Peer Review File

Article information: http://dx.doi.org/10.21037/tcr-20-3221

Reviewer A

General comments

Comment 1: Overall, the study is a nice overview of the power of using natural language processing to cluster keywords together to infer relationships. The Authors make some strong conclusions that are lacking robust evidence as keyword relationships do not provide context. This is a limitation I would like to see mentioned in the study. There are some things that are not clear, such as why PubMed was not also included, was there any exclusion criteria, what are the categories of the citations (articles, reviews, letters to the editor, etc), and what MALAT1 does specifically.

Reply 1: Thanks for your comments. In this research, first, the cluster analysis was implemented via research about "MALAT1" and "cancer" in *Web of Science. Web of Science* is a typical citation database, in addition to the abstract of the document, it also contains citation information, which can be used for the analysis of citations between articles and the relationship between citations. Therefore, in the early stage of this research, we only used the relevant information in *Web of Science* for cluster analysis. Secondly, in order to further explore the correlation between "MALAT1" and "breast cancer", relevant data published in *GEO* were collected for meta-analysis. In short, two kinds of databases were utilized to conduct analysis in different directions.

Comment 2: I think a strong revision can help this study, starting with stating the hypothesis and focusing the text on that. It would really improve the article if the hypothesis were stated and the results were provided in the context of the hypothesis.

Reply 2: Thank you for your suggestions. We have revised the article according to the questions you raised, and given the specific revision process.

Major Comments

Comment 3: Tense is an issue here. The article keeps switching from past to present and future tense. The whole article needs to pick one (in accordance with the journal) and revise. I also suggest copy editing be done, as there are a few areas that are weak and several phrases are repeated ("adopted", "therefore").

Reply 3: Sorry for this mistake and thank this good suggestion. We reviewed this paper and found the according sentences with confusing tense. Thus, we modified the tense all over the paper. The quoted examples were cited in past tense and the descriptive statements were cited in simple present tense, which all could be seen in red marker. The tense of this article has been revised. In addition, The repeated phrases were adjusted and weak areas were improved accordingly (such as Page 3, line 51).

Comment 4: Section 1. Introduction: a nice overview of the topics, but I would like to see some specific examples of why the reader should care about MALAT1. The examples provided are quite generic, so specific findings would help convince the reader that MALAT1 is important and worth studying.

Reply 4: Thanks for your valuable advice. Based on your comments, we have added some specific examples on the correlation between MALAT1 and cancer in the introduction of the article, for highlighting the importance of MALAT1 gene in cancer diseases.

The specific examples are as follows and as in revised manuscript:

" Metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) is a highly conserved lncRNA, which is expressed in mammalian multi-tissue cells and can participate in epigenetic regulation and cell cycle regulation (6-8). In addition, many studies confirmed that MALAT1 was closely related to malignant tumors such as lung cancer, liver cancer, gastric cancer, and colorectal cancer. The high expression of MALAT1 was mostly found in tumor cells, which can promote the proliferation, metastasis, and invasion of tumor cells (9,10). Research of Zhang et al. proved that the expression of MALAT1 was correlated with the overall survival of patients after surgery (11). Studies also revealed that MALAT1 can regulate angiogenesis through the extracellular regulated protein kinase / matrix metalloprotease pathway, thereby promoting cancer metastasis (12)." (see Page 4, lines 77-87)

Comment 5: Section 2.1 in the Methods. I am not sure this really belongs here, as it is not outlining what was done by the authors. It is just a general introduction to the software, and if it is software they are promoting, then it should be moved to the introduction. If it is not their software that they are promoting, then this section should be cut down to the parameters that they used, and a reference made to the original paper for CiteSpace. Overall, it feels a bit strange as CiteSpace itself does not appear to be the focus of the study, yet an in-depth description was provided.

Reply 5: Thank you for your kind comment. The analysis principle and process of CiteSpace software were briefly described in section 2.1. Indeed as you said, it was not the software we promoted, but only a general introduction to the software. So we should have moved out in the introduction. In this revised manuscript, the introduction of the software was deleted and only the analysis process of the software was retained (see Page 5, line 112).

Comment 6: It was also hard to follow what was being described in section 2.1. I think if this is not their software, it should be cut down and a reference to a paper where it describes these details can be provided.

Reply 6: Thanks for the comment. Section 2.1 mainly described the principle and specific process of CiteSpace for document clustering. In this revision, it has reduced the content in Section 2.1 based on your comments. Only the specific flow chart of using CiteSpace for literature information analysis was retained (see Page 5, line 99).

Comment 7: Line 81-83: While a reference to the Web of science is provided, it is not clear if the authors had to download or acquire additional data, or if this data is provided in CiteSpace. If the former, please provide links or reference numbers to exact place to get the data. If the latter, please state that all data can be found in CiteSpace. If it is a search engine, consider rewording lines 81 to 90 to make that clearer. Additionally,

there are multiple databases called “Web of Science” including from institutions, so a direct link is crucial for someone to repeat this study.
Reply 7: Thank you for the comment. *Web of Science* is a kind of database. We have provided the link (http://webofscience.com) in the revised manuscript (see Page 5, line 109). Most of the literature used for the construction and analysis of the knowledge graph in this article came from *Web of Science*. After retrieving the required information by searching the database, the relevant information was saved in a specific format, which were imported into CiteSpace to construct a knowledge graph. The documents that could be clustered in this study have been specifically cited in the article. The content of the subsequent knowledge map results came from different types of analysis of the documents downloaded in *Web of Science*. The process was described in detail in section 2.2 in the methods.

