Peer Review File

Article information: https://dx.doi.org/10.21037/gs-23-124

Reviewer A

The main idea of the study was to use HPO on large-scale gene expression data for the purpose of a (somehow meaningful) subclustering of PTC, which sounds quite interesting, However, the concrete implementation (at least as described in the current draft) is so confusing that even having read the paper several times I cannot actually follow the approach used to generate and validate the subclusters. There are so many doubts and questions that I cannot list all of them. Here just a few most serious:

1. The authors used data from two very unequal databases/cohorts (GEO and TCGA). The GEO cohort was small (just 40 patients), so its representativity is questionable. For example, the M:F ratio in this cohort was appr 1:1, which does not reflect the natural situation (M:F = 1:3 to 1:4 in the practice) and suggests that especially female PTCs might be significantly undersampled. Furthermore, the gene expression data in 3 out of the 4 GEO datasets used are quite old (published 2004 to 2006) and based on older versions of Affymetrix RNA microarray chips, which proved to be quite unreliable. The TCGA cohort, by contrast, is large, features realistic M:F ratio and contains relatively new and very solid data, so it is much more trustworthy.

Reply: Thank you for your comment. It should be emphasized that this study mainly takes the samples of TCGA database as the main body of analysis, and the samples of GEO databases are only used to verify the expression status of the subtype characteristic genes of TCGA samples in cancer and normal tissues of other data sets. Moreover, the 40 patients in GEO are from 4 different databases and combined to improve the universality of verification. In addition, in the combined data set, the number of women is still greater than that of men, which can reflect the natural situation.

2. The way how the authors used these two different cohorts in the study is extremely confusing. In principle, one possible (and very prudent) approach would be to use one cohort (especially the less reliable one, i.e. GEO) as the ìlearning setî and the other cohort (reliable TCGA set) as a completely independent ìvalidation setî for the clustering. Another approach would be to ìpoolî both cohort to increase the sample size (which is also always a good idea). Yet, instead of taking any of these two approaches, the authors followed a very strange way that I cannot understand at all. First, they obviously used the cohorts SEPARATELY to screen for the significantly up- and downregulated genes (as stated in lines 137-139), the results being then somehow ìjoinedî (line 140: ìFinally, the intersection of the arrayExpress and bulk RNA-seq results was determinedî) ñ does it mean ìvalidationî of DEG? Subsequently, an unsupervised clustering was performed again SEPARATELY on both GEO and TCGS data sets using different packages and procedures (lines 144-147), for a reason that I cannot understand either. So the way how the authors arrive at the 4 final PTC clusters is thus absolutely ominous at the end.

Reply: Thank you for your comment. In this study, only the subtype analysis was carried out on the data of TCGA. Bulk RNA-seq refers to TCGA, while the data sets of GEO are all array data rather than RNAseq data (lines 144-147). HPO is composed of thousands of gene sets. In this paper, the HPO subgene sets with common differences in HPO scores between tumor and normal were screened, rather than the intersection and screening of differential genes (lines 137-140). To prevent ambiguity, these explanations are explained in the “Methods”.

3. What was actually the advantage of using HPO scores as the basis for clustering, instead of using the gene expression data directly? Does one need HPO at all? Being the main idea of the study, this topic remained completely forgotten in the paper. To answer this question, two

separate analyses (one based on HPO and another based on the direct genomic data) had to be performed, but this was not done.
Reply: Thank you for your comment. HPO can provide comprehensive logical criteria for describing and calculating phenotypic abnormalities found in human diseases. Phenotypic refers to these aspects of the disease: an etiology (whether identified or as yet unknown), a time course, a set of phenotypic features, treatments exist, there is a characteristic response to them, refer to the article (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7778952/).
Using multiple gene sets as sample characteristics for subtype clustering can improve the robustness of subtype clustering. The characteristics of the subtype belong to a functional set rather than a single molecule. Since only mRNA level analysis is available, it is often not robust to describe a subtype by a single molecule, unless the molecular characteristics are described by multiomics around the central rule. Often, the expression of some characteristic molecules at the level of protein, RNA and DNA is not consistent, while the function of gene sets obtained through enrichment analysis is often consistent. Before doing this research, we consulted the ideas of two articles (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6697729/; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9046667/).

