The 2nd International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/4/20220614

Association between the Prognosis of Stomach Adenocarcinoma (STAD) Patients and the Hotspots of Gene Copy Number Variation

Yilin Shi

Guangzhou No.2 High School, Guangzhou,510530, China

shiyilin2010@163.com

Abstract. Stomach Adenocarcinoma, with complex mechanisms about progression, has the biggest proportion in Stomach Cancer, which is one of the most common causes of death in the world. It is widely acknowledged that gene mutations have relationship with Stomach Adenocarcinoma. Copy number variation (CNV), producing identical sequence in a large amount, is associated with Stomach Adenocarcinoma. There are some hotspots of gene mutations in chromosome, CNV is no exception. Their positions may potentially have association with prognosis of patients, which is poorly understood. In this research, windows with fixed length were set; the number of start and end points of CNV in the windows were calculated. The rates of Poisson Distribution in different windows were calculated and the hotspots were found. Then by using proportional hazards model, The author learned the significance of the impact of CNV in certain areas on prognosis with data from TCGA. After analyzing 180 windows, the author found 14 significant windows. These discoveries may reveal new methods of targeted therapy, promoting the precise treatment of Stomach Adenocarcinoma patients.

Keywords: prognosis, stomach adenocarcinoma, copy number variation, hotspots, association

1. Introduction

1.1 A brief introduction of Stomach Adenocarcinoma

Stomach Cancer is one of the most common causes of death in the world. About 95% percent of Stomach Cancer is Stomach Adenocarcinoma(STAD)[1]The patients may feel abdominal pain, loss of appetite and so on[2]. Although some genetic mutations are known to be related to STAD, the mechanisms about cancer progression are still hard to be analyzed[3].

1.2 CNV and prognosis of patients of Stomach Adenocarcinoma

One of the most common mutations is called copy number variation, which lead to cancer initiation and progression[4]. Copy number variation(CNV) produces heavily duplicated, highly identical sequence in chromosome, having impact on human disease and other character[5].In fact, this mutation copy and insert coding or non-coding DNA segments randomly in the genome[6]. It is important for us to assess

DOI: 10.54254/2753-8818/4/20220614

the CNV systematically for its role in sporadic genomic disorders[7]. There are some areas that are easy for CNV to take place called hotspots of CNV[8]. Analyzing the hotspots of CNV and the prognosis of STAD patients may promote the development of targeted therapy on STAD, indirectly lowering the death rate of STAD patients. Recent research report that CNV affect disease susceptibility, proposing studies of different structural variations[9], but the association between start, end point of CNV and the prognosis of STAD patients is poorly understood.

1.3 The Poisson Distribution in this research

There are significant differences in possibilities of mutations in different points[8]. Therefore, in this research, windows with fixed length are set; the number of start and end points of CNV in the windows are calculated. The rates of Poisson Distribution in different windows are calculated and the hotspots are found.

2. Main body

2.1 Methodology

All the data come from TCGA (The Cancer Genome Atlas Program - NCI).

Determining the hotspots (Figure 1). A window with fixed length of 200,000 db is set at the point 0 in the chromosome, the starting point of chromosome (Figure 2). Then calculate the number of start or end points in this window. Consider the midpoint of the window as the coordinate of the window. Then calculate the rate of Poisson Distribution in this window. Next, move this window 200,000 db to the right (Figure 3). Repeat step 2, step 3 and step 4 until the end of the window exceeds or reaches the end of the chromosome. By implementing these steps, all the area in the chromosome will be analyzed. Finally, draw the graph of the coordinates of windows and the rate of Poisson Distribution (Figure 4 is examples of graphs in chromosome 8) according to different chromosomes. The graph of each chromosome will be the standard for determining the hotspots. By doing so in what positions can windows contain more start or end points of CNV can be learned. For example, if there is a series of data of start and end points of CNV: 46509, 200010, 304501, 403921, 495091, 509172, 698192, 798271, 980192, there are 2 points in the window 300000, which is an interval of (200000,400000). Then the rate of the Poisson Distribution can be calculated. The rate of the Poisson Distribution represents the possibility of the mutation. If the rate of the Poisson Distribution is large, it is more likely to happen CNV in this window. And that is a possible hotspot.

