General response: Firstly, we would like to thank you for your kind letter and for reviewers' constructive comments concerning our article (Submission ID: 87022). These comments are valuable and helpful for improving our article. All the authors have seriously discussed about all these comments. According to the reviewers' comments, we have tried best to modify our manuscript to meet with the requirements of your journal. In this revised version, changes to our manuscript within the document were highlighted by using yellow colored text. Point-by-point responses to the reviewers are listed below this letter.

Point-by-point response

Reviewer #1:

1. In the background section of your abstract, CNS? (explicate this abbreviation).

Thank you for your reminder. CNS is an abbreviation for the central nervous system. We have incorporated this revision into the background section and the final list of abbreviations.

2. In the manuscript background after your 3rd reference, please cite the rate of this incidence in development countries.

Thank you for your nice comments on our article. Endemic BL is primarily found in equatorial African countries. The estimated annual incidence of endemic BL is 3-6 cases per 100,000 children in African countries, which is approximately 50 times higher than that in the United States. Sporadic BL predominantly occurs in the United States and Western Europe. In the United States, the annual incidence of BL is approximately 3 cases per 1 million individuals, while the annual incidence in Europe stands at around 2.2 cases per 1 million people. Immunodeficiency-associated BL mainly affects HIV-infected patients and typically involves those with relatively high CD4 counts and no opportunistic infections.

We have incorporated this revision into the background section.

3. Why the LDH rate which is a relevant biological marker have not been integrated in your nomogram?

Thank you for your reminder. The nomogram in the article is based on the GSE69051 cohort, which does not include LDH. Consequently, the integration of LDH rates has been omitted from the nomogram.

4. GS and MM do not appear in the final list of abbreviations.

Thank you for your reminder. GS is an abbreviation for gene significance, and MM is an abbreviation for module membership. We have incorporated this revision into the "Materials and Methods" section and the final list of abbreviations.

5. *Your nomogram have not been discussed, despite your cites in you bibliography a reference about a nomogram used in BL, Lu, J., et al., Status and prognostic nomogram of patients with Burkitt lymphoma. Oncol Lett, 2020. 19(1): p. 972-984.)*

Thanks for your help. In accordance with your suggestion, we have extensively discussed the nomogram in the dedicated discussion section of the manuscript.

Reviewer #2:
Specific Comments to Authors:

(1) Abstract: most probably "CNS" abbreviation might need an explanation. It is used only two times throughout the paper, so alternatively you can put the full description.

Thank you for your reminder. CNS is an abbreviation for the central nervous system. We have incorporated this revision into the background section and the final list of abbreviations.

(2) Abstract: In "Chips and sequencing information", by "Chips" you meant microarray data? If yes, please change it. Same is on page 4 in the section "Background".

Thanks for your help. We feel really sorry for our carelessness. The article has been appropriately modified in accordance with the required changes.

(3) Abstract, Aims: you can add to the first sentence that you wanted to find hub genes in perform gene ontology specifically in BL. In the second sentence, most likely the "carry out" would be "carried out" and "construct" would be "constructed". Some parts of the manuscript might need to be reviewed by a native speaker; please double-check the language quality in your paper.

Thanks for your careful checks. We are sorry for our carelessness. Based on your comments, we have made the corrections to make the unit harmonized within the whole manuscript.

(4) Abstract, Methods: The second sentence starting from a full name of WGCNA is not capitalized.

Thanks for your help. We feel really sorry for our carelessness.

(5) Abstract, and the entire paper: Did you mean "cytoHubba" instead of "cytoHub" by any chance?

Thank you for your reminder. We feel really sorry for our carelessness. Our preference lies in utilizing cytoHubba instead of cytoHub, and the article has been appropriately modified.

(6) Why hub genes were identified using a tool that is separated from WGCNA toolkit? WGCNA does provide Module Membership (MM) values that can help in the investigation of hub genes, and combining it with the Gene Significance (GS) values can result in the identification of a so-called "driver genes". What was the reason of not using WGCNA in this step? Moreover, did you verified cytoHub/cytoHubba results using MM values in WGCNA, or were they not used at all in the workflow?