Comment 8: Section 2.3. This section nicely walks the reader through how to use CiteSpace, but CiteSpace is not the focus of this study (if it is, then it is not clear). I think this section could be shortened a bit to just explain which parameters were used. Also, there is a few areas where the writing is not scientific.
Reply 8: Thank you for your comments. It has appropriately reduced the content in section 2.3, and only listed the main parameter settings (see Page 6, lines 122-132).

Comment 9: Section 2.4: There is insufficient explanation as to what was done here. What software was used to process the arrays? How were the data cutoffs determined? What were the exact cutoff values used? How were the proportional hazard models developed? Which ones? What were the cutoff values for significance? There are a few things mentioned, but it is not clear what was used for what.
Reply 9: Thank you for your suggestions for amendments. Detailed information was supplemented in section 2.4 (see Pages 6-7, lines 134-150). This study adopted the data in Affymetrix human genome U133 plus 2.0 array and U133A array. Then, a total of six data sets were finally selected, namely GSE6532, GSE9195, GSE20711, GSE31448, GSE42568, and GSE1456. The value of “third quantile as cut-off value” was specifically classified according to the expression results obtained from the detection of different chip data, so a definite cut-off value cannot be given. The proportional hazard regression model is a semiparametric regression model proposed by British statistician D.R.Cox (1972). The analysis method “R language” of the proportional risk model was provided in this revision.

Comment 10: Section 3.1. The authors comment on MALAT1 and cancer by country and make the statement “It means that Chinese are more interested in studying the mechanism of MALAT1 gene in cancer, which has provided a good research atmosphere and laid a solid foundation for exploring the function of MALAT1 gene in China.” That is a strong conclusion to make based on the limited data provided in this study. There is not enough evidence provided here to come to that conclusion.

Reply 10: Thank you for the kind suggestion. In the previous study, we intended to show that the Chinese region had a relatively large share of such articles published. After consideration, we deleted this scentence.

Comment 11: Additionally, how are international collaborations categorized? Is the country of publication based on the journal, the lead author's institution, or the last author's institution?
Reply 11: Thank you for your comments. After carefully checking, we were sure that the collaborative articles were classified based on the information of the first author and have added in the revised manuscript (see Page 7, lines 164-165).

Comment 12: The same can be applied to the conclusions on line 164-168
Reply 12: Thank you for your comments. After carefully checking, we were sure that the collaborative articles were classified based on the information of the first author and have added in the revised manuscript (see Page 8, lines 181-182).

Comment 13: Section 3.2. The authors make the conclusion (line 176) that the expression of MALAT1 is closely related to proliferation of cancer cells or metastasis. There is no evidence provided at this point to come to that conclusion as there are no expression data sets accompanying this observation. The only thing that Figure 8 shows is that there is a relationship between those terms, but there is no context provided, especially not expression. Please provide a gene expression analysis with this observation or remove the conclusion.
Reply 13: Thanks for the comment. From section 3.2, CiteSpace was employed to construct the knowledge graph of the keywords published in 1996-2020, it can be seen that MALAT1, expression, metastasis, proliferation, cancer, etc. were frequently occurring keywords. We pointed out that CiteSpace can point out the closeness of the topic vocabulary relationship through "keyword co-occurrence analysis". But as you said, we did not provide enough convincing evidence and we have deleted this scentence in the revised manuscript (see Page 8, line 190).

Comment 14: In the following paragraph (lines 179-187), the conclusions are not as strong, but the evidence is not convincing either. This may be because the importance of centrality is not explained or what it means. Also, Table 1 is poorly explained. Most of the centrality scores are 0, so what does that mean? MALAT1 is 0.57, why is it so far? Is there a cutoff value for when something is not considered related?
Reply 14: Thanks for the comment. Centrality is a unique term in CiteSpace, which is an indicator of the importance of detecting nodes in the network. The larger the Centrality, the more important the node was in the network. As shown in Page 8, lines 192-193: "Table 1 suggested the higher the centrality, the more studies around the keyword." Table 1 was the analysis result of using CiteSpace to analyze the subject terms "MALAT1" and "Cancer" related articles. The range of Centrality in the software was [0,1]. Most of the centrality scores were 0, which meant that nodes were relatively unimportant in the network . It was probably because the number of network

structure nodes in the component map was too large (over 500), which caused the software to stop the calculation of Centrality. In this modification, the parameters were adjusted and the complete Centrality value was given (see Table 1).

Comment 15: Line 200-206: Tags are given but no context is provided. The author comments on the relationship between the keywords, but both "regulates" and "suppressor" are there. A statement about the limitation of this kind of study by not being able to understand the context of these keywords should be provided. IE is MALAT1 a regulator of tumorigenesis and migration, or a suppressor?

Reply 15: Thanks for your comments. We previously stated that the use of CiteSpace to construct a knowledge map was only to analyze the current research front-burner issue of MALAT1 in cancer diseases. It was hoped to analyze the potential mechanism of MALAT1 regulating cancer through information extraction from a large amount of literature. Therefore, the results in lines 211-219 were only to show other keywords that MALAT1 affected the cancer process in the current published literature, in order to speculate on the current research front-burner issue and provide research directions for subsequent research.