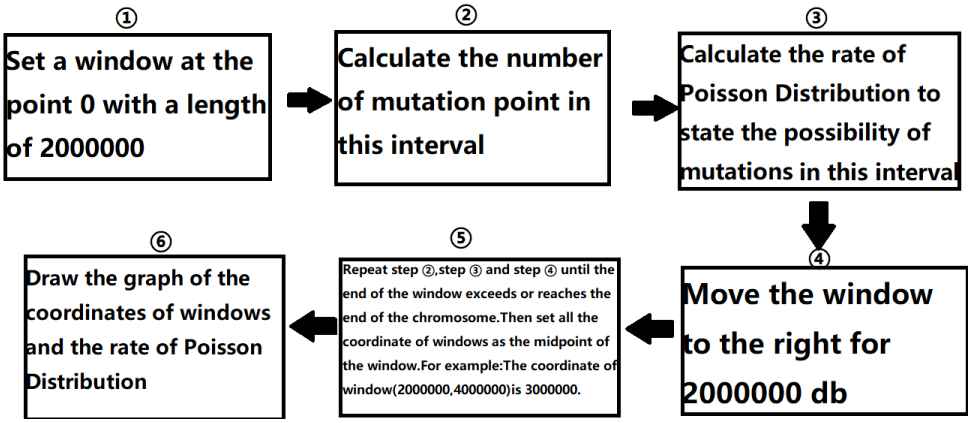

Figure 1. The steps of determining the hotspots (credit: original).

DOI: 10.54254/2753-8818/4/20220614

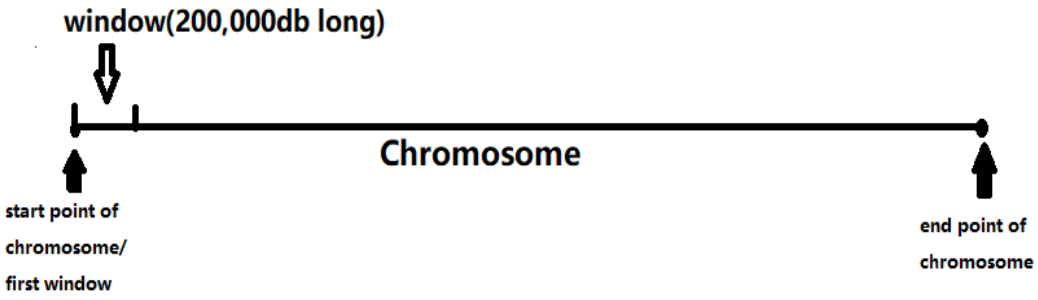

Figure 2. Layout of window in the chromosome (credit: original).

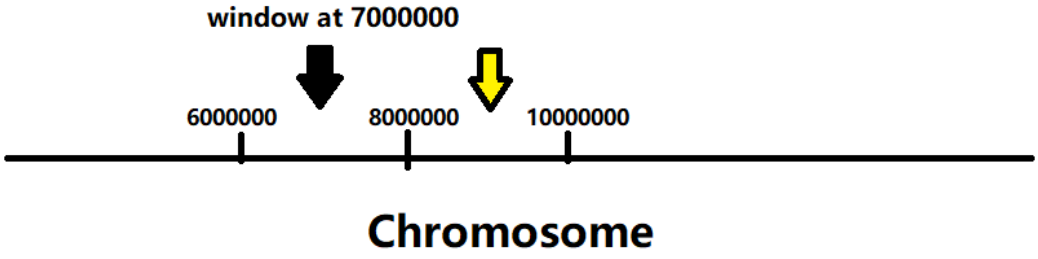

Figure 3. How the window moves (credit: original).