4. The ìfunctional characterizationî of the clusters (starting with line 211) actually discloses that the identified gene sets do not make ANY sense. One finds virtually everything, from brachydactyly over dermatological and jaw muscle disorders and up to bloody diarrhea. Are the authors really serious about these results?? Honestly, such clustering is not plausible to me.
Reply: Thank you for your comment. So that's the point of the HPO, you can see some associations between disease phenotypes. Using multiple gene sets as sample characteristics for subtype clustering can improve the robustness of subtype clustering.

5. There are several well recognized and established histopathological PTC subtypes substantiated by the corresponding genetic background (WHO classification). At least in TCGA, these PTC subtypes are annotated for each tumor. Comparing the clustering results from this study with the established WHO subtyping would maybe contribute to more plausibility of the results. However, the authors did not undertake any effort towards such a comparision.
Reply: Thank you for your comment. In this study, we first identified subclusters in PTC based on HPO and found that patients with distinct subclusters have different prognoses. We then identified and validated the characteristic genes in the 4 subclusters. These findings are expected to serve as a crucial reference that will improve our understanding of PTC heterogeneity and the use of novel targets.
The classification of thyroid tumors has evolved based on classic histopathology and molecular pathogenesis. Most encapsulated/circumscribed thyroid tumors with a predominant follicular growth pattern exhibit a RAS-like molecular profile. On the other hand, most thyroid tumors with a $BRAF^{V600E}$-like molecular profile have papillary and/or infiltrative growth and florid nuclear atypia. Both $BRAF^{V600E}$-like and RAS-like thyroid cancers can gain additional genetic alterations and, as a result, progress to high-grade cancers. However, data of braf_gene_genotyping or RAS_gene_genotyping in TCGA data is missing a lot, so we cannot be compared the clustering results from this study with the WHO subtypes.

Prior to survival analysis, it would be necessary to look at the possible associations of the clusters with other clinical and pathological parameters (age, sex, stage, treatment etc.). Although these data are summarized in Table S4, no further analysis was performed. As a consequence, the univariate survival analysis for the clusters in this study is not reliable and cannot be trusted to.
Given these major flaws, I cannot recommend this study for publishing. It must be completely re-designed and the analysis performed anew, in order to address the main idea appropriately.
Reply: Thank you for your comment. We tried to do this analysis before, and the results are shown in the Table S4. However, we see no clear correlation between PTC subclusters and histopathological subtypes. In this study, we identified subclusters based on HPO and found

statistically significant differences in survival among the 4 subclusters, patients in Clusters 1 and 3 had better survival times, while those in Cluster 4 had the worst survival time (P=0.028). These results suggest that subclusters based on HPO can be used to predict the clinical prognosis of PTC. Despite the promising results, a number of questions remain. For example, more experimental studies need to be conducted to establish predictive models based on HPO. Additionally, further research should be undertaken to investigate why the patients in Clusters 1 and 3 had better prognoses and to explore the unreported function of such characteristic genes in PTC. We hope our manuscript will receive your consideration.

Reviewer B

The study identifies subclusters of papillary thyroid cancer (PTC) using the Human Phenotype Ontology (HPO) platform. The manuscript is well-written and interesting, especially for the readers focused on endocrine-related cancers. I recommend that this paper be accepted after major revision. Suggested changes are listed below.

1. The inscriptions on some figures are difficult to read (e.g., Figs. 1 and 2). Please make the font bigger if possible.

Reply: Thank you for your comment. We revised the figures according to your comment.

2. The Authors mention that SLC5A5 expression level was specifically downregulated in clusters 1 and 3. It is worth mentioning what is the physiological function of the protein encoded by this gene (NIS) in the light of the use of radioiodine-based therapy in PTC. Additionally, I would suggest mentioning that SLC5A8 gene encodes apical iodide efflux transport protein.

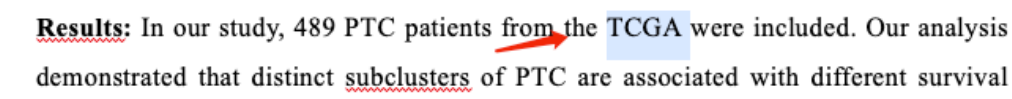
Reply: Thank you for your comment. We revised the manuscript according to your comment.

