nature portfolio

Peer Review File

Digital profiling of gene expression from histology images with linearized attention

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author): Expert in cancer genomics, spatial transcriptomics, and bioinformatics

In this study, the authors proposed a transformer model named SEQUOIA which employs grouped vision attention for predicting gene expression levels from WSIs. The authors' approach to pretraining on normal tissue histology images and bulk RNA-seq data, followed by fine-tuning on cancer-specific datasets, is reported to outperform the HE2RNA method across various tumor datasets. They also showed that the predicted expression levels of certain genes can be used to predict the recurrence risk of breast cancer patients. SEQUOIA is then extended for spatial gene expression prediction using a slide-window approach, which is the only part in the entire manuscript that I feel interesting. Despite their claims, I have serious concerns about the method's practicality and effectiveness, for reasons that will be enumerated in my detailed review.

Major comments:
1. The study does not adequately address the fundamental issue of cell type composition variance in bulk RNA-seq data. Given that gene expression can be significantly influenced by the sampled cell population, the methodology must account for the disparity between cell composition in H&E slides and the corresponding RNA-seq analyzed tissue. The manuscript should provide a clear explanation of how SEQUOIA's predictions from H&E slides are representative of the actual gene expression in the varied tissue sections, considering the potential discrepancies arising from the way these slides are prepared.
2. The scenario of predicting gene expression from two H&E slides from the same tumor sample with differing cellular compositions raises serious concerns about SEQUOIA's consistency and reliability. If one slide is predominantly composed of normal cells and the other of tumor cells, the predictions would differ substantially, which undermines the model's applicability. If this issue is not addressed, it could lead to significant biases when applying SEQUOIA to new samples, limiting its practical use.
3. The manuscript falls short in providing a comprehensive comparative analysis. SEQUOIA is compared with HE2RNA, but not with tRNAformer, another transformer-based model mentioned in the Introduction. A systematic evaluation of SEQUOIA against all relevant tools is essential for validating its purported superiority and should be included in the Results.
4. The manuscript lacks clarity regarding the application of training and test sets within the model. It is imperative to specify whether the model was trained exclusively on the training dataset and then independently predicted on the test set. Such details are crucial for assessing the model's performance and should be clearly outlined.
5. The selection of data from only 9 cancer types for training and testing SEQUOIA is inadequately justified, especially when compared to HE2RNA's training on 28 cancer types. The rationale behind this selection must be provided, or a broader range of cancer types should be used to ensure a comprehensive evaluation of SEQUOIA's performance.
6. The assertion that pre-training on normal tissue data confers an advantage is not convincingly supported by the results. The model pretrained on a smaller dataset (1,802 slides) does not consistently outperform a model trained from scratch when fine-tuned on a larger dataset (4,331 TCGA slides). Only in 4 out of 9 tumor types does the pre-trained model exhibit superior performance. For a robust validation of the pre-training approach, the authors should compare the pre-trained model's

performance directly with a model trained on the entire TCGA dataset, which is larger and more representative of the actual cancer prediction task:

(a) Generally, pre-training data is expected to be much larger than fine-tuning data. However, in this model, the pre-training dataset consists of 1,802 slides, while the TCGA dataset for fine-tuning consists of 4,331 slides. The imbalance between the pre-training and fine-tuning datasets raises questions about the presumed benefits of pre-training. It is essential for the authors to demonstrate the efficacy of pre-training by comparing it with a model trained solely on the comprehensive TCGA set, particularly since the latter includes a greater number of samples.

(b) The application of Z-score normalization per gene within each tissue type during pre-training could introduce errors, potentially affecting model performance. For genes with tissue-specific expression patterns, this normalization may artificially inflate or reduce their signal in an unrepresentative manner. The authors should investigate and discuss the impact of this normalization technique on the model's predictive accuracy.

(c) The lack of differentiation between tissue types in the GTEx dataset used for pre-training contrasts with the tumor-type specificity applied during fine-tuning. The authors should explore whether pre-training on data from a single tissue type followed by fine-tuning on a single cancer type would affect the model’s performance, to provide a clearer understanding of the benefits and limitations of their pre-training strategy.

7. The results depicted in Fig. 2a-d are incomplete as they lack a control group for comparison. While pathways enriched with accurately predicted genes are highlighted, there is no mention of pathways with inaccurately predicted genes, which may also exhibit enrichment. The significance of accurately predicted genes, especially in the context of pathway or GO term analysis, hinges on the confirmation of predicted gene expression values. Without this validation, any biological interpretations drawn from these genes are speculative at best. The authors should elucidate the defining features of accurately versus inaccurately predicted genes and provide empirical verification of predicted gene expression values to substantiate any biological insights.

8. HE2RNA is also capable of predicting gene expression with spatial resolution. It is necessary for the authors to offer a direct comparison between SEQUOIA and HE2RNA in this aspect. Furthermore, the use of Earth Mover's Distance (EMD) as a metric for assessing differences in predicted gene expression is atypical. To aid readers in accurately evaluating the tool's performance, conventional metrics like Pearson Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE) should be employed alongside or instead of EMD.

9. Fig. 5 showcases SEQUOIA's ability to predict the expression of specific genes with spatial resolution, yet it remains unclear if this accuracy is consistent across different genes and slides. A statistical summary detailing the number of genes predicted with high accuracy across all spatial transcriptomic slides, alongside their proportion relative to the total number of predicted genes, would be highly informative. This would provide a clearer picture of the model's overall performance and reliability in spatial gene expression prediction.

10. Clarity is required regarding whether the TCGA slides used in the prediction of breast cancer recurrence were included in the training set of the SEQUOIA model. To ensure the model's generalizability and to avoid dataset bias, it would be prudent for the authors to construct the gene signature using a more diverse and publicly accessible dataset and then validate it using the TCGA dataset, reflecting a more realistic application scenario.

Minor comments:

1. In Fig. 3a, the authors present only the count of accurately predicted genes. To comprehensively assess the model's performance, it is imperative to include both PCC and RMSE for the predictions versus ground truth on an independent validation set, akin to the metrics provided in Fig. 1d & e.
2. here is inconsistency in the text case between Fig. 2a and Fig. 2c. To maintain uniformity, 'luad' in Fig. 2a should be amended to the capitalized form 'LUAD'.
3. The authors demonstrate SEQUOIA's superior prediction accuracy over HE2RNA using PCC and RMSE. However, a detailed analysis of the overlap and distinct preferences of well-predicted genes by both SEQUOIA and HE2RNA is necessary to contextualize the extent of improvement.
4. There is a discrepancy in Fig. 1b which cites '2,242 slides from 7 normal tissues'; this should be corrected to '1,802 slides from 6 normal tissues' for accuracy, as per the supplementary material. Additionally, the number of TCGA slides should be rectified to '4,331 slides' to reflect the data presented in Supplementary Table A1.
5. Technical difficulties were encountered when setting up and running the code from the provided Github repository:

(a) Incompatibilities exist within the specified library versions in the requirements.txt file. Despite successful installation of their latest versions which support execution, the file lacks necessary dependencies such as scikit-image, opencv-python, and py-lz4framed.
(b) The script at ./pre_processing/patch_gen_hdf5.py is missing a comma at line 150.
(c) The instructions for Step 5 mention importing both vit.py and vit_new.py, but only vit_new.py is available in the repository.

6. There appears to be a bibliographic error: References [22] and [26] are listed as separate entries but refer to the same article.