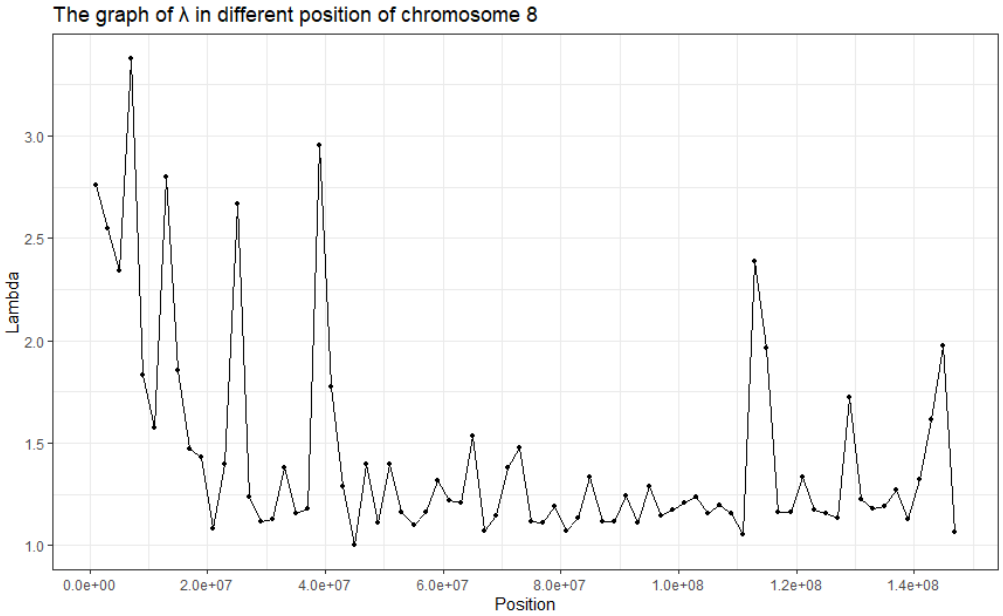

Figure 4. Examples of graphs (credit: original).

According to the observation, the following methods are adapted. Add the maximum value of rate of Poisson Distribution and the minimum value of rate of Poisson Distribution. Then divide the sum by 3 to calculate the Minimum Rate of Hotspots (MRH). If over two-thirds of points do not have rates that are higher than MRH, then any points with rates higher than MRH are considered as hotspots; if over one-thirds of points have rates that are higher than MRH, I add the MRH with 0.35 to calculate the Fixed Minimum Rate of Hotspots (FMRH). No matter how many points have rates higher than FMRH, the points that have rates higher than FMRH are all considered as hotspots. That means it is acceptable even

DOI: 10.54254/2753-8818/4/20220614

if only one point can be considered as hotspots among 40 or more points in one chromosome. All the hotspots are recorded as coordinate of the window, then the record is transformed to the intervals, which makes the analysis of the hotspots easier.

Analyze the association between the hotspots and prognosis. After knowing the hotspots window of the CNV, the author match the code of the patients in the data of CNV start and end points with other data to find their days of death. Now consider all the patients whose variation points (start or end points) of analyzed chromosome are in one hotspot as experimental group, the patients whose variation points of analyzed chromosome aren't in that hotspot as control group. All hotspots are analyzed by proportional hazards model (Cox regression model). The hotspot that has a p-value smaller than 0.05 is considered to be significant. That means the CNV in that point has ahad a significant correlation with the overall survival of patients . Considering that some data about day of death in TCGA is not applicable and the way to include last day of contact into the analysis is complex, the author only analyze the samples of patients whose days of death are clear. This adjustment of analysis won't affect the accuracy of the result because the number of patients analyzed is much bigger than 100, from 300 to 400. That is a big sample.