3. I would suggest mentioning the numerous studies aimed at identifying the molecular subtypes of PTC (e.g., Cell. 2014 Oct 23;159(3):676-90; (ii) Oncogene. 2022 Nov;41(47):5121-5132; (iii) J Transl Med. 2023 Mar 20;21(1):206; (iv) Exp Mol Med. 2022 Mar;54(3):263-272)). Please discuss the obtained results in the light of the available literature.

Reply: Thank you for your comment. We discussed the obtained results in manuscript. We hope our manuscript will receive your consideration.

Reviewer C

1. Please ensure all abbreviated terms are defined the first time they appear in the Abstract.


Results: In our study, 489 PTC patients from the TCGA were included. Our analysis demonstrated that distinct subclusters of PTC are associated with different survival

Response: Thank you for reminding. The abbreviated term of TCGA was defined. And we confirmed all abbreviated terms are defined the first time they appear in the Abstract.

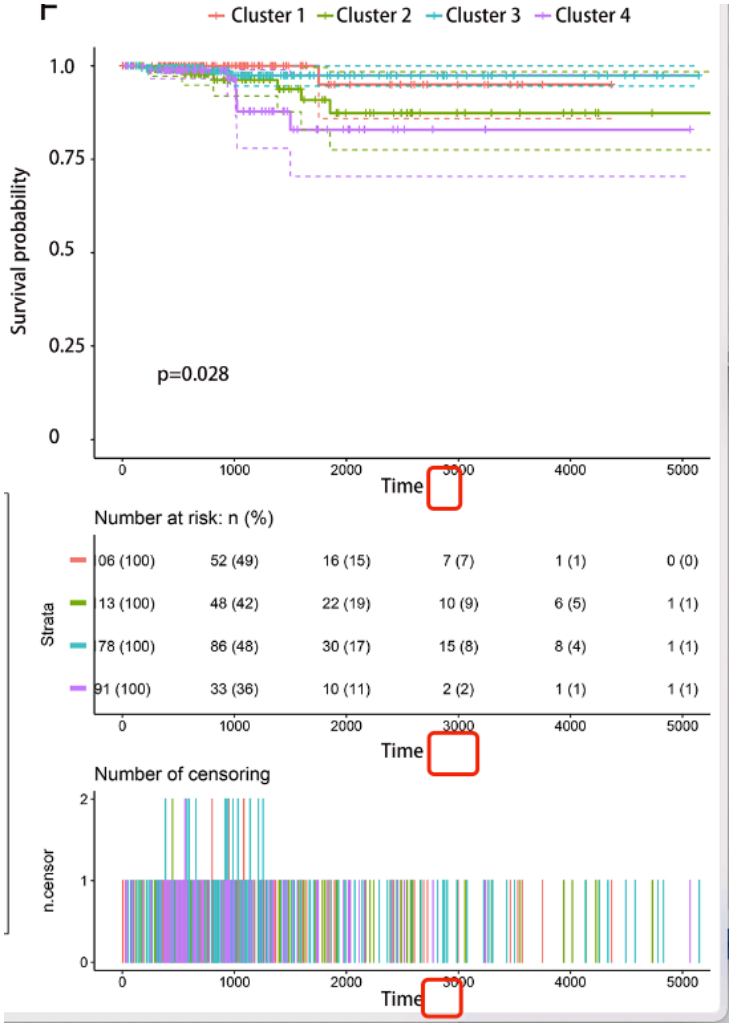
2. You've mentioned "studies", while only one reference was cited in the below sentences. Please check.

Previous studies have noted that HPO is widely available for the differential diagnosis of rare diseases, phenotype-driven investigations based on next-generation sequence-variation data, and translational studies (30). To the best of our knowledge, no previous

and *ARHGAP36* in PTC have been reported. For example, previous studies have reported that *SLC5A5* is expressed at a lower level in PTC, and the lower expression of *SLC5A5* is correlated with aggressiveness and *BRAF*, *NRAS*, and *TERTp* mutations (35).