Reviewer #1 (Remarks on code availability):

Technical difficulties were encountered when setting up and running the code from the provided Github repository:

(a) Incompatibilities exist within the specified library versions in the requirements.txt file. Despite successful installation of their latest versions which support execution, the file lacks necessary dependencies such as scikit-image, opencv-python, and py-lz4framed.
(b) The script at ./pre_processing/patch_gen_hdf5.py is missing a comma at line 150.
(c) The instructions for Step 5 mention importing both vit.py and vit_new.py, but only vit_new.py is available in the repository.

Reviewer #2 (Remarks to the Author): Expert in cancer genomics and imaging, bioinformatics, machine learning, and digital pathology

Summary

The authors present SEQUOIA, a deep learning model intended for predicting RNA-Seq gene expression from whole-slide histology images, contributing towards personalized cancer care. SEQUOIA utilizes

pretraining on normal tissue data and incorporates an attention-based mechanism to potentially outperform existing approaches like ST-Net and HE2RNA in gene expression prediction. The model is evaluated across various cancer types, showcasing its capability to accurately interpret complex biological information crucial for personalized diagnosis and treatment. Additionally, SEQUOIA's performance is compared with other algorithms, addressing previous challenges such as the integration of contextual information between image tiles and the limitations of models trained on specific gene expressions or cancer types. In exploring spatial transcriptomics, SEQUOIA predicts locoregional gene expression, offering insights into tumor biology. The research also involves the development of a 50-gene signature aimed at predicting breast cancer recurrence, leveraging SEQUOIA's gene expression predictions to enhance disease classification and treatment strategies. This study highlights SEQUOIA's role in advancing the understanding of cancer pathology by potentially improving upon existing deep learning models and methodologies in the field.
While paper holds merits, there are multiple important parts (outlined below) that will require authors' attention.

Major Comments
1. While the authors show that their model is performing well across several cancer types, there is a lack of baseline comparisons as they only compare to HE2RNA model. Authors should compare their model's performance with several recent baselines such as tRNAsformer.
2. In the "A digital signature for breast cancer recurrence prediction" section, the authors should also compare their survival analysis results with the recent models that directly predict outcome from WSIs.
3. The application of the model for spatial transcriptomics is interesting. However, the selection of the window size can directly impact this analysis. The authors should also evaluate their model using different window sizes (the current window size is 10x10 tiles) and provide potential rationale for the behavior.
4. As shown in figure 5a, slides have variations in the staining. Are there any techniques being used to address this problem? If so, it should be explained and if not, the authors should at least comment on it within the text (however, performing color normalization experiments is more desirable).
5. Paragraph 4 of page 16 – using a simple averaging over the patch features residing within the same cluster losses the count and variation information of the features of the cluster. More intelligent mechanisms such as adding std alongside the mean or even using a learnable attention mechanism should also be tested as they can potentially improve the performance even further.
6. Are there any positional encodings included in the transformer architecture? Authors should evaluate their model with and without the positional encoding and comment on the performance as well as the rationale.

Minor Comments
1. Line 3, page 5 – while the authors refer to the performance as "accurately predicted", they have to provide a convincing quantitative metric to back up this claim within the text.
2. As a follow-up to the previous comment, authors refer to RSME values in the text. However, this metric is not easily interpretable. I would suggest that the authors either provide the possible range of the metric for the data, normalize it compared to the range, or at least report the RSME of the baseline in that case.
3. Paragraph 5, page 5 – in this paragraph, the authors provide a list of pathways enriched by the

correctly predicted genes. However, I would recommend describing how these pathways are related to the disease based on the known knowledge. This improves the story-telling flow of this section.
4. Are there any mechanisms to measure the uncertainty of the predictions for each gene? If it's possible for the model to provide the confidence of predictions, the author should clarify the mechanism. Otherwise, it would be good to add it to the text as a potential future direction.
5. Paragraph 5 of page 16 – the part the authors talk about ViT is unrelated to the model or the context. I recommend either removing this part or describing it within the context of the original Transformer architecture.

Reviewer #3 (Remarks to the Author): Expert in cancer genomics and imaging, bioinformatics, machine learning, and digital pathology

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #1 (Remarks to the Author): Expert in cancer genomics, spatial transcriptomics, and bioinformatics

In this study, the authors proposed a transformer model named SEQUOIA which employs grouped vision attention for predicting gene expression levels from WSIs. The authors' approach to pretraining on normal tissue histology images and bulk RNA-seq data, followed by fine-tuning on cancer-specific datasets, is reported to outperform the HE2RNA method across various tumor datasets. They also showed that the predicted expression levels of certain genes can be used to predict the recurrence risk of breast cancer patients. SEQUOIA is then extended for spatial gene expression prediction using a slide-window approach, which is the only part in the entire manuscript that I feel interesting. Despite their claims, I have serious concerns about the method's practicality and effectiveness, for reasons that will be enumerated in my detailed review.

We appreciate the time and effort invested by the reviewers in evaluating our manuscript. Please see our point-to-point response below.

Major comments:

1. The study does not adequately address the fundamental issue of cell type composition variance in bulk RNA-seq data. Given that gene expression can be significantly influenced by the sampled cell population, the methodology must account for the disparity between cell composition in H&E slides and the corresponding RNA-seq analyzed tissue. The manuscript should provide a clear explanation of how SEQUOIA's predictions from H&E slides are representative of the actual gene expression in the varied tissue sections, considering the potential discrepancies arising from the way these slides are prepared.

We thank the reviewer for their comment. While there could be potential differences in cell-type composition between H&E slides and RNA-seq analyzed tissues, published studies have shown that the RNA-seq data can accurately reflect tissue composition as seen in matched H&E slides. In our previous work (PMID: 37433817, Fig.2g-i), we demonstrated that the proportions of transcriptional phenotypes of tumor cells inferred from H&E slides are significantly correlated with those estimated from computational deconvolution of bulk RNA-seq data in the TCGA glioblastoma cohort. Similarly, recent work (PMID: 33712588) from other colleagues has also revealed a significant correlation between the cell-type proportions captured from H&E slides and those estimated from bulk RNA-seq across five evaluated TCGA cancer types.

In the current study, we have taken several measures to address tissue-level heterogeneity. Firstly, we selected a large number of tiles (*N* = 4,000) from each slide to generate slide-level feature representation, which enabled us to capture the tissue-level heterogeneity (Methods: Preprocessing of Whole Slide Images). Secondly, we employed a transformer architecture, leveraging attention-based mechanisms to determine which image features were most relevant for gene expression signals obtained from the matched RNA-seq samples (Methods: SEQUOIA architecture). During the training process, image features strongly correlated with the gene expression signals will be assigned with a higher weight than those not strongly related. Thirdly, our model was trained on a substantial number of tissue slides and matched RNA-seq samples (Supplementary Table A1), enabling it to capture relevant histological features associated with gene expression signals. Lastly, we validated our model across independent datasets (Figure 3 and Supplementary Table A2). The results from these measures collectively indicate that our models learnt the relevant histological features corresponding to the gene expression levels.

2. The scenario of predicting gene expression from two H&E slides from the same tumor sample with differing cellular compositions raises serious concerns about SEQUOIA's consistency and reliability. If one slide is predominantly composed of normal cells and the other of tumor cells, the predictions would differ substantially, which undermines the model's applicability. If this issue is not addressed, it could lead to significant biases when applying SEQUOIA to new samples, limiting its practical use.