2.2 Result

After analyzing 180 hotspots determined by using the methods of 2.1, 14 hotspots that may be significant are found in 9 chromosomes. They are chromosome 5, 6, 8, 10, 13, 15, 17, 18, 20. The number of hotspots and position of hotspots in these chromosomes are shown in the following table (table 1). The λ and p-value of each window is also shown in the following table (table 2). There is an example graph of chromosome 8 in window 7000000, namely the interval (6000000.8000000). (Figure 5).

Table 1. The hotspots that have a significant correlation with the overall survival of patients (credit: original).

Chromosome	windows of hotspots	number of hotspots
5	103000000, 151000000	2
6	163000000	1
8	7000000, 25000000, 115000000, 145000000	4
10	125000000	1
13	49000000, 57000000	2
15	93000000	1
17	35000000	1
18	21000000	1
20	53000000	1

DOI: 10.54254/2753-8818/4/20220614

Table 2. The λ value of the hotspots which have a significant correlation with the overall survival of patients (credit: original)

chromosome	windows of hotspots	p-value	λ value (approximate)
5	103000000	0.04	1.75
5	151000000	0.034	1.7
6	163000000	0.015	1.99
8	7000000	0.04	3.35
8	25000000	0.0092	2.65
8	115000000	0.01	1.98
8	145000000	0.017	1.99
10	125000000	0.039	1.6
13	49000000	0.044	1.5
13	57000000	0.031	1.68
15	93000000	0.0091	1.4
17	35000000	0.022	1.87
18	21000000	0.025	1.36
20	53000000	0.049	1.6

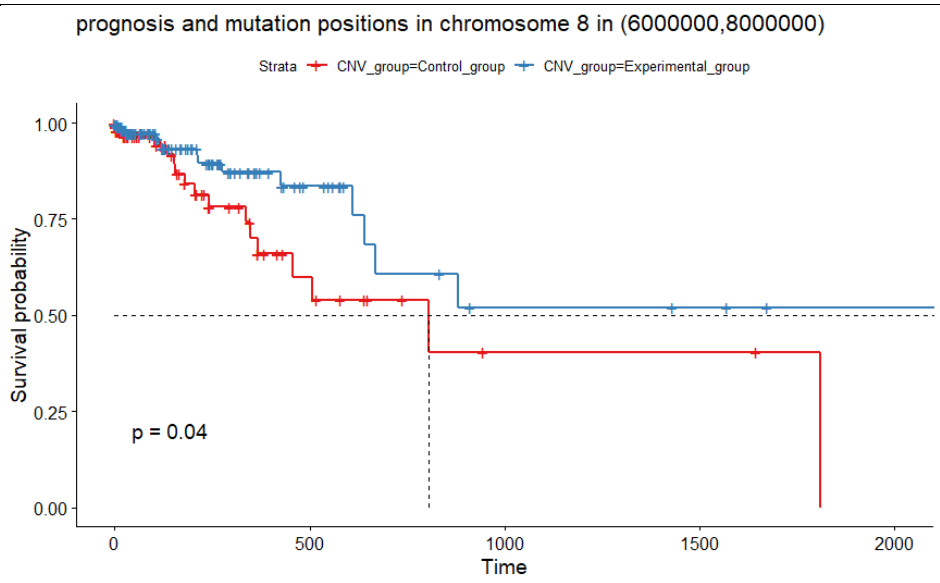

Figure 5. Example graph of chromosome 8 in window 7000000 (credit: original).