Thank you for your reminder. The MM-GS diagram, depicted in Figure 4, illustrates the association between genes and phenotypic traits. Each circle represents a gene, with the horizontal axis indicating the correlation of the gene with its corresponding module, while the vertical axis represents its correlation with the phenotypic trait (age). Notably, this visualization reveals that genes exhibiting high significance in relation to a specific trait often play crucial roles within modules significantly associated with that particular trait.

The network serves as a valuable tool for presenting diverse biological data, encompassing protein-protein interactions, gene regulations, cellular pathways, and signal transductions. By evaluating nodes based on their network features, we can infer their significance within the network and effectively identify central elements of biological networks. CytoHubba offers a user-friendly interface for the exploration of pivotal nodes in protein-protein interaction (PPI) networks. Therefore, in this study, we used cytoHubba tool to screen the hub genes.

(7) One of your aims was to perform a survival analysis, then why you only focused on overall survival and no other endpoints such as disease-free survival etc.? Events caused by disease recurrence occur earlier than death from the disease, so it may be beneficial to include such an endpoint instead of overall survival. Preferably, one could include more than one survival endpoint, especially if intending to perform a survival analysis.

Thanks for your suggestions. Due to the limited sample size of GSE4475, we opted for GSE69051 as the dataset for survival analysis. Although it would be advantageous to incorporate PFS in our analysis, unfortunately, GSE69051 solely provides information on OS and lacks data on PFS. Consequently, our survival analysis only focused on OS.

(8) Remember to italicize all gene symbols throughout the paper.

Thanks for your help. The article has been appropriately modified in accordance with the required changes.

(9) Abstract, results: Consider changing "Moreover, we found 2 hub genes associated with OS" to "Within these hubs, two genes were associated with OS", next in the part "we combined the two hub genes", "the" could be "these". Lastly, rewrite the sentence starting with "And" and try to avoid repetition of "we".

Thank you for your nice comments on our article. The article has been appropriately modified in accordance with the required changes.

(10) Abstract, results: Once I completed reading the paper, I am unsure if you "found several potential therapeutic targets for BL with poor prognosis". Therapeutic targetability of these findings were not investigated in your study, or at least not in a proper way. The best bet would be to focus on prognostic significance of investigated age-related biomarkers. Consequently, I would avoid "Therapeutic target" keyword.

Thanks for your reminder. On the investigation of therapeutic targets, we only screened drugs associated with hub genes in the DGIdb database. Due to limitations in experimental conditions, further research was not conducted. Consequently, this aspect served as a supplementary component to the primary content of this article, aiming to provide potential avenues for future related investigations. We have incorporated these modifications into the results section of the Abstract.

(11) Abstract, conclusion: add "that" before "might"

Thanks for your help. We feel really sorry for our carelessness.

(12) Throughout the text, some space marks are missing. Please double check the entire manuscript and checked whether words are separated from brackets, citations, etc.

Thank you for your nice comments on our article. We feel really sorry for our carelessness.

(13) Background: The part "while chronic EBV (Epstein-Barr virus, EBV) infection plays an important role in BL" could be "with chronic Epstein-Barr virus (EBV) infection playing an important role in BL". Moreover, you can specify in which clinical type of disease.

Thanks for your help. The article has been appropriately modified in accordance with the required changes.

(14) Background: The part "MYC regulates the expression of target genes which regulate a variety of cellular processes" could be "MYC orchestrates the expression of target genes, regulating a variety of cellular processes".

Thank you for your nice comments on our article. The article has been appropriately modified in accordance with the required changes.

(15) Background: In the part "and other forms of high-throughput functional genomic data, which submitted by research communities", change "which submitted" to "that are submitted".

Thanks for your help. The article has been appropriately modified in accordance with the required changes.

(16) Background: whether the short synopsis about GEO and WGCNA is necessary in this section is a matter of debate. In my opinion you can shorten these descriptions and move what is left to the methodological section.