Since this study was based on big data, analysis software was adopted to explore potential associations, and it did not conduct substantial experimental research, we could not keep check and inspection of each message. Clustering tags might not be accurate due to the particularity of published articles and other reasons. Therefore we could only get a trend or assumption, guiding role for subsequent experiments, but still need to experiment, which are limitations (see Page 9, lines 217-219: "Since this study was based on big data, analysis software was adopted to obtain assumptions, experiments were still needed for verifying"). As shown in Figure 8 (original Figure 9), "regulates" was more convincing than "suppressor". MALAT1 was previously described by many papers as a cancer-promoting and metastasis-promoting lncRNA, while other reports suggested a tumor-suppressing role of MALAT1. Therefore, based on the current studies, it is more accurate that MALAT1 is a regulator of tumorigenesis and migration (refer to : doi: 10.3390/cancers11020216).

Comment 16: Line 207 to 224 are a series of fragmented paragraphs and do not really provide any explanation of the results. It is unclear how they relate to the hypothesis of the study. Please revise or remove.

Reply 16: Thank you for your kind comment. According to your comments, the content of original lines 207-224 has been adjusted appropriately. Some content has been deleted (see Pages 9-10, lines 217-228).

Comment 17: Throughout section 3.2 the authors continuously refer to papers that are highly cited, but at the same time talk about how more publications are coming out in recent years. As newer papers would have fewer citations due to the length of time they have been available, the authors need to have a way of considering time in their citation frequency. Otherwise, this produces bias towards articles published years back. This

creates a disjunction between the observations focusing on more publications coming out in recent years as well as focusing on the citation index of older articles.

Reply 17: Thanks for the your comments. In the early stage of this research, we determined the time span from 1996 to 2020 and used the relevant information in *Web of Science* for cluster analysis about "MALAT1" and "cancer" Then, in order to further explore the correlation between "MALAT1" and "breast cancer", relevant data published in *GEO* were collected for meta-analysis. We mainly focused on the association between MALAT1 and breast cancer during the period from 1996 to 2020, and the selected materials could basically cover all the MALAT1 related articles published within this time range. Therefore, Our study was based on big data to obtain assumptions, which still needed experimental verification in the future.

Comment 18: Section 3.3: The authors state that the "results in Figure 12 show that high expression of MALAT1 is…" but there are no expression values in Figure 12. Is it high expression or low expression that is correlated with relapse-free survival? Figured 12 just shows that expression level is correlated, it does not show if it is low or high expression levels. Please provide the analysis showing the correlation with expression values.

Reply 18: Thank you for the kind comment. It was pointed out in section 2.4 that the "third quantile value" was used to classify the expression values "high", "medium", and "low". The expression results obtained by different chip analysis were different, so they cannot be unified into the same data range. The proportional hazard model was utilized to calculate the 95% CI value of expression data and survival rate and recurrence-free rate, and then Review Manager 5.3 was employed for meta-analysis. From the forest plot on the right side of Figure 10 (original Figure 12), it was obvious that both the survival correlation and the recurrence-free correlation were biased towards the "High" expression of MALAT1, so that the results in section 3.3 were acquired.

Comment 19: Section 4 Conclusion: On lines 261-263 the authors state "The results reveal that the abnormal expression of lncRNA MALAT1 is closely related to the occurrence and progression of cancer and the poor prognosis of patients." There was no comparison to normal expression levels of MALAT1 in healthy patients so there is not enough evidence to state that it is abnormal expression. Furthermore, there is not enough evidence relating to occurrence and progression of cancer as there was not a time series provided or healthy samples. If the authors wish to make these conclusions, they should supply healthy samples and/or a before and after incident dataset. Otherwise revise the conclusions.

Reply 19: Thank you for your kind suggestions. According to your comments, the conclusion of "The results reveal that the abnormal expression of lncRNA MALAT1 is closely related to the occurrence and progression of cancer and the poor prognosis of patients." in Section 4 has been revised (see Page 12, lines 271-273). The results of the previous knowledge map of MALAT1 and cancer showed that most keywords were clustered into keywords such as "expression". It was speculated that the expression of

MALAT1 was closely related to the progression of cancer diseases. Then, the meta-analysis results showed that the high expression of MALAT1 was closely related to the survival rate and recurrence rate of breast cancer patients. Therefore, it was speculated that the expression of MALAT1 was closely related to the prognosis of cancer patients. The limitation of the research proposed was that it had not been verified by experiments, and it proposed that the follow-up needed to be verified by experiments.

Comment 20: The next conclusion is "Moreover, lncRNA MALAT1 may play its role in regulating breast cancer progression through the PI3K/AKT/m-TOR signaling pathway." The evidence provided does not support that so strongly, as there are just keywords provided without proper context. This conclusion should be drawn in to focus on this limitation, as the evidence provided does not have a measurable way of demonstrating the relationship between MALAT1 and PI3K/Akt/m-TOR.
Reply 20: Thanks for the kind advice. It has deleted the conclusion that "lncRNA MALAT1 may play its role in regulating breast cancer progression through the PI3K/AKT/m-TOR signaling pathway." (see Page 12, line 273)

Figures and Tables
Comment 21: All the figure legends need to be improved and provide more description (especially Figure 12). The authors should provide enough description of each figure to walk the reader through so that they understand the image. At the moment, there are a lot of components of each figure that is not explained. Please revise.
Reply 21: Thank you for your valuable comments. To address this problem, we tried to define each figure legend more clearly. Specific explanations are as follows (Based on the previous revisions, we have deleted Figure 1, Figure 10 and Table 4, so the order of the charts has partially changed) :

Figure legends
Figure 1. CiteSpace analysis process. The analysis process using CiteSpace was mainly divided into the following process. I, the time node and time slicing mode were decided; II, the analysis node type and the link type were selected; III, the similarity or proximity of the input data was calculated; IV, corresponding networks for each time slice were constructed; V, whether to perform network scaling was determined. If so, network scaling was performed; if not, the time series of the network were merged; VI, the merge of network time series after network scaling; VII, finally, the analysis process could be ended according to whether to perform network scaling and the network after scaling with merged time series.