2.3 Analysis

Interestingly, the hotspots often appear when the λ of the windows are smaller than 2.65. Maybe that is because if the λ is bigger than 2.65, the control group won't have enough samples to draw the survival curve, or the number of samples in control group is far from the number of experimental group, I can't identify effect of the CNV happen in this window. The hotspots also often appear near the end of the chromosome. The window 163000000 in chromosome 6, the window 145000000 in chromosome 8, the window 125000000 in chromosome 10, the window 93000000 in chromosome 15 are close to the end of the chromosome. The chromosome 8 has the biggest number of hotspots windows in 9 chromosomes. Maybe there are more special genes that can control the growth or the death of cell in chromosome 8 than other chromosomes analyzed in this research. In some of the survival curve graphs

DOI: 10.54254/2753-8818/4/20220614

the end of one or two curves decrease dramatically or disappear suddenly instead of remaining horizontal or decreasing slightly. This is caused by the lost of data from previous step (deleting the samples of patients whose days of deaths are unknown). The lost of data may cause some of the inaccuracy in this analysis, but won't have significant impact on the result. In the initial stage in some of the graphs, the curve of the experimental group has many intersect points control group, indicating that the effect of this window may be less significant, but due to a low enough p value, the significance is still acceptable.

2.4 Discussion

The discussion of the method of MRH. The method of determining hotspots applied in this research depends on the observation of the graphs. If I rank the rates of points to determine the hotspots, I can't figure out because there is no fixed proportion of hotspots in chromosomes. Ranking and sorting can't be conducted in a fixed way. In addition, in some chromosomes, there are only 2 or 3 points that have relatively high rates, while others are almost the same. The method of ranking will consider points that have relatively small rates of the Poisson Distribution as hotspots. During the research, I have considered to write an algorithm to examine whether the point is a peak. Only the peaks that have relatively high values can be recognized as hotspots. But this method was rejected by me soon because some of the graphs have an area which include 3 to 4 points with relatively high rates. These points should be considered as hotspots, but in this method they will be ignored because they are not peaks. This will decrease the accuracy of the research. The method of taking the top 3 windows was also rejected because the chromosomes are in different length. It is irrational to consider only 3 points as hotspots while there are 60 points or more in the graphs.

The discussion of the method of FMRH. If FMRH isn't calculated when the number of hotspots reach the condition, half of the analysis in chromosome can't be produced. That is because in some chromosomes, only one point can reach a rate which is relative very high, while other rates of peeks in the graph are only half of the maximum. In some chromosomes only one point can reach a rate which is relatively low, influencing the analysis of the hotspots. Sometimes the division value is very small. It may be apparent that the maximum value and the minimum value do not have large difference. This may lead to a MRH that is even smaller than the minimum value of rate in the graph. These factors disrupt the calculation of MRH. The calculation of FMRH will assure that the number of hotspots in one chromosome can be limited approximately from 5% to 15% of the total number of windows. For example, if a chromosome has 40 windows, the number of hotspots of the chromosome may be from 2 to 6. The method used in this research is worth developing. What is important is that this method, including the calculation of MRH and FMRH, is based on observation, not on a formula that was proposed by essays. The further studies should focus on the accuracy of the calculation of MRH and FMRH. The further studies should also focus on the principles of determining hotspots, trying to improve the methods of determining hotspots with graphs of windows and the λ.

The discussion about methods of setting windows and investigating start and end points. Currently, there are methods of calculating the patterns of CNV by constructing a 48-dimensional matrix. The characteristics of patterns include ploidy, fLOH, cLOH and so on. The patterns will be associated with the prognosis of patients. But these methods are complex, and they take many variables into account such as the number of copies[4]. This research only focuses on the start and end points of CNV, investigating the association between position of CNV and the prognosis of patients. That tremendously simplify the research. However, I can't figure out what impact other variables may have on the prognosis. Due to the limit of time, I only analyze one variable: positions of CNV.