Thank you for your nice comments on our article. The article has been appropriately modified in accordance with the required changes.

(17) Background: In the part "was carried out to identify a mRNA signature which significant associated with prognosis. Finally, a prognostic nomogram was established based on the combination", you can change "which significant associated" to "that was significantly associated with" or "presenting a significant association with". If you would like to avoid two -ed in "established based", you can change it to the "established on the basis of".

Thanks for your help. The article has been appropriately modified in accordance with the required changes.

(18) Materials and Methods: you can consider deleting "s" and the end of "raw gene expressions". I think "expression" would be better. Double-check the entire paper.

Thanks for your help. We feel really sorry for our carelessness.

(19) What was the reason of selecting these two (GSE4475 and GSE69051) GEO datasets? Once I completed reading the paper, it turned out that GSE69051 was only used after GSE4475 because GSE4475 was too small to perform survival analysis. Alternatively, the entire workflow could have been performed only on GSE69051 since there was no verification of results from one dataset in the other. Moreover, if you aimed to perform survival analysis, why not searching for a dataset that included more than one survival endpoint? Were there some problems in finding such datasets?

We sincerely appreciate your invaluable advice. Our study utilized a publicly available database; however, due to the low incidence of BL and limited data sets, we encountered challenges in identifying alternative suitable free and open datasets for validating our findings. Furthermore, the unavailability of pathological specimens and experimental conditions posed limitations on conducting verification experiments, which is a constraint in our study design. In future endeavors, if we are able to locate relevant databases or obtain appropriate experimental conditions for validation purposes, we will strive to enhance the robustness of our research.

(20) What kind of tool was used to perform this step: "The mRNA sequencing data annotation information was used to match the probe with the corresponding gene to transform the gene name into gene symbol". I presume it was not done manually. Have you tried g:Profiler or SYNGO, or something similar?

Thank you for your reminding. We obtained the GPL annotation file for the GSE dataset and subsequently performed chip ID to gene symbol mapping using R.

(21) Section 2.2: The part "The top 5,000 variant of expression profiles were used" could be "The top 5,000 most variable genes were used", assuming I am correct in understanding of your workflow.

Thanks for your help. The article has been appropriately modified in accordance with the required changes.

(22) Section 2.2: In the descriptions of WGCNA, you did not specify if you selected unsigned, signed, or signed hybrid approach.

Thanks for your help. In this study, the soft threshold β was 12, and the networkType = "signed". The article has been modified.

(23) In methods, instead of adding links to websites, consider adding citations for each tool to help authors gain attention. If no preferred citation is available, URL links are okay.

Thank you for your reminding. We have added citations for each tool in methods.

(24) Section 2.2: "Samples cluster analysis was performed using the hclust tool" means that you did not use a built-in clustering options from WGCNA?

Thanks for your suggestions. We apologize for the misunderstanding. We utilized the hclust function from the WGCNA package to perform sample cluster analysis. To ensure clarity, we have omitted this statement from the original article.

(25) Section 2.2: Beta-power used in WGCNA is mentioned in Results (it was 12 if I understood correctly), but please mention it also in the methodology.

Thanks for your help. In this study, the soft threshold β was 12, and the article has been appropriately modified in accordance with the required changes.

(26) Should section 2.3 be a part of WGCNA toolkit? Especially if you refer to MM and GS values at the end of this section. Moreover, the usefulness of GS/MM values in your study is not evident. Currently it appears that you omitted GS because you did not investigate driver genes but instead focused on hub genes, and secondly MM values from WGCNA were also not used because hub genes were indicated by a separate tool (cytoHub/cytoHubba) and not using values provided within WGCNA toolkit. Moreover, in the same section – what is "thermal mapping kit"? Did you mean module-trait relationship? If yes, then as I said at the beginning – all these descriptions should be a a part of WGCNA, and thus you could merge them with section 2.2.

Thank you for your reminder. Section 2.3, which is an integral part of the WGCNA analysis, has been integrated with section 2.2.