Figure 2. CiteSpace parameter setting interface.
After the CiteSpace software was running, *Web of Science* was selected and a new program in the "Projects" column was created. The "project" folder created in the previous step was taken as the "Project Home" and the "data" folder as the "Data Directory". I. Time Scaling: time scaling was for setting and presenting the temporal evolution of the research data, and the time span could be selected based on the time range of the data. II. Text Processing: this module was divided into two parts: "Term Source" and "Term Type". III. Network Configuration: this module was divided into

three parts: “Node Types”, “Links”, and “Selection Criteria”. IV. Pruning: the “Pathfinder and Pruning the merged network” option was chosen. V. Visualization: the “Cluster View-Static” and “Show Merged Network” options were chosen.

Figure 3. Statistics on the number of publications on MALAT1 gene and Cancer in different years. The number of publications with the keywords “MALAT1” and “Cancer” from 1996 to 2020 was derived from the core database of *Web of Science*. Horizontal and vertical axes represented the year and number of publications respectively.

Figure 4. Statistics on the number of publications on MALAT1 genes and Cancer in different countries. The number of publications with the keywords “MALAT1” and “Cancer” from 1996 to 2020 was derived from the core database of *Web of Science*. Horizontal and vertical axes represented the country and number of publications respectively.

Figure 5. Statistics on the number of publications on MALAT1 genes and Breast Cancer in different years. The number of publications with the keywords of “MALAT1” and “Breast Cancer” from 1996 to 2020 was derived from the core database of *Web of Science*. Horizontal and vertical axes represented the year and number of publications respectively.

Figure 6. Statistics on the number of publications on MALAT1 genes and Breast Cancer in different countries. The number of publications with the keywords of “MALAT1” and “Breast Cancer” from 1996 to 2020 was derived from the core database of *Web of Science*. Horizontal and vertical axes represented the country and number of publications respectively.

Figure 7. Clusters of core keywords of MALAT1 gene and Cancer. In CiteSpace, it set the “Top N%” and “per slice” as 1% and 100, respectively, to visualize the knowledge map of keywords.

Figure 8. LLR-based clusters of core highly cited publications on MALAT1 gene and Cancer. The highly cited literature clustering was performed by the LLR method, and a total of nine clustering tags (sunitinib, regulates, tumorigenesis, characterization, migration, metastasis-associated, transition, suppressor, and mir-9) were obtained.

Figure 9. LLR-based clusters of core highly cited publications on MALAT1 gene and Breast Cancer. Through the clustering of highly cited publications by the LLR method, 5 clustering tags (breast, long, PI3K/AKT/mTOR, coding, and -coding) were obtained.

Figure 10. Meta-analysis of the correlation between MALAT1 expression and breast cancer survival. A was the correlation between MALAT1 expression and recurrence-free survival; B was the correlation between MALAT1 expression and overall survival rate. The 95% CI value was the correlation value between MALAT1 expression obtained by proportional hazard model analysis and patient survival rate and recurrence-free rate.”

Comment 22: The figure quality, for example Figures 9 and 10, need to be improved (resolution).
Reply 22: Thank you for your kind suggestions. We have attempted to improve the quality of figures in our manuscript. For example, the resolution of Figure 8 (original 9) has been improved to 300 dpi (see Figure 8) and the original Figure 10 has been deleted for the previous revisions.

Comment 23: What are the years referring to in Table 1? Please include a description
Reply 23: Thanks for your comments. The year in Table 1 referred to the year when the keyword appeared frequently, and the notes have been given at the bottom of Table 1 (see Table 1).
"Year: the year when the keyword appears frequently."

Questions
Comment 24: Section 2.2, is there a reason why PubMed was not included in this search? Could this explain the results found on lines 147-153
Reply 24: Thanks for your valuable comments. Since the focus of this article was on the construction of a knowledge map of published documents and their citation frequency, and the potential associations between different documents derived from *PubMed* cannot be analyzed, only the data derived from the *Web of Science* were used to construct the knowledge graph. The latter part of the article performed meta-analysis to explore the relationship between MALAT1 expression and the prognosis of breast cancer patients. This part of the data came from the analysis of the chip data included in the *GEO* database of NCBI.

Comment 25: What is the vetting process for Web of Science?
Reply 25: Thanks for the comment. For the selection criteria of *Web of Science*, it was held at the Clarivate Analysis headquarters in Philadelphia on the morning of May 30, 2019. Ms. Mariana Boletta (Acting Executive Editor, *Web of Science* Group) and Ms. Chang Liu (Editor, Editorial Selection, Scientific and Academic Research) reported the *Web of Science* Core Collection journal selection process and standards.

First, the subject terms "MALAT1" and "Cancer", and "MALAT1" and "Breast Cancer" were searched in the core database of *Web of Science*. Then, "Article" was selected as the document type, and time span from 1996 to 2020 was determined. Finally, a total of 1,180 articles with the subject word "MALAT1" and "Cancer", and 215 with "MALAT1" and "Breast Cancer" were retrieved. Each record was saved as a plain text document, the record format was as follows "author, title, source publication, and abstract", and the file name was "Download.txt". Then, the data-import/export function in CiteSpace was taken to transform the input.