As described in the Methods section (“Patient cohorts and ethics”), our models were trained exclusively on H&E slides of tumor tissues, while the adjacent normal tissues were excluded. It is important to note that for the majority of cancer types (except for GBM), there is only one available H&E slide of tumor tissues per patient (Supplementary Table A1). Only in rare cases (less than 10%) where two diagnostic slides were available for a patient , we included both slides in the training set for better capture of the tissue heterogeneity. While there could be discrepancies in cell-type compositions between the two slides, our transformer-based model leverages self-attention mechanisms to determine which image features were most relevant to the gene expression signals. By training our models on tissues from a large number of patients and validating them in independent cohorts, we demonstrated that the predicted gene expression signals reflected underlying biological processes across patients rather than discrepancies between different slides from the same patient.

3. The manuscript falls short in providing a comprehensive comparative analysis. SEQUOIA is compared with HE2RNA, but not with tRNAformer, another transformer-based model mentioned in the Introduction. A systematic evaluation of SEQUOIA against all relevant tools is essential for validating its purported superiority and should be included in the Results.

We appreciate the reviewer for this valuable suggestion. In the revised manuscript, we have included a benchmark with the tRNAsformer model. As shown in Figures 1d-e, SEQUOIA demonstrated superior performance in 8 out of 9 evaluated cancer types regarding the number of significantly well-predicted genes as determined by Pearson correlation analysis and RMSE values. In contrast, tRNAsformer only showed comparable performance with SEQUOIA in breast cancer, where the highest number of training samples were available. The Pearson correlation coefficients obtained from SEQUOIA were significantly higher compared to tRNAsformer in 7 out of 9 evaluated cancer types (Figure 1d and Supplementary Figure A2b, Mann-Whitney U test, *P* < 2E-09). Regarding the RMSE values, SEQUOIA outperformed tRNAsformer in all evaluated cancer types with statistically lower RMSEs (Figure 1e and Supplementary Figure A2c).

4. The manuscript lacks clarity regarding the application of training and test sets within the model. It is imperative to specify whether the model was trained exclusively on the training dataset and then independently predicted on the test set. Such details are crucial for assessing the model's performance and should be clearly outlined.

We thank the reviewer for this comment. In the revised manuscript, we have added clarification that the model was trained exclusively on the training dataset and then independently predicted on the test set (Methods: Training and evaluation on the TCGA dataset), and we added Supplementary Figure A1 to illustrate our approach.

5. The selection of data from only 9 cancer types for training and testing SEQUOIA is inadequately justified, especially when compared to HE2RNA's training on 28 cancer types. The rationale behind this selection must be provided, or a broader range of cancer types should be used to ensure a comprehensive evaluation of SEQUOIA's performance.

Assessing the generalization capacity of our deep-learning models trained on the TCGA dataset to independent data cohorts is a critical component of our experimental design. However, there is a lack of existing data resources that contain both H&E images and matched RNA-seq data. To validate our models across independent data cohorts, we focused on seven cancer types that are common to both TCGA and CPTAC datasets. In addition, we also considered two additional cancer types (i.e., PRAD and KIRP) from TCGA due to their prevalence. In the revised manuscript, we have added clarification of this rationale into the Methods section ("Patients cohorts and ethics").

6. The assertion that pre-training on normal tissue data confers an advantage is not convincingly supported by the results. The model pretrained on a smaller dataset (1,802 slides) does not consistently outperform a model trained from scratch when fine-tuned on a larger dataset (4,331 TCGA slides). Only in 4 out of 9 tumor types does the pre-trained model exhibit superior performance. For a robust validation of the pre-training approach, the authors should compare the pre-trained model's performance

directly with a model trained on the entire TCGA dataset, which is larger and more representative of the actual cancer prediction task.

We appreciate the critique from this reviewer. We agree that pretraining the model on normal tissues improved the performance only in specific cancer types. However, it is worth mentioning that the improvement was most significant on pancreatic cancer (PAAD), the cancer type with the smallest number of training samples available (n = 202 slides). It is expected that, in cancer types where the training cohort is large (e.g. BRCA), the effect from pre-training will be reduced. While only in 4 out of 9 cancer types did we observe an increase when evaluating the performance on individual genes, 6 out of 9 (66.7%) cancer types showed an improvement in the correlation coefficient between the predicted pathway activation levels and the ground truth when evaluating the performance at the pathway level (Figure 2e and page 6 the last paragraph). To optimize the effect from pretraining, we tested several strategies suggested by this reviewer.

(a) Generally, pre-training data is expected to be much larger than fine-tuning data. However, in this model, the pre-training dataset consists of 1,802 slides, while the TCGA dataset for fine-tuning consists of 4,331 slides. The imbalance between the pre-training and fine-tuning datasets raises questions about the presumed benefits of pre-training. It is essential for the authors to demonstrate the efficacy of pre-training by comparing it with a model trained solely on the comprehensive TCGA set, particularly since the latter includes a greater number of samples.

We appreciate the comments from this reviewer. It is important to note that we pre-trained the model on the GTex dataset combining all tissue types (n = 1,870 slides), and then finetuned the model on each specific cancer type. The number of slides from each specific cancer type is much smaller than the pre-training dataset (Supplementary Table A1, except for BRCA). This approach follows the logic suggested by this reviewer.

We agree with this reviewer that pretraining the model on the entire TCGA dataset may offer additional benefits, and we have tested this strategy in our pilot study. However, pretraining on the entire TCGA dataset resulted in a worse performance than simply training it from scratch. This is likely due to the enormous cell morphological and histological differences across cancers from different tissue origins. By pretraining the model on data from different cancer tissues, the model simply learns the tissue-type labels, instead of “cancer-type-specific” gene expression signals. Based on results from our pilot study, we eventually abandoned this strategy.

(b) The application of Z-score normalization per gene within each tissue type during pre-training could introduce errors, potentially affecting model performance. For genes with tissue-specific expression patterns, this normalization may artificially inflate or reduce their signal in an unrepresentative manner. The authors should investigate and discuss the impact of this normalization technique on the model's predictive accuracy.

In response to this reviewer's suggestion, we have experimented with pretraining our model on gene expression data without z-score normalization. As shown in Supplementary Figure A2e, incorporating z-score normalization improved the prediction performance in 8 out of 9 evaluated cancer types regarding the number of well-predicted genes and Pearson correlation coefficient. Exception was found for COAD, where the model achieved a comparable performance than the non-z-score-normalized model. Therefore, we incorporated z-score normalization to the pretraining data.

(c) The lack of differentiation between tissue types in the GTEx dataset used for pre-training contrasts with the tumor-type specificity applied during fine-tuning. The authors should explore whether pre-training on data from a single tissue type followed by fine-tuning on a single cancer type would affect the model's performance, to provide a clearer understanding of the benefits and limitations of their pre-training strategy.

In response to this suggestion, we experimented with pretraining the model on each specific tissue type separately, and then finetuned it on the corresponding cancer type. As shown in Supplementary Figure A2e, this strategy did not offer additional benefits compared to our original approach. While pretraining the model on tissue-specific data significantly increased the correlation coefficient in two cancer types (BRCA and KIRP), it fell short in all remaining cancer types. In contrast, pretraining the model on combined data from all available tissue types significantly increased the correlation coefficients in 6 out of 9 cancer types than the tissue-specific model.

7. The results depicted in Fig. 2a-d are incomplete as they lack a control group for comparison. While pathways enriched with accurately predicted genes are highlighted, there is no mention of pathways with inaccurately predicted genes, which may also exhibit enrichment. The significance of accurately predicted genes, especially in the context of pathway or GO term analysis, hinges on the confirmation of predicted gene expression values. Without this validation, any biological interpretations drawn from these genes are speculative at best. The authors should elucidate the defining features of accurately versus inaccurately predicted genes and provide empirical verification of predicted gene expression values to substantiate any biological insights.