(27) Section 2.5: Most probably the GeneMania includes both PPI and GI data. Have you filtered out PPI data from the server so to obtain GI only? In the same section, the last sentence might be misleading ("The statistical significance was expressed as a collective score of >0.15"). I would change "statistical significance" to something else, or just write that the threshold of collective score of 0.15 was applied on the server, because in the next section you refer to the "real" statistical significance that was a typical p<0.05, which could introduce uncertainty among Readers.

Thanks for your help. As you are concerned, we filtered out PPI data from the server so to obtain GI only. Besides, according to your suggestions, we have modified the last sentence in this section in the revised version.

(28) Section 2.6: In the part "According to the 50th percentile cut-off value of each hub gene mRNA, patients were divided into the high-expression and low-expression groups" – that means that all hub genes were in fact protein-encoding? Were non-coding data included in the GEO datasets? Moreover, applying median cut-off is not always a proper way from the biological and clinical point of view. Have you tried applying a cutpoint using relevant tools? Some genes in Figure 8 could be statistically significant (like in subfigure B, C, H) if other cut-off would be applied.

Thank you for your valuable comments. In this section, we conducted an analysis of the expression data of hub genes. Specifically, all the 10 hub genes (SRC, TLR4, CD40, STAT3, SELL, CXCL10, IL2RA, IL10RA, CCR7 and FCGR2B) we focused on were protein-coding gene. Prior to the analysis process, non-coding data was filtered out during data sorting and thus excluded from our investigation. The rationale behind employing a median cut-off in this study stems from its widespread usage in previous literature, as well as there is limited availability of original bioinformatic articles on BL which could have provided us with a more suitable cut-off value.

(29) Section 2.6: The last sentence states that "Additionally, P <0.05 was statistically significant unless otherwise indicated". I did not see any other part that indicates statistical significance and threshold, so maybe the part "unless otherwise indicated" is unnecessary?

Thank you for your nice comments on our article. The article has been modified and delated the part "unless otherwise indicated".

(30) Section 2.7: The name of the database "DSigDB" is not in line with the website to which a link is provided next to the name (URL link refers to DGIdb tool). Please double-check and correct whichever is wrong. Moreover, from what I understood, this part cannot be technically described with more details, because once the webtool is accessed, only gene symbol is provided and all results are automatically provided in the new tab? There are not filtering, thresholding, etc.?

Thank you for your assistance. We deeply regret our oversight regarding the database name, which should be referred to as "DGIdb". While it is true that only gene symbols were provided and all results were automatically generated without thresholding, we have included this information as a supplementary addition to the main content.

(31) Section 3.1: Change "In this study, we obtained the BL dataset in GSE4475, A total of 13, 514 gene expression values were derived from the raw file" to "In this study, we obtained the BL dataset from the GSE4475, resulting in a total of 13,514 gene expression values".

Thanks for your help. The article has been appropriately modified in accordance with the required changes.

(32) Section 3.1: In the part "Then, we selected a total of 5, 000 genes with the greatest average expression values for cluster analysis", should it be about most variable genes, not the ones with the greatest average expression? Focusing on greatest expression would be inappropriate since you can have a biologically meaningful change in the expression that is relatively small compared to others that are not so crucial.

Thanks for your advice. In WGCNA analysis, the selection of genes to be analyzed is a crucial decision when faced with an overwhelming number of genes, as an excessive gene pool can lead to intricate and challenging interpretation of the co-expression network. Common strategies of choosing a subset of genes exhibiting highest expression levels or selecting those demonstrating the greatest variation in expression.

In this article, we have selected genes with the highest expression for analysis due to several reasons: 1. Ensuring a stable signal: Genes with higher expression levels often yield more reliable signals, enabling a more accurate reflection of biological processes. By selecting these genes, we can mitigate the impact of noise and effectively capture co-expression patterns. 2. Reducing computational complexity: Opting for highly expressed genes helps simplify network complexity and streamlines the calculation and analysis process. 3. Enhancing biological significance: Genes displaying high expression frequently play pivotal roles in biological processes; thus, their selection increases the likelihood of identifying biologically relevant modules and associations pertinent to research.