Comment 26: Section 2.2 Why was 1996 to 2020 chosen? Why not earlier?
Reply 26: Thanks for the comment. From Figure 3 (original Figure 4) and Figure 5 (original Figure 6), the research on MALAT1 gene in cancer only had a large number

of reports after 2008. Therefore, the research scope selected in this article from 1996 to 2020 can fully include the related research of MALAT1 gene.

Minor comments

Comment 27: Line 12 to 14 is a bit hard to follow. Perhaps consider shortening it to avoid repeating the same phrases multiple times in a sentence.

Reply 27: Thanks for the kind advice. According to this comment, the sentence has been shortened as follows.

"The relevant researches on MALAT1 in cancers published in recent years are collected and integrated. CiteSpace is employed to draw a knowledge map of MALAT1 in breast cancer, to evaluate the research front-burner issues (see Page 3, lines 50-53)."

Comment 28: Line 19: I believe the genes are acronyms and should be capitalized (MTOR = Mechanistic/Mammalian Target of Rapamycin, for example). Also does the author mean Akt, not Art?

Reply 28: Thank you for your careful work. "pi3k/art/mtor" in line 58 was revised to "PI3K/AKT/mTOR" (see Page 3, line 58).

Comment 29: Line 21, I am not sure what "It" is referring to.

Reply 29: Thanks for the comment. The sentence was revised as "It is hoped to …" (see Page 3, line 60).

Comment 30: Line 30: I am not familiar with "gene chip", does the author mean DNA Microarray?

Reply 30: Thanks for the comment. The "DNA Microarray" you proposed is what we often call "Gene chip", which can also be called "biochip". Therefore, in the revised version, we still use "gene chip".

Comment 31: Line 32, I am not sure "aroused" is part of scientific writing in this context.

Reply 31: Thanks for the comment. This word has been replaced with "drawn" (see Page 4, line 72).

Comment 32: Line 33: first tile lncRNA is mentioned in the main article but the definition of the abbreviation is not provided.

Reply 32: Thanks for the comment. The full name of "lncRNA" "ong non-coding ribonucleic acid (lncRNA)" has been added (see Page 4, line 73).

Comment 33: Line 34: References missing.

Reply 33: Thanks for the kind comment. The references related to lncRNA have been added (see Page 4, line 74).

Comment 34: Line 37: References missing. Also first time MALAT1 is mentioned in the main article but the definition of the abbreviation is not provided.

Reply 34: Thank you for your comments. The full name of "MALAT1" "Metastasis-associated lung adenocarcinoma transcript 1 (MALAT1)" was added in line 77. The role of MALAT1 gene in cancer has been modified accordingly (see Page 4, lines 77-87).

Comment 35: Recommend changing to "This study provides a new…" makes the sentence less sticky
Reply 35: Thanks for the suggestion. According to the modification you proposed, the expression "This study intends to provide a new..." in line 108 has been changed to "This study provides a new way..." (see Page 4, lines 94-95).

Comment 36: Line 53: "amazing" is not scientific writing.
Reply 36: Thanks for the comment. This word has been revised (see Page 5, line 100).

Comment 37: Line 57: "hot issues" is not scientific writing
Reply 37: Thanks for the comment. It was revised to "front-burner issues" (see Page 5, line 104).

Comment 38: Line 70: "and so on" is awkward
Reply 38: Thanks for the comment. It was deleted and adjusted (see Page 5, line 105).

Comment 39: Line 91: It is not necessary for the reader to know what the file name was. Same can go for the record format.
Reply 39: Thanks for the comment. Because of the particularity of the CiteSpace software, it has special requirements for the naming of the sample files of the input software, so we have given the specific name of the input software in the article.

Comment 40: Line 93-96: This information is unnecessary for the reader to be able to perform these experiments. (line 92-93 is fine as it describes what function was used).
Reply 40: Thanks for your kind suggestions. According to your comments, in this modification, unnecessary content was deleted, and the steps for data conversion using CiteSpace software were retained, which was merged with the previous paragraph (see Page 5, lines 119-120).

Comment 41: Line 99-101: This is a bit confusing, it would be easier to describe what output from the previous step was used for what parameter in this step. Stating "folder" is a bit abstract and is not explaining what is needed.
Reply 41: Thanks for your valuable suggestions. According to your comments, the content of original lines 99-101 has been adjusted and deleted accordingly (see Page 6, lines 122-123).

Comment 42: Line 106-109: While this area nicely describes the process that was done, "is clicked" as well as the rest of that sentence needs a rework to be more scientific.

Reply 42: Thanks for the comment. The content after the "is clicked" in original lines 106-109 was adjusted (see Page 6, lines 125-127).

Comment 43: Line 119 is more like instructions than explaining what was done. Please revise
Reply 43: Thanks for the comment. The description in lines 131-132 was the final operation flow after all the parameter settings in section 2.3 were completed (see Page 6, lines 131-132).

Comment 44: Line 125: I am not sure what "study-specific third-quantile distribution" means or is referring to. Please explain the context
Reply 44: Thanks for the comment. "Third-quantile distribution" meant that the MALAT1 expression in the chip data was arranged from large to small, it was classified into low (values from 0% to 33.3%) and medium (33.3 % To 66.6%) and high (66.6% to 100%) (see Page 6, lines 138-141).

Comment 45: Line 160: 10.698% of what? What is the total referring to? Please revise
Reply 45: Thanks for the comment. Lines 169-179 were the analysis of the annual distribution characteristics of the number of articles published on the subject headings "MALAT1" and "Breast Cancer" from 1996 to 2020 from the core database of *Web of Science*. Therefore, articles about "MALAT1" and "Breast Cancer" published in 2020 accounted for 10.698% of the total number of articles with the keywords of "MALAT1" and "Breast Cancer" from 1996 to 2020 (see Page 8, lines 174-177). Figure 5 showed the percentage of related articles published in different years more intuitively.