In response to this comment, we added results from gene-set enrichment analysis for the inaccurately genes in the revised manuscript (Supplementary Figures A3d-e-f). The results are twofold. First, the gene sets enriched with accurately predicted genes were *not* enriched with the inaccurately predicted genes (i.e., genes that did not pass our significant thresholds), indicating the presented gene sets in Figures 2a-d were specific to the accurately predicted genes. Second, the gene sets enriched with the inaccurately predicted genes (e.g. synaptic transmission, bile secretion) were not strictly related to cancers or were not interpretable in the context of the disease (Supplementary Figures A3f). These results indicate that the well-predicted genes from SEQUOIA were primarily and specifically related to the regulation of cancer development and progression. It is important to note that the "accurately predicted" genes refer to those that have passed our statistical thresholds (Page 3, paragraph 3), thus their predicted expression values have been statistically validated by both Pearson correlation and RMSE analyses. In addition, we have also performed single-sample gene set enrichment analysis using the predicted gene expression values of each sample and compared the predicted pathway activation levels with the groundtruth (Figure 2e-f). These results validated the predicted gene expression values at the pathway level.

8. HE2RNA is also capable of predicting gene expression with spatial resolution. It is necessary for the authors to offer a direct comparison between SEQUOIA and HE2RNA in this aspect. Furthermore, the use of Earth Mover's Distance (EMD) as a metric for assessing differences in predicted gene expression is atypical. To aid readers in accurately evaluating the tool's performance, conventional metrics like Pearson Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE) should be employed alongside or instead of EMD.

We appreciate this valuable suggestion from the reviewer. In the revised manuscript, we have added a benchmark of SEQUOIA with HE2RNA. In addition, we have also added Pearson correlation coefficient (PCC) as an additional evaluation matrix. As shown in Supplementary Figure A5, SEQUOIA outperformed HE2RNA in 15 out of 18 evaluated slides in terms of PCC, and in 10 out of 18 slides in terms of EMD values.

9. Fig. 5 showcases SEQUOIA's ability to predict the expression of specific genes with spatial resolution, yet it remains unclear if this accuracy is consistent across different genes and slides. A statistical summary detailing the number of genes predicted with high accuracy across all spatial transcriptomic slides, alongside their proportion relative to the total number of predicted genes, would be highly informative. This would provide a clearer picture of the model's overall performance and reliability in spatial gene expression prediction.

In the revised manuscript, we have added a statistical summary for the number of genes that can be well-predicted across all spatial transcriptomics slides as Supplementary Data 7 and Supplementary Data 8.

10. Clarity is required regarding whether the TCGA slides used in the prediction of breast cancer recurrence were included in the training set of the SEQUOIA model. To ensure the model's generalizability and to avoid dataset bias, it would be prudent for the authors to construct the gene signature using a more diverse and publicly accessible dataset and then validate it using the TCGA dataset, reflecting a more realistic application scenario.

In the revised manuscript, we have added clarification in the Results section (Page 11, the last paragraph) that the TCGA slides used for predicting breast cancer recurrence were from the test sets, which prevented any leakage of information. To demonstrate the generalizability of our gene expression signature, we validated it on both the SCANB cohort (n = 5,034 patients) and the METABRIC cohort (n = 2,262 patients). The results from these validations on a total of 7,296 patients indicate that the signature can be generalized to diverse and publicly accessible datasets (Figures 4b-c).

Minor comments:

1. In Fig. 3a, the authors present only the count of accurately predicted genes. To comprehensively assess the model's performance, it is imperative to include both PCC and RMSE for the predictions versus ground truth on an independent validation set, akin to the metrics provided in Fig. 1d & e.

In response to this reviewer's comment, we have added PCC and RMSE for the predictions versus ground truth on the independent validation sets. The results are now presented in Figure 3a and Supplementary Figure A4 of the revised manuscript.

2. here is inconsistency in the text case between Fig. 2a and Fig. 2c. To maintain uniformity, 'luad' in Fig. 2a should be amended to the capitalized form 'LUAD'.

We appreciate this reviewer for the careful reading of our manuscript, we have corrected the text labels in our revised manuscript.

3. The authors demonstrate SEQUOIA's superior prediction accuracy over HE2RNA using PCC and RMSE. However, a detailed analysis of the overlap and distinct preferences of well-predicted genes by both SEQUOIA and HE2RNA is necessary to contextualize the extent of improvement.

We thank the reviewer for this suggestion. In our revised manuscript, we have added a Venn diagram as Supplementary Figure A2d, which illustrates the overlapped and distinct genes well-predicted across different models (i.e., SEQUOIA, HE2RNA, tRNAsformer). Notably, approximately 60% to 76% of the genes well-predicted by HE2RNA and tRNAsformer models were also well-predicted by SEQUOIA. In addition, SEQUOIA extended the well-predicted gene sets that were not captured by HE2RNA and tRNAsformer. Exceptions were found for GBM and PAAD, where HE2RNA and tRNAsformer predicted a distinct set of genes compared to SEQUOIA. This discrepancy is expected, as only a few genes predicted by HE2RNA and tRNAsformer in these two cancer types can be validated in the independent cancer cohorts (see Figure 3b). These results underscore the superior performance of SEQUOIA compared to existing architectures.

4. There is a discrepancy in Fig. 1b which cites '2,242 slides from 7 normal tissues'; this should be corrected to '1,802 slides from 6 normal tissues' for accuracy, as per the supplementary material. Additionally, the number of TCGA slides should be rectified to '4,331 slides' to reflect the data presented in Supplementary Table A1.

We appreciate this reviewer for the careful reading of our manuscript. We have corrected the slide numbers in the revised manuscript to align them between Fig.1b and Supplementary Tables.

6. There appears to be a bibliographic error: References [22] and [26] are listed as separate entries but refer to the same article.

We apologize for the duplication due to two different versions of the reference. In the revised manuscript, we have removed the duplicated entries.

Reviewer #1 (Remarks on code availability):

Technical difficulties were encountered when setting up and running the code from the provided Github repository:
(a) Incompatibilities exist within the specified library versions in the requirements.txt file. Despite successful installation of their latest versions which support execution, the file lacks necessary dependencies such as scikit-image, opencv-python, and py-lz4framed.
(b) The script at ./pre_processing/patch_gen_hdf5.py is missing a comma at line 150.
(c) The instructions for Step 5 mention importing both vit.py and vit_new.py, but only vit_new.py is available in the repository.

We thank the reviewer for these comments. We have now updated the project requirements to fix incompatibilities and to include all dependencies. We have also renamed the legacy files, and structured the code in different project folders for better readability and improved structure.

Reviewer #2 (Remarks to the Author): Expert in cancer genomics and imaging, bioinformatics, machine learning, and digital pathology

Summary

The authors present SEQUOIA, a deep learning model intended for predicting RNA-Seq gene expression from whole-slide histology images, contributing towards personalized cancer care. SEQUOIA utilizes pretraining on normal tissue data and incorporates an attention-based mechanism to potentially outperform existing approaches like ST-Net and HE2RNA in gene expression prediction. The model is evaluated across various cancer types, showcasing its capability to accurately interpret complex biological information crucial for personalized diagnosis and treatment. Additionally, SEQUOIA's performance is compared with other algorithms, addressing previous challenges such as the integration of contextual information between image tiles and the limitations of models trained on specific gene expressions or cancer types. In exploring spatial transcriptomics, SEQUOIA predicts locoregional gene expression, offering insights into tumor biology. The research also involves the development of a 50-gene signature aimed at predicting breast cancer recurrence, leveraging SEQUOIA's gene expression predictions to enhance disease classification and treatment strategies. This study highlights SEQUOIA's role in advancing the understanding of cancer pathology by potentially improving upon existing deep learning models and methodologies in the field. While paper holds merits, there are multiple important parts (outlined below) that will require authors' attention.