(33) Section 3.1: I would move this sentence to the figure's description "Red indicated more gene expression, white less, and gray indicated deletion (Fig. 1)" and moved reference to Fig1 to previous sentence in the text. Moreover, you should use other words than "indicated deletion" for gray in this context, it would be better to say that it represents an unknown status for some samples. This is equivalent to "unknown" you used in Table 1.

Thanks for your advice. The article has been appropriately modified in accordance with the required changes.

(34) Figure S1 and S2 are unavailable for me to assess – I cannot identify them in the system or in the manuscript file.

Thanks for your help. We feel really sorry for our carelessness. We have incorporated the supplementary figures into both the manuscript and the system.

(35) Section 3.1: Please standardize the use of the word "gray" or "grey".

Thanks for your help. We feel really sorry for our carelessness.

(36) Section 3.1: In the part "Genes in gray were not included in any module, then we analyzed", the word "thus" would fit better than "then". Right after this sentence, I would delete "After docking with clinical character data".

Thanks for your advice. The article has been appropriately modified in accordance with the required changes.

(37) Section 3.1: Rationale for selecting the age as a clinical trait of interest is mentioned once in Discussion, but its relevance is not mentioned in Results. Moreover, why focusing only on age when there were also other traits that were found significantly correlated with gene modules, as shown in Figure 3?

Thanks for your reminder. We appreciate your positive feedback on our article. In response to your valuable suggestions, we have provided additional justification for focusing solely on age in section 3.1. We exclusively focused on age due to its confirmed prognostic effect in previous literature, while other traits lacked clear prognostic value. Therefore, we believed that analyzing the module most closely associated with age would yield more meaningful outcomes.

(38) Figure 1 could be better described (see my comment no. 33), with more details. Moreover, CCS and Ki67 are not mentioned in Table 1, while they constitute clinical data similar to age, sex, and stage.

We express our gratitude for your valuable advice. In accordance with your insightful suggestions, we have modified this part in the manuscript.

(39) Figure 2: The last sentence of figure description ("The yellow brightness of the middle part represented the strength of connections between modules") could mention about darker shades of yellow turning into brown or orange.

Thanks for your advice. The article has been appropriately modified in accordance with the required changes.

(40) Figures in general would benefit much from increasing font size.

Thanks. We have increased font size in figures based on your suggestions.

(41) Figure 4 represents scatter plot with both GS and MM. If applying 0.6 threshold, a few driver genes would be identified but this threshold is not ideal. Have you at least tried to identify if genes most correlated to both trait and module were indicated as hub genes using cytoHub/cytoHubba?

Thank you for your inquiry. This question bears resemblance to the one discussed in detail in my response to question 6.

(42) It might be hard to increase font in Figure 5; thus, please consider moving it to the supplementary materials. In the same figure, if top hubs the ones that are in the center of the network? If yes, then I see 12 nodes, whereas you mentioned about 10 top hubs. How can we identify top hubs in this figure before you focused on them in further steps?

Thank you for your invaluable assistance. In this illustration, each node represents a gene, and we have arranged the gene nodes based on their degree of connectivity. The centrally positioned nodes in the image corresponds to genes with higher degree, although it may not precisely align with the hub genes. Following your insightful suggestions, we have made revisions to our manuscript and relocated Figure 5 to the supplementary materials section.

(43) Figure 6: what is "with a common goal" or "that implemented common goals"? Moreover, in the same figure the slash is barely visible, please change it something more evident or increase the quality of the figure. The figure would generally benefit from small legends for each subfigure. For example, in subfigure A you can show that edges represent co-expression. Figure's description should be updated afterwards.

We express our gratitude for your valuable advice. The manuscript has been modified in accordance with your suggestion.

(44) Section 3.3: the part "were shown in Fig. 7A" or the equivalent for 7B could be put in brackets and moved earlier in sentences, next to the "top 10 GO/KEGG terms", so that the remaining part of the sentences would be only the name of terms.

