhang C, Zheng X, *et al*. PEST domain mutations in Notch receptors comprise an oncogenic driver segment in triple-negative breast cancer sensitive to a gamma-secretase inhibitor. Clin Cancer Res 2015;21:1487-96.