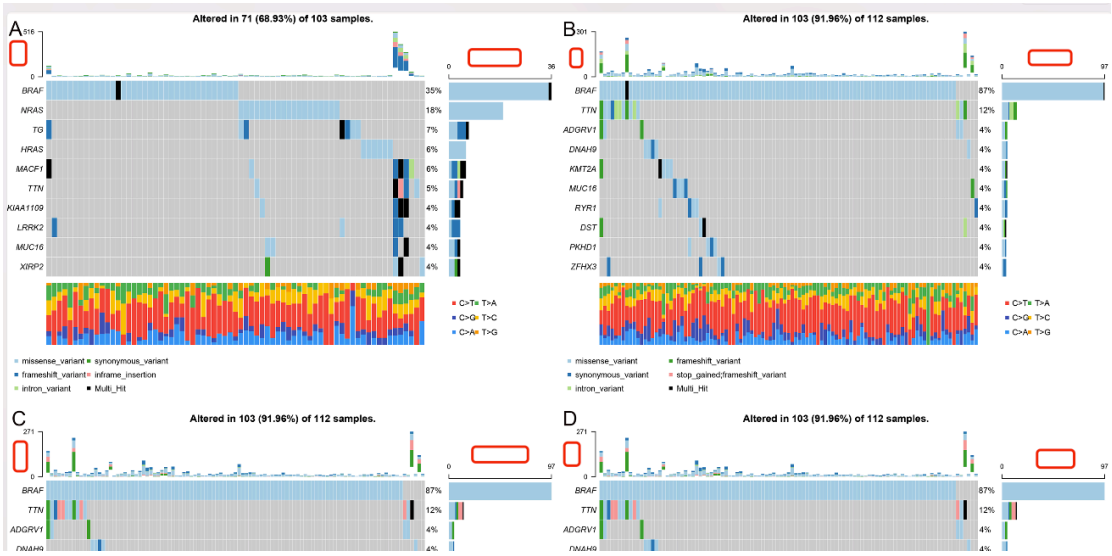
Response: Thank you for your suggestion. We revised the manuscript.

3. Figure 1: Please add units for x-axis in figure 1F.



Response: Thank you for reminding. We added units for x-axis in figure 1F.

4. Figure 3: Please supplement the descriptions of X- and Y-axis and resend us updated figure.



Response: Thank you for reminding. We updated figure 3.

5. Please check the legends of Figure 5D-5E, they are not Venn diagrams.

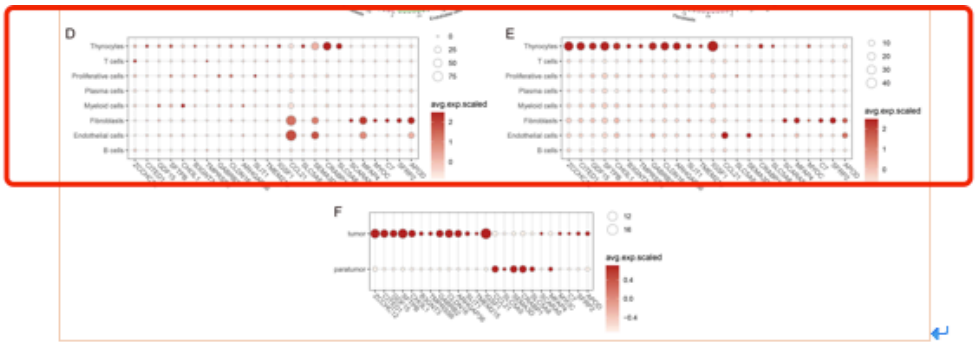


Figure 5 Verification of the characteristic genes in PTC using the scRNA-sequencing data set. (A) Comparison of various cell proportions in the PTC and normal samples. (B) Uniform Manifold Approximation and Projection (UMAP) of the normal samples. (C) UMAP of the PTC samples. (D) Venn diagram of the characteristic genes in the different cell types of the normal samples. (E) Venn diagram of the characteristic genes in the different cell types of the PTC samples. (F) Comparison of the characteristic gene

Response: Thank you for your suggestion. We revised the manuscript.

6. Table S1-S4: Please define ALL abbreviations shown in the tables in their corresponding table footnotes.

Response: Thank you for your suggestion. We confirmed all abbreviations shown in the tables in their corresponding table footnotes.