We appreciate the reviewer's effort in evaluating our manuscript. Please find our poin-to-point response below.

Major Comments

1. While the authors show that their model is performing well across several cancer types, there is a lack of baseline comparisons as they only compare to HE2RNA model. Authors should compare their model's performance with several recent baselines such as tRNAsformer.

We appreciate the reviewer for this valuable suggestion. In the revised manuscript, we have included a benchmark with the tRNAsformer model. As shown in Figures 1d-e, the SEQUOIA model demonstrated superior performance in 8 out of 9 evaluated cancer types regarding the number of well-predicted genes as determined by Peason correlation analysis and RMSE values. In contrast, tRNAsformer only showed comparable performance with SEQUOIA in breast cancer, where the highest number of training samples were available. The correlation coefficients obtained from SEQUOIA were significantly higher compared to tRNAsformer in 7 out of the 9 evaluated cancer types (Figure 1d and Supplementary Figure A2b, Mann-Whitney U test, *P* < 2E-09). Regarding the RMSE values, SEQUOIA outperformed tRNAsformer in all evaluated cancer types with statistically lower RMSEs (Figure 1e and Supplementary Figure A2c).

2. In the “A digital signature for breast cancer recurrence prediction” section, the authors should also compare their survival analysis results with the recent models that directly predict outcome from WSIs.

We appreciate the reviewer for this valuable suggestion. Following this suggestion, we trained a separate model based on SEQUOIA’s architecture for predicting recurrence-free survival directly from WSIs (Methods: Recurrence-free survival prediction from histology images). For training and evaluating the model, we kept the input features consistent with the model trained for gene expression prediction. The model was trained to predict a risk score for recurrence using Cox loss as detailed in our published studies (PMID: 37433817 and PMID: 36991216). As shown in Figure 4f, patients assigned with high risk scores had a worse prognosis compared to those with low risk scores, but this association failed to reach statistical significance (Cox regression: HR = 1.01, .95CI = 0.995-1.02, *P* = 0.215; Log-rank test: *P* = 0.065). This result highlights the effectiveness of using our gene expression-based signature for patient stratification.

Furthermore, our gene-expression-based method offers additional benefits for interpretability compared to the model trained directly from histology images. As demonstrated in our gene set analysis (Figure 4d), the gene expression signature contains a set of genes regulating cell apoptosis, migration, and metabolism. The predicted expression levels of these genes may offer mechanistic insights into the clinical outcomes of a patient. However, a model trained directly from histology images will lack such mechanistic information.

3. The application of the model for spatial transcriptomics is interesting. However, the selection of the window size can directly impact this analysis. The authors should also evaluate their model using different window sizes (the current window size is 10x10 tiles) and provide potential rationale for the behavior.

We agree with the reviewer that the selected window size can directly impact results from spatial analysis. However, the rationale for our selected window size (10x10 tiles) stems from the model architecture: the model requires 100 feature vectors as input (Methods: SEQUOIA architecture). Different window sizes would hence be sub-optimal: choosing a smaller window size would necessitate introducing ‘empty’ feature vectors. However, the model was never trained to make predictions in case some feature vectors are left empty, so the prediction behavior in this case would be unreliable. A larger window size of >10x10 tiles would result in the loss of spatial granularity for the prediction, since gene expression measured at a higher number of spots would have to be aggregated in this manner. In the revised manuscript, we have added this rationale into the Methods section ("Spatial gene expression prediction at tile level").

4. As shown in figure 5a, slides have variations in the staining. Are there any techniques being used to address this problem? If so, it should be explained and if not, the authors should at least comment on it within the text (however, performing color normalization experiments is more desirable).

We thank the reviewer for this comment. We did not include staining normalization into our procedures. Since our model was trained on a large number of tissue slides and also validated in independent data cohorts, the model is expected to be robust to handle stain variation. In the revised manuscript, we have added exploring the potential benefits of color normalization as a potential future effort (Discussion, page 15 the last paragraph).

5. Paragraph 4 of page 16 – using a simple averaging over the patch features residing within the same cluster losses the count and variation information of the features of the cluster. More intelligent mechanisms such as adding std alongside the mean or even using a learnable attention mechanism should also be tested as they can potentially improve the performance even further.

We thank the reviewer for this suggestion. While we agree that using a simple averaging can remove some information provided by the individual patches, it can also omit noise that might be incorporated within the cluster. While the learnable attention mechanism has been proposed in recent digital pathology literature, it would be infeasible to be introduced into our architecture. It would imply introducing attention layers for each of the 100 clusters (each one composed of a maximum of 4,000 feature vectors), exponentially increasing the memory requirements and parameters of the model. This would make training the model infeasible in a reasonable amount of time. Furthermore, it must be noted that in benchmarking works of multiple aggregation methods in digital pathology, the average has been one of the best performant methods over attention mechanisms (see Table 2 https://www.sciencedirect.com/science/article/pii/S1361841523001457). For these reasons, we have decided to use an average of the features in our work.

6. Are there any positional encodings included in the transformer architecture? Authors should evaluate their model with and without the positional encoding and comment on the performance as well as the rationale.

We thank the reviewer for this question. Since our data were preprocessed in a way that does not have an explicit positional restriction, we did not include positional encoding in our transformer architecture. While transformers for natural language processing require positional embeddings since they introduce the structure that language has, this is not the case in digital pathology slides. Furthermore, since tumors are heterogeneous and patches from distant regions in the slide can have similar characteristics (which is why we clustered them), we reckon that imposing any kind of positional restriction would damage the model performance, especially because this restriction does not exist.

Minor Comments

1. Line 3, page 5 – while the authors refer to the performance as “accurately predicted”, they have to provide a convincing quantitative metric to back up this claim within the text.

Thanks for the detailed reading of our manuscript. We have now added clarification in the corresponding section. Our quantitative metric combines both Pearson correlation analysis and root mean square error (RMSE) to identify genes that can be significantly well predicted for their expression values, which was detailed in the Results section (Page 3, paragraph 3) and the Methods.

2. As a follow-up to the previous comment, authors refer to RSME values in the text. However, this metric is not easily interpretable. I would suggest that the authors either provide the possible range of the metric for the data, normalize it compared to the range, or at least report the RSME of the baseline in that case.

We appreciate the reviewer for this valuable critique. In the revised manuscript, we have derived a “normalized RMSE value” as it is now presented in Figure 1e and Figure 2f. This value was calculated using a two-step method (Page 19, first paragraph). First, since the absolute gene expression values varied across different genes, we performed quantile normalization of the RMSE values. For each gene, the RMSE value between the prediction and ground truth was divided by the interquartile range of its absolute expression values across all patients. This normalization ensures that the RMSE values were comparable between different genes. Second, we performed min-max normalization for the quantile-normalized RMSE values across all genes calculated in each specific cancer type. This step scaled the quantile-normalized RMSE values to a range between 0 and 1, where smaller values indicate a closer correspondence between ground truth and the predicted values.

3. Paragraph 5, page 5 – in this paragraph, the authors provide a list of pathways enriched by the correctly predicted genes. However, I would recommend describing how these pathways are related to the disease based on the known knowledge. This improves the story-telling flow of this section.

We thank the reviewer for this valuable comment. In the revised manuscript, we have added descriptions of how the presented pathways were related to the disease (Page 6 Paragraph 2). The corresponding text reads: "Published studies have revealed the critical roles of glioblastoma-associated macrophages in promoting the malignancy of tumor cells by secreting pro-tumorigenic cytokines that enhance immunologic tolerance and increase epithelial-mesenchymal transition. Similarly, variations in the activation levels of the endothelial cell development pathway may reflect the difference in hypoxia levels across tumors, as hypoxia is a known factor that drives the clonal evolution of glioblastoma cells…These findings agree with the essential roles of collagen remodeling, immune-cell regulation and epigenetic alterations in affecting lung cancer development and progression."

4. Are there any mechanisms to measure the uncertainty of the predictions for each gene? If it's possible for the model to provide the confidence of predictions, the author should clarify the mechanism. Otherwise, it would be good to add it to the text as a potential future direction.

To assess the uncertainty of predictions, we can utilize ensemble methods, which involves training multiple instances of our models with different initialized parameters or different training/test dataset splits. We can then calculate the variance or standard deviation for the predictions across the test results for each gene from different models. While measuring the uncertainty of predictions can help assess the reliability of predictions, in this study we have incorporated robust statistical measurements into the evaluation of our models and also validated them on independent datasets. The results have confirmed their reliability. In the revised manuscript, we have added measuring the uncertainty of prediction as a potential future direction (Discussion, page 15 the last paragraph).

5. Paragraph 5 of page 16 – the part the authors talk about ViT is unrelated to the model or the context. I recommend either removing this part or describing it within the context of the original Transformer architecture.

We appreciate the comment from this reviewer. We have shortened this paragraph to focus on the relevant context (Methods: SEQUOIA architecture).

Reviewer #3 (Remarks to the Author): Expert in cancer genomics and imaging, bioinformatics, machine learning, and digital pathology

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We appreciate the time and effort invested by this reviewer in evaluating our manuscript. Please see our point-to-point response above.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Overall, I commend the authors for their thoughtful responses to all the concerns raised and their systematic comparison of different methods and strategies. However, I still have concerns about the actual performance of SEQUOIA in predicting the spatial distribution of gene expression. Based on the results provided in the revised manuscript, SEQUOIA's performance in this aspect appears to be poor, which significantly impacts the tool's effectiveness and value. Specific comments are as follows.

1. The authors mentioned that the model pre-trained on the entire TCGA dataset performs worse than a model trained from scratch but did not provide any results or data to support this claim. The authors should present the relevant results in their response.

2. In the GBM spatial transcriptomics dataset, which consists of 18 slices, SEQUOIA did not predict any gene with a PCC greater than 0.1 across all slices, and only one gene had a PCC greater than 0.1 in 10 slices. Although SEQUOIA performed slightly better than HE2RNA in the comparison, this result still indicates that SEQUOIA is largely ineffective in predicting the spatial expression of genes, making it difficult to apply the tool accurately to real slices.

4. Related to the previous point, the authors chose GBM as the tumor type for their study. However, based on Figures 1d-e and 3a, GBM does not appear to be a tumor type with good prediction results. I recommend the authors study a tumor type with better evaluation results, such as BRCA, and discuss whether the prediction performance in spatial gene expression is related to the model evaluation performance.

3. On Page 3, in the last paragraph, the authors claimed that the number of accurately predicted genes is related to the sample size, but the evidence from comparisons across different cancer types is indirect. I suggest the authors conduct downsampling of the sample size within the same cancer type to obtain direct evidence of the relationship between the number of accurately predicted genes and the sample size.

5. In Figure 5b, when showing example genes, I suggest that the authors annotate the PCC or EMD values in the figure to help readers understand the actual performance of the tool.

6. The authors developed a web application to display the model's predictions on TCGA data. However, the current web page can only read and display slices from the TCGA database sequentially. I suggest the authors add a feature to select or search for slice IDs, which would make the website more user-friendly.

Reviewer #1 (Remarks on code availability):

The revised code is now operational, but users need to install `openslide` (>v3.4.0) in addition to the dependencies listed in `requirements.txt`. This requirement should be specified in the GitHub repository.

Reviewer #2 (Remarks to the Author):

I would like to thank the authors in addressing this reviewer's comments.
There are two additional suggestions that this reviewer believes will further strengthen the study.

1- It is understandable that adding learnable attention mechanism will increase the complexity of the model (#of parameters, memory requirements). However, a simple mean, although can remove noise, can still obscure variation in features. I would recommend adding std alongside the mean as this strategy will not add too much complexity and may in fact improve results.

2- With respect to reviewer #1 comment related to the number of datasets utilized in this study, it is true that there is lack of existing data resources that contain both H&E images and matched RNA-seq data. However, TCGA contains other large cohorts that could be utilized in this study. I understand that there won't be independent datasets to validate findings but a cross-validation strategy is something to consider. The reason that I suggest the above analysis is that I see a lot of merit in this study and would like to see how the model generalizes to other cancers. It is perhaps not possible to perform the suggested analysis on all TCGA cohorts, but the authors can select the cohorts in TCGA with large # of cases to do this extra analysis.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We appreciate the time and effort that the reviewers have invested in evaluating our manuscript. We are grateful for their suggestions which we believe have enabled us to greatly improve our work. Below we provide a detailed response to each of the reviewers' questions.

In addition, driven by the reviewers' comments regarding performance and pre-training strategies, we have included two improved components into the model which have independently caused a significant boost in performance compared to our previous submission, both in terms of the bulk-level prediction as well as the spatial prediction performance:

1) Instead of using a ResNet pre-trained on ImageNet as tile feature extractor (which does not contain medical images), we included UNI [1], a recently released state-of-the art foundation model that was trained using more than 100 million images from over 100,000 diagnostic H&E-stained WSIs (>77 TB of data) across 20 major tissue types.
2) Next to trying to solve the added computational complexity of the self-attention weights in the transformer with pre-training strategies, we tackled the problem at the root and implemented a linear alternative to this self-attention component. This allowed us to model contextual relations across tiles at linear instead of quadratic complexity, thereby reducing overfitting and improving performance.

To benchmark the added value of both components, we performed a thorough ablation study across all cancer types. In each experiment, we kept the pre-processing steps constant (in the same way as we did before), and we benchmarked the following combinations (see Figure 1):

1) ResNet features + MLP aggregation (as implemented in HE2RNA)
2) ResNet features + transformer aggregation (our previous SEQUOIA model)
3) ResNet features + linearized transformer aggregation
4) UNI features + MLP aggregation
5) UNI features + transformer aggregation
6) **UNI features + linearized transformer aggregation: new SEQUOIA model**

Since tRNAsformer is comparable to setting (2) in terms of feature extraction (CNN pre-trained on ImageNet) and tile aggregation (transformer encoder), we did not include a separate benchmarking for their exact implementation anymore into our results, but we included the model into the description in the intro and discussion.

[1] Chen, R.J., Ding, T., Lu, M.Y. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat Med* 30, 850–862 (2024). https://doi.org/10.1038/s41591-024-02857-3

Reviewer #1 (Remarks to the Author):

Overall, I commend the authors for their thoughtful responses to all the concerns raised and their systematic comparison of different methods and strategies. However, I still have concerns about the actual performance of SEQUOIA in predicting the spatial distribution of gene expression. Based on the results provided in the revised manuscript, SEQUOIA's performance in this aspect appears to be poor, which significantly impacts the tool's effectiveness and value. Specific comments are as follows.