Thanks for your advice. The article has been appropriately modified in accordance with the required changes.

(45) Table 2 might need to be moved to the supplementary materials due to its size. Consequently, Table 3 might be moved too, even though its size it not that large. However, top 10 terms from GO or KEGG are visible in Figure 7, so it is not a big deal to make both tables as supplements. Another question, how "top" terms established? Based on p-value, number of annotated genes ("count") or what? While on the topic, you can also explain the meaning of "count" column.

Thanks for your advice. The article has been appropriately modified in accordance with the required changes. "Top" denoted the sorting criterion based on p-values, and "count" means the number of genes.

(46) Section 3.6: Enrichr website was first-time mentioned in this section; it is not present in methodology. Are you sure that association with drugs for IL2RA and CXCL10 was investigated using Enrichr built-in tool? Because methodology stated a specific URL link that is outside Enrichr. Please justify and correct. For the same section, description is rather weak, so is the part of Discussion related to it. You can mention that you focused only on IL2RA and CXCL10 because these were the only significant results from survival analysis. I also found that Discussion is lacking details on current drugs that are used in BL and maybe appeared in the results of this analysis. Some drugs common for IL2RA and CXCL10 might be further discussed. In general, current description is too short. Druggability of IL2RA and CXCL10 should be discussed thoroughly if you would like to leave some prospects for the future about therapeutic potential etc. Mentioning, e.g., that "This makes CXCL10 a 'key driver chemokine' and a valid target for therapy" entails a proper justification.

Thanks for your help. we were really sorry for our careless mistakes. We have delated the 'Enrichr website' part based on your suggestions. Moreover, we have added a description of the drugs for standard treatment of BL in the first paragraph of the discussion section. However, upon extensive literature review, we discovered a paucity of studies investigating IL2RA and CXCL10 related drugs in previous research. Furthermore, due to its supplementary nature rather than being the focal point of this article, we refrained from delving into detailed discussions on this topic with the intention of providing a foundation for future drug research in BL treatment.

(47) Discussion: In the part "between c-MYC and the gene for either the kappa or lambda light chain", would "of" be better than "for"?

Thanks for your advice. The article has been appropriately modified in accordance with the required changes.

(48) Discussion: Is apoptosis really that high in BL, similar to proliferation? The part "The proliferation rate and apoptosis rate of BL tumor cells are extremely high" suggests so.

We appreciate your inquiry. It has been extensively documented that a high rate of apoptosis is a common feature in BL.

(49) Discussion: "Ten hub genes (SRC, TLR4, CD40, STAT3, SELL, CXCL10, IL2RA, IL10RA, CCR7 and FCGR2B) and several pathways were identified by WGCNA". What pathways were identified using WGCNA?

Thank you for your help. We sincerely apologize for our carelessness. Key modules were identified through WGCNA analysis, hub genes were determined using cytoHubba, and pathways were analyzed via KEGG analysis. Appropriate revisions have been made to the manuscript.

(50) Discussion: In the part "And then, we used nomogram to find a new risk assessment system", the part "And then" could be "Afterwards". Later on, "What's more" should be "What is more".

Thanks for your advice. The article has been appropriately modified in accordance with the required changes.

(51) Conclusion: mentioning about "driving genes" only in this part is not appropriate way of referring to "hub genes". GS values were practically omitted in your WGCNA approach and MM values were probably not used because external tool was used for hubs identification. The part "that might be new therapeutic targets" could not conclude your findings. I would rather extensively enrich the paper with therapeutic methodology workflow and discuss it properly, or alternatively provide statements more related to prognostic significance of identified biomarkers. In the last sentence of Conclusion, "A" before "nomogram" must be lowercased.

Thanks for your nice suggestions. We have corrected these mistakes based on your suggestions.

(52) Another study limitation is the use of only public datasets (you can mention it in the relevant part of Discussion). However, once properly presented and discussed, it is acceptable.

Thank you for your nice comments on our article. According to your suggestions, we have supplemented this limitation in our previous draft.