Comment 46: Lines 169-171: I am not sure what the point of this paragraph is. As more recently published articles are going to have fewer citations, this paragraph should be removed.
Reply 46: Thank you for your comments, the paragraphs on original lines 169-171 have been deleted (see Page 8, line 185).

Comment 47: Line 188: LLR is used but not defined
Reply 47: Thanks for the comment. LLR is one of the main analysis algorithms in CiteSpace. In this modification, it has added the full name of LLR "Log-likelihood rate" in line 198, as well as the principle of the algorithm (see Page 9, lines 198-200).

Comment 48: Line 190-192: It is unclear where that conclusion comes from. If it is from another study, reference is needed. If it is this study, then data supporting this claim is needed.
Reply 48: Thanks for the comment. The statement in lines 202-205 about "Thus, MALAT1 gene is a type of long non-coding RNA, which may regulate miRNAs associated with certain diseases, and then play a role in regulating the clinicopathological changes of cancer" came from the LLR clustering results. The selected articles were obtained by searching the keyword of "MALAT1" and "cancer".

The high-frequency keywords “miRNA”, “clinical pathology”, and “long chain” obtained by clustering suggested that the current research was focused on the regulation of cancer disease process by MALAT1 through miRNA, or MALAT1 affected the clinicopathological changes of cancer diseases, etc. This part of the result was speculated based on the clustering results in this article.

Comment 49: Line 196: is a repeat of line 173. As is 211
Reply 49: Thanks for the comment. According to your proposed amendment, the repeated statements in line 209 and the original line 211 have been deleted (see Page 9, line 209).

Comment 50: Line 196-199: this information does not appear to be relevant to the study, please remove or revise to highlight importance.
Reply 50: Thanks for the comment. The information in lines 209-210 was the specific parameters set during the analysis of highly cited authors. In this modification, it has deleted the parameter setting process and only retained the result analysis of the highly cited authors (see Page 9, lines 209-210).

There are some spelling and grammatical errors.
Comment 51: “researches” (line 11). Research is not a countable noun, so it is just research or studies.
Reply 51: Thank you so much for your comments and kindly for patience. As you said, these minor spelling and grammatical errors should have been avoided before submission. In this mistake, we have revised “researches” as “research”.

Comment 52: Furthermore, line 11-12 uses conflicting tense; “recent years” and “are collected”. Past vs present tense.
Reply 52: It was was revised as “published in recent years were”.

Comment 53: Same with line 12- 13 with “is” and “adopted”.
Reply 53: It was revised as “employed”.

Comment 54: Line 34: “was also” implies that it is no longer involved. Perhaps add “found to be” after also
Reply 54: The sentence was revised.

Comment 55: Line 39-40, “also” is used twice in one sentence
Reply 55: One “also” has been deleted.

Comment 56: Line 44 “Therefore” is a bit strange to start a new paragraph with and was just used in the previous sentence.
Reply 56: According to your opinion, the sentence was improved.

Comment 57: Line 44: same confusion with “is” and “adopted”, same in line 46.

Reply 57: It was revised as “employed” and “performed”.

Comment 58: “Generally” at the end of a sentence is awkward
Reply 58: It has been deleted.

Comment 59: “(uniformity)” is unnecessary. Homogeneity explains it clear enough
Reply 59: This word has been revised.

Comment 60: Suggested change “This study adopts the data in…”
Reply 60: It has been revised according to your suggestion.

Comment 61: Line 84: “on” should be “in”
Reply 61: It was revised.

Comment 62: Line 88: “Article” should be plural
Reply 62: It was revised.

Comment 63: Line 89: Tense confusion with “are” and “retrieved”. “Respectively” is not necessary here.
Reply 63: The tense is unified, and “respectively” was deleted.

Comment 64: Line 90: “should be” should be “was as follows”
Reply 64: It was revised.

Comment 65: Line 98: Tense confusion
Reply 65: The tense is revised.

Comment 66: Line 114: This is where “respectively” would fit nicely at the end.
Reply 66: “Respectively” was added here.

Comment 67: Line 123-124 is a bit awkwardly written. Please revise
Reply 67: They were revised.

Comment 68: Line 126: a conjunction word after the first comma is missing
Reply 68: It was revised as “Using the study-specific third-quantile distribution as the cutoff value, the MALAT1 expression data was finally divided into 3 categories ”.

Comment 69: Line 130: could be “In this study we used…”
Reply 69: It was revised as “In this study, we used *R* language to build a proportional hazard model and used proportional hazard regression analysis to compare ‘High’ and ‘Low’”.

Comment 70: Line 145: “Hotspot” is not scientific in this context, also on line 162
Reply 70: This word was replaced by “focus” and “front-burner issue”.

Comment 71: Line 155: Is "Web of Sciences" supposed to be plural? It is not in other areas.
Reply 71: It was revised to "*Web of Science*".

Comment 72: Line 173: "software" is not necessary.
Reply 72: "Software" was deleted.

Comment 73: Line 175: missing a word like "are" and "and"
Reply 73: It was revised according to your suggestion.

Comment 74: Line 238: Disclosed is an improper term here.
Reply 74: The word was replaced with "revealed".