The discussion about CNV in general population. Currently, there are difficulty distinguishing CNV that will make people ill from those that seldom occur in the general population but are generally benign. The CNV in normal tissue is so rare that no equipment can detect those mutations and gain accurate data[10], so I can't find whether the cells of normal issue have the CNV in the hotspots determined by this research. As a result, it is hard to identify whether the effect of the CNV is increasing the survival days of patients or not. This research shows the positions in chromosomes that should be

analyzed more thoroughly, giving a way to developing potential tumor therapeutic targets. But this research can only state that there are association between the prognosis of Stomach Adenocarcinoma and the hotspots of the CNV. Because of limited time, I can't analyze the gene in the hotspots. I only research on the start and end points of the CNV. I do not take the numbers of the copies and the positions these copies finally reach into account. Maybe normal cells also have these CNV and the mutations have no effect on these cells. Maybe the number of the copies and the positions these copies finally reach have impact on the survival rates of the patients. But what can't be denied is that these hotspots are worth studying. Further research should focus on how to detect the CNV in normal cells. The points of CNV in normal cells should be recorded and compared with the data of patients. Specific genes associated with CNV and their structures should be taken into account to know the real effect of the CNV in the hotspots. The genes at the destination of copies should also be learned to know whether the copies interrupt or stimulate the functions of the genes. The next step of my research is analyzing more variables once a time, then concluding all the results into patterns. I will also study the structures of the genes and the distribution of genes in chromosomes.

3. Conclusion

There are 180 hotspots of CNV in 24 chromosomes of patients with Stomach Adenocarcinoma. There are 14 hotspots where the CNV have a significant impact on the prognosis of the patients. The 14 hotspots are windows 103000000, 151000000 in chromosome 5, window 163000000 in chromosome 6, windows 7000000, 25000000, 115000000, 145000000 in chromosome 8, window 125000000 in chromosome 10, windows 49000000. 57000000 in chromosome 13, window 93000000 in chromosome 15, window 35000000 in chromosome 17, window 21000000 in chromosome 18, window 53000000 in chromosome 20. The specific effects of the CNV happen in these windows are unknown.

Acknowledgment

First of all, I would like to give my heartfelt thanks to Mr. Qiyuan Li who provided me with idea of investigating Stomach Adenocarcinoma and hotspots of CNV. He gave me practical instructions and imparted basic but important knowledge to me. In addition, I wanted to express my gratitude towards my parents, especially my father. As a cancer doctor, he gave me suggestions on the determination of hotspots. Although the suggestions proved to be impractical, they were inspiring, stimulating me to form a method that would be practical in this research.

References

[1] Nyren, O. and H.-O. Adami, Stomach cancer. Textbook of cancer epidemiology, 2008. 2: p. 239-74.
[2] Torpy, J.M., C. Lynm, and R.M. Glass, Stomach Cancer. JAMA, 2010. 303(17): p. 1771-1771.
[3] Zhang, X., et al., The Somatic Mutation Landscape and RNA Prognostic Markers in Stomach Adenocarcinoma. Onco Targets Ther, 2020. 13: p. 7735-7746.
[4] Steele, C.D., et al., Signatures of copy number alterations in human cancer. Nature, 2022. 606(7916): p. 984-991.
[5] Sudmant, P.H., et al., Diversity of Human Copy Number Variation and Multicopy Genes. Science, 2010. 330(6004): p. 641-646.
[6] Conrad, D.F., et al., Origins and functional impact of copy number variation in the human genome. Nature, 2010. 464(7289): p. 704-12.
[7] McCarroll, S.A. and D.M. Altshuler, Copy-number variation and association studies of human disease. Nat Genet, 2007. 39(7 Suppl): p. S37-42.
[8] Rogozin, I.B. and Y.I. Pavlov, Theoretical analysis of mutation hotspots and their DNA sequence context specificity. Mutation Research/Reviews in Mutation Research, 2003. 544(1): p. 65-85.
[9] Almal, S.H. and H. Padh, Implications of gene copy-number variation in health and diseases. J Hum Genet, 2012. 57(1): p. 6-13.
[10] Girirajan, S., C.D. Campbell, and E.E. Eichler, Human copy number variation and complex genetic disease. Annu Rev Genet, 2011. 45: p. 203-26.