1. The authors mentioned that the model pre-trained on the entire TCGA dataset performs worse than a model trained from scratch but did not provide any results or data to support this claim. The authors should present the relevant results in their response.

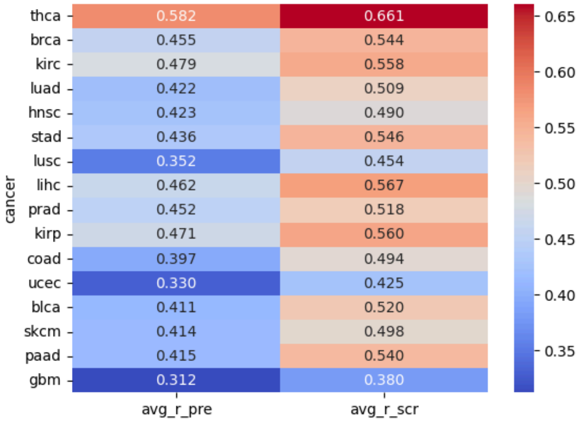
We apologize for the insufficient motivation we provided to omit the specific results from the response. We examined several variations of model architectures, training schemes and hyperparameters during model development. Due to the large computational time required to run the SEQUOIA model (±12h per cancer type on average = 108h for 9 cancer types we had at the time), it was infeasible to check each variation in model architecture/hyperparameter setting with several different pre-training strategies.

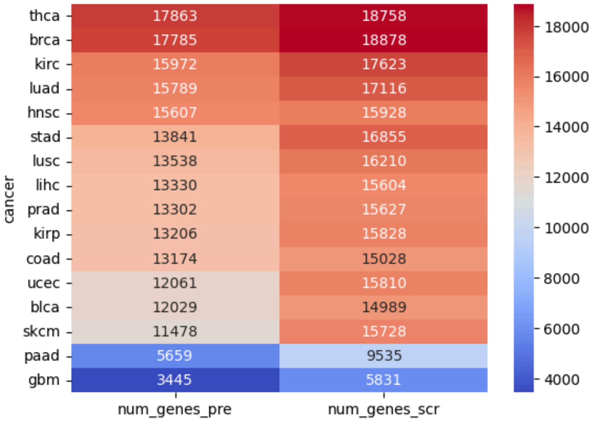
Although we can parallelize some of the runs on our GPU cluster, runtime is still a significant bottleneck. Since we tested pre-training strategies in early development stages with different architectures, those results were not directly comparable to the ones we presented in the paper. Due to the computational time required for model training, we could not re-run those experiments in time.

This time, we still did not include pre-training on TCGA because we prioritized to run the experiments 1,2,4 below and because of our insights from point 3:

1) We included 8 more cancer types, resulting in 96h additional computational runtime for each tested model combination. This means in total 192h runtime per model across the 16 cancer types.
2) We ran four new benchmarking combinations (points 3-4-5-6 from page 1)
3) Although we included the UNI tile feature extractor, which contains enormous pre-training, we still verified whether pre-training on GTex could further boost the performance of the new model.
   However, we found this made the prediction slightly worse, see below for comparison of the average Pearson correlation coefficients after fine-tuning the new SEQUOIA model from a pre-trained version (“avg_r_pre”) versus training from scratch (“avg_r_scr”) and the same with the number of well-predicted

genes. With more advanced pre-training/fine-tuning strategies or by performing training parameter optimizations there may be a way to make this work, but we decided not to pursue this strategy further. Instead we decided to focus on the two new components, each of which showed much larger and more consistent improvements across all cancer types than the GTex pre-training did for the previous SEQUOIA model.





4) we prioritized including the additional BRCA spatial dataset of 92 samples because we considered this to be of higher priority than the TCGA pre-training. This spatial prediction came with an additional ±10h of runtime per slide for extracting the tile-level predictions.

2. In the GBM spatial transcriptomics dataset, which consists of 18 slices, SEQUOIA did not predict any gene with a PCC greater than 0.1 across all slices, and only one gene had a PCC greater than 0.1 in 10 slices. Although SEQUOIA performed slightly better than HE2RNA in the comparison, this result still indicates that SEQUOIA is largely ineffective in predicting the spatial expression of genes, making it difficult to apply the tool accurately to real slices.

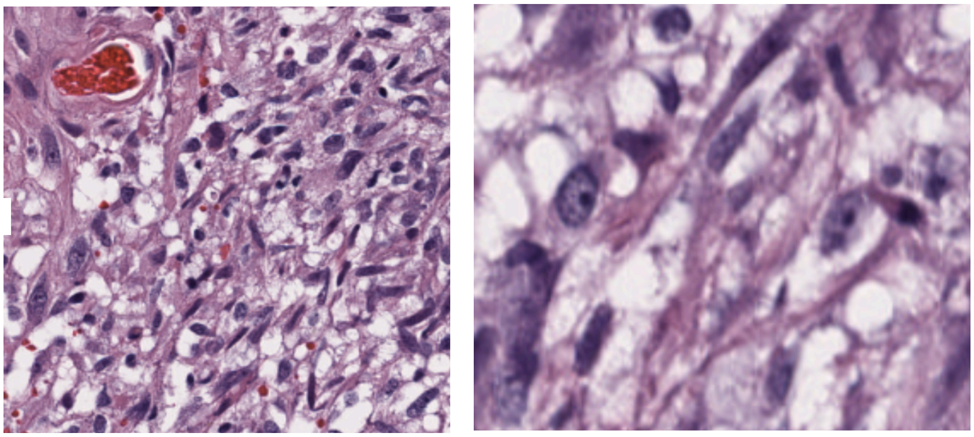
While we agree with the reviewer that the performance on the spatial GBM is modest, we want to explain a few limitations of the spatial dataset and the used metrics that cause this drop in performance (see Supplementary Figure A6 and new comments on PCC limitations and low H&E quality in spatial section). Despite these limitations, we were able to considerably improve the spatial prediction results. We elaborate on these findings below:

Reasons for worse performance on spatial compared to bulk:

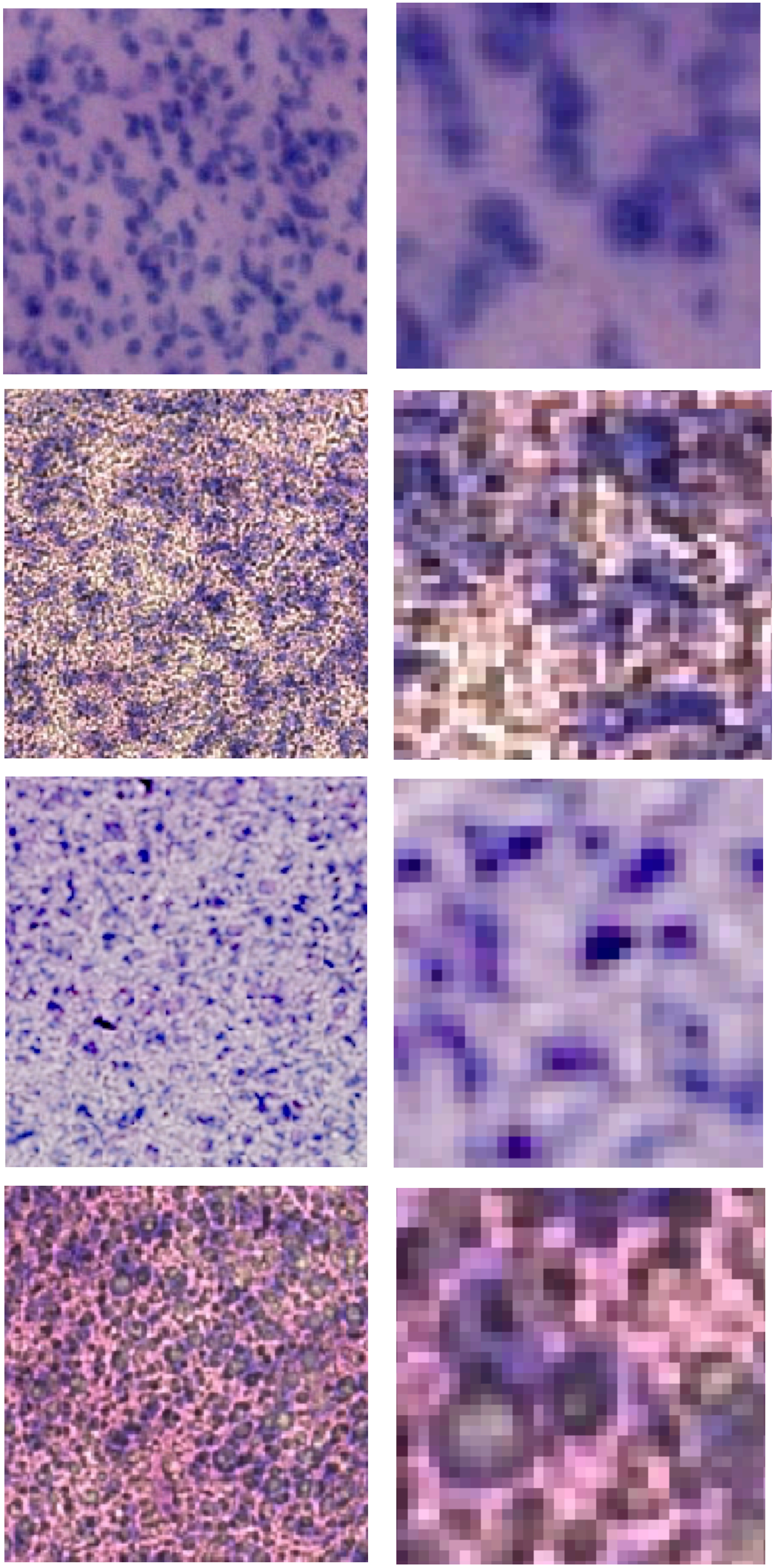
1) The quality of the H&E slides in the spatial GBM cohort is considerably worse compared to the slides in our bulk cohorts from TCGA/CPTAC/TEMPUS (see below). Hence, our spatial performance is most likely a lower bound on achievable performance if you would have the full-resolution H&E available.

   Unfortunately, currently available spatial datasets do not contain the high quality H&E. Specifically, In all 18 slides, nuclei are blurred and nucleoli are not visible within the patches. Further, there are blurring/ringing/blocking artifacts present.

   For reference, this is a TCGA-GBM sample (zoomed in version on the right where nuclei and nucleoli details are visible)

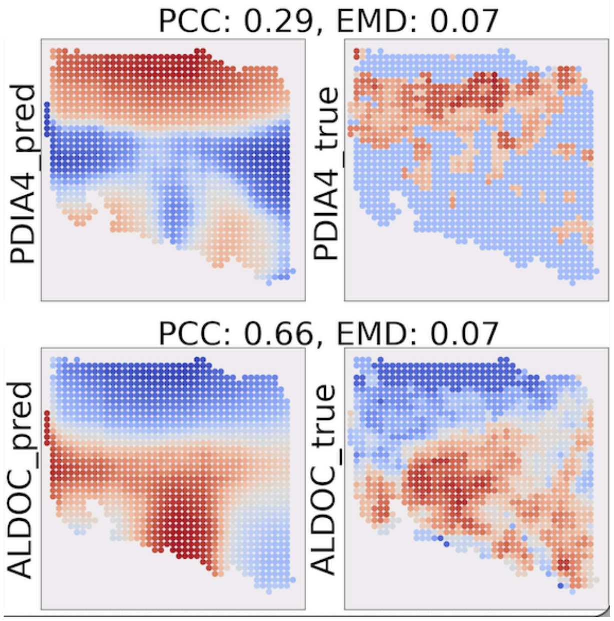


For comparison, these are examples from four slides in the spatial cohort: (with the last example representative of the best quality slides in the cohort)



2) PCC is not always a good evaluation metric for spatial prediction performance:
    a) The spatial slides are significantly smaller and more homogeneous than TCGA slides, in which case, PCC cannot fully capture true performance. Specifically, the slides contain only between 210 (min) and 1550 (max) tiles, while the TCGA slides consistently contain > 4000.
    b) PCC does not consider the spatial aspect of the predictions: if the prediction shifts with a few pixels, this heavily impacts the PCC, while in reality this small shift may not be that noticeable (see below for example).
    c) These two downsides were the reason for including EMD next to PCC as evaluation metric, since EMD is not sensitive to slide size/heterogeneity and it takes into account the spatial distance between ground truth and prediction (as described in Methods).

       A concrete example of two cases with very different PCC but same EMD is present in Figure 5 (enlarged snippet below). In both cases, the prediction highlights a region that is close to the ground truth, but slightly shifted in some places. Although the visual assessment of performance for both would be similar (reflected in equal EMD), the difference in PCC is very large (0.37).



3) Further, the low number of genes with PCC>0.1 across many slides is likely attributed to the slides varying in terms of size, heterogeneity, staining and quality. Still, we were able to improve this part as well. We can now predict 60 genes across at least 8 slides, compared to 18 before. The number of genes that validate across *N* slides (N>=2) was included in Supplementary Figure A8. Specific gene lists that validate to at least 5 slides were included in Supplementary Table A13.

Improved performance with new model:
While we cannot improve the quality of the available H&E, we were able to improve the SEQUOIA model by making use of the more robust feature extractor UNI and improved attention mechanism. This considerably improved the spatial prediction performance, showing that relevant information can still be extracted from the H&Es despite the lower quality (see Supplementary Table A12, Supplementary Figure A7, Supplementary Figure A8 and updated spatial section). In addition, this observation suggests that the spatial prediction performance can be further improved upon in the future when better feature extractors or other architecture improvements become available.

4. Related to the previous point, the authors chose GBM as the tumor type for their study. However, based on Figures 1d-e and 3a, GBM does not appear to be a tumor type with good prediction results. I recommend the authors study a tumor type with better evaluation results, such as BRCA, and discuss whether the prediction performance in spatial gene expression is related to the model evaluation performance.

We agree with this observation from the reviewer, and have therefore included results on a spatial BRCA dataset of 92 samples (see Supplementary Figure A10, Supplementary Tables A14, A15, A16 and text in spatial section). Although our model performs significantly better on BRCA than on GBM for bulk prediction, the spatial performance metrics within slides turned out similar. The performance in both cases is most likely primarily limited by characteristics of spatial transcriptomics datasets described above (worse H&E quality, smaller and more homogeneous slides). Same as for GBM, we included examples of the quality of the H&E images from the spatial BRCA in Supplementary Figure A9.

3. On Page 3, in the last paragraph, the authors claimed that the number of accurately predicted genes is related to the sample size, but the evidence from comparisons across different cancer types is indirect. I suggest the authors conduct downsampling of the sample size within the same cancer type to obtain direct evidence of the relationship between the number of accurately predicted genes and the sample size.

We appreciate this suggestion by the reviewer, and have now extended our analysis to also include a downsampling experiment within a constant cancer type. We chose BRCA since it has the most samples, allowing us to test a range of different dataset sizes