Reviewer B
Comment 1: The authors present an interesting article that can help in this field. However, the manuscript has major deficiencies. The introduction should be expanded with a more clinical approach and adequately justifying the spirit of its study. Bibliographic references must be more current.
Reply 1: Thanks for the comment. We have supplemented the related research of MALAT1 in cancer in the introduction, in order to point out the influence of MALAT1 abnormal expression on the progress of cancer disease and the prognosis of patients, so as to highlight the importance of MALAT1 in cancer research. In this revision, relevant references in recent years have been added (see Page 4, lines 77-87, and Page 13, line 307). In addition, we deleted some conclusions that were not fully supported. Some of the conclusions in the paper have also been improved and marked.

Comment 2: The methodology should specify the inclusion and exclusion criteria of the patients themselves, the tumor grades and the manner of diagnosis in a clear and concise manner.
Reply 2: Thank you for your kind comments. As you said, the inclusion and exclusion criteria of the patients themselves, the tumor grades and the manner of diagnosis were the important factors in the methodology, which should be in a clear and concise manner. However, in this research, first, the cluster analysis was implemented via research about "MALAT1" and "cancer" in *Web of Science*. *Web of Science* was a typical citation database, in addition to the abstract of the document, it also contained citation information, which could be used for the analysis of citations between articles and the relationship between citations. Therefore, in the early stage of this research, we only used the relevant information in *Web of Science* for cluster analysis. We hoped to use CiteSpace to draw a knowledge map of MALAT1 in breast cancer in order to evaluate the research front-burner issues, but did not include the tumor grades and diagnostic methods of patients. Secondly, in order to further explore the correlation between "MALAT1" and "breast cancer", relevant data published in *GEO* were

collected for meta-analysis to evaluate the relationship between MALAT1 and breast cancer survival rate. In short, two kinds of databases were utilized to conduct analysis in different directions. The innovation of our research is visualize the correlation of different keywords via knowledge map, which can be more intuitive and provide a new way of thinking for the follow-up study of MALAT1 in the occurrence and metastasis of breast cancer and its effect on the prognosis of patients. However, the limitation lies in that since this study is based on big data, analysis software is adopted to explore potential associations, and we have not studied tumor grading and diagnosis methods for patients.

The analysis method of our study is as follows:

"**2.2 Data collection and preprocessing**

The data used for CiteSpace analysis could be collected from the Web of Science (http://webofscience.com), Chinese Social Science Citation Index, and China National Knowledge Infrastructure database (14). This study adopted the data in the Web of Science database to study the regulation of MALAT1 in cancer and breast cancer.

First, the subject terms "MALAT1" and "Cancer", and "MALAT1" and "Breast Cancer" were searched in the core database of Web of Science. Then, "Article" was selected as the document type, and time span from 1996 to 2020 was determined. Finally, a total of 1,180 articles with the subject word "MALAT1" and "Cancer", and 215 with "MALAT1" and "Breast Cancer" were retrieved. Each record was saved as a plain text document, the record format was as follows "author, title, source publication, and abstract", and the file name was "Download.txt". Then, the data-import/export function in CiteSpace was taken to transform the input.

2.4 Meta-analysis

Breast cancer data sets containing MALAT1 gene expression were extracted from the *GEO* database (https://www.ncbi.nlm.nih.gov/gds). This study adopted the data in Affymetrix human genome U133 plus 2.0 array and U133A array. Then, a total of six data sets were finally selected, namely GSE6532, GSE9195, GSE20711, GSE31448, GSE42568, and GSE1456. Third-quantile distribution meant that the MALAT1 expression in the chip data was arranged from large to small, it was classified into low (values from 0% to 33.3%) and medium (33.3 % To 66.6%) and high (66.6% to 100%). Using the study-specific third-quantile distribution as the cutoff value, the MALAT1 expression data was finally divided into 3 categories, which were "High" (MALAT1 expression> high tertile cutoff), "Mid" (low tertile cutoff ⩽ MALAT1 expression<high tertile cutoff), and "Low" (MALAT1 expression ⩽ low tertile cutoff) Deadline).

In this study, we used R language to build a proportional hazard model and used proportional hazard regression analysis to compare "High" and "Low". RevMan 5.3 software was employed for meta-analysis, and in the meta-analysis, a random effects model (DerSimonian and Laird method) was used to calculate the combined hazard ratio and 95% confidence interval."

Comment 3: The results must be explained and justified further. Figures must have a linear relationship with the manuscript. Figures must be modified. Authors should

consider increasing the quality of the figures, unifying colors. Figures should be more self-explanatory, with larger figure legends.

Reply 3: Thank you for your valuable comments. We have improved the quality of the figures in our manuscript. For example, the resolution of Figure 8 (original 9) has been improved to 300 dpi (see Figure 8). Since the images in the article were directly exported by analysis software, the color of the image could not be modified and adjusted, so the color of the image could not be unified. Specific explanations are as follows (Based on the previous revisions, we have deleted Figure 1, Figure 10 and Table 4, so the order of the charts has partially changed) :

"Figure legends

Figure 1. CiteSpace analysis process. The analysis process using CiteSpace was mainly divided into the following process. I, the time node and time slicing mode were decided; II, the analysis node type and the link type were selected; III, the similarity or proximity of the input data was calculated; IV, corresponding networks for each time slice were constructed; V, whether to perform network scaling was determined. If so, network scaling was performed; if not, the time series of the network were merged; VI, the merge of network time series after network scaling; VII, finally, the analysis process could be ended according to whether to perform network scaling and the network after scaling with merged time series.

Figure 2. CiteSpace parameter setting interface.

After the CiteSpace software was running, *Web of Science* was selected and a new program in the "Projects" column was created. The "project" folder created in the previous step was taken as the "Project Home" and the "data" folder as the "Data Directory". I. Time Scaling: time scaling was for setting and presenting the temporal evolution of the research data, and the time span could be selected based on the time range of the data. II. Text Processing: this module was divided into two parts: "Term Source" and "Term Type". III. Network Configuration: this module was divided into three parts: "Node Types", "Links", and "Selection Criteria". IV. Pruning: the "Pathfinder and Pruning the merged network" option was chosen. V. Visualization: the "Cluster View-Static" and "Show Merged Network" options were chosen.

Figure 3. Statistics on the number of publications on MALAT1 gene and Cancer in different years. The number of publications with the keywords "MALAT1" and "Cancer" from 1996 to 2020 was derived from the core database of *Web of Science*. Horizontal and vertical axes represented the year and number of publications respectively.

Figure 4. Statistics on the number of publications on MALAT1 genes and Cancer in different countries. The number of publications with the keywords "MALAT1" and "Cancer" from 1996 to 2020 was derived from the core database of *Web of Science*. Horizontal and vertical axes represented the country and number of publications respectively.

Figure 5. Statistics on the number of publications on MALAT1 genes and Breast Cancer in different years. The number of publications with the keywords of "MALAT1" and "Breast Cancer" from 1996 to 2020 was derived from the core database of *Web of Science*. Horizontal and vertical axes represented the year and number of

publications respectively.

Figure 6. Statistics on the number of publications on MALAT1 genes and Breast Cancer in different countries. The number of publications with the keywords of "MALAT1" and "Breast Cancer" from 1996 to 2020 was derived from the core database of *Web of Science*. Horizontal and vertical axes represented the country and number of publications respectively.

Figure 7. Clusters of core keywords of MALAT1 gene and Cancer. In CiteSpace, it set the "Top N%" and "per slice" as 1% and 100, respectively, to visualize the knowledge map of keywords.

Figure 8. LLR-based clusters of core highly cited publications on MALAT1 gene and Cancer. The highly cited literature clustering was performed by the LLR method, and a total of nine clustering tags (sunitinib, regulates, tumorigenesis, characterization, migration, metastasis-associated, transition, suppressor, and mir-9) were obtained.

Figure 9. LLR-based clusters of core highly cited publications on MALAT1 gene and Breast Cancer. Through the clustering of highly cited publications by the LLR method, 5 clustering tags (breast, long, PI3K/AKT/mTOR, coding, and -coding) were obtained.

Figure 10. Meta-analysis of the correlation between MALAT1 expression and breast cancer survival. A was the correlation between MALAT1 expression and recurrence-free survival; B was the correlation between MALAT1 expression and overall survival rate. The 95% CI value was the correlation value between MALAT1 expression obtained by proportional hazard model analysis and patient survival rate and recurrence-free rate."

Comment 4: The discussion should be expanded, with a more translational character and with the implications that your research could have. The authors should also mention male breast cancer. The survival of these patients is something that is not very clear in the manuscript.

Reply 4: Thank you for your kind comments. We have expanded the discussion according to your suggestions. The revised discussion are as follows (see Pages 10-12, lines 243-278):

"**4. Discussion**

Studies have revealed that elevated MALAT1 expression levels are closely related to the prognosis of breast cancer patients. Zheng et al. (2018) found that the expression of MALAT1 in breast cancer tissues increased significantly, and the overall survival rate was negatively correlated with the expression of MALAT1, indicating that MALAT1 was a potential prognostic factor (31). Sun et al. (2020) detected a significant increase in the expression level of MALAT1 in the serum of breast cancer patients (32). Then, they found that the MALAT1 level of patients was significantly reduced after treatment, and that the MALAT1 level of patients with good prognosis was significantly lower than that of patients with poor prognosis. In our study, MALAT1 expression data from the *GEO* database were used, and the results showed that high MALAT1 expression was significantly associated with relapse-free survival, but not with overall

survival. This conclusion is slightly different from previous studies and may be caused by the selection and sample size of MALAT1 data in *GEO* database. This association may remain significant after adjustment for age at surgery, disease stage, tumor grade, and hormone receptor status due to the complexity of breast cancer staging and typing. Totally, it could be inferred that the increase of MALAT1 expression level could lead to poor prognosis of breast cancer patients, and then increased the mortality rate. Unfortunately, there has been no report on the role of MALAT1 in male breast cancer. The reason for this is that the incidence of male breast cancer is much lower than in women. As a result, in the actual research process, the isolated tumor samples of male breast cancer are difficult to obtain and is relatively of small number. Besides, the research value is far lower than that of female breast cancer. However, it is possible from existing studies that MALAT1 may also be a pro-tumor regulator in male breast cancer. Of course, this question is worth exploring further.

In conclusion, the core data of Web of Science were taken to collect the literature related to MALAT1, Cancer, and Breast Cancer in recent years. CiteSpace was adopted to visualize the correlation of different keywords. The results revealed that the lncRNA expression level of MALAT1 was closely related to cancer progression and prognosis of cancer patients. Moreover, it was also found that the expression of MALAT1 was closely related to the survival rate of recurrence-free breast cancer. However, it only adopted the existing literature for visual analysis, and subsequent experiments needed to be done to verify this conclusion. In conclusion, the results can provide evidence for lncRNA MALAT1 as a biological marker for diagnosis and treatment of breast cancer."

Comment 5: Authors must adequately improve the English grammar of their manuscript.

Reply 5: Thank you for the comment. We also note the deficiency in the quality of writing, thus we have invited two professor to do the modification of writing. We hope the revised manuscript would be more readable (such as Page 3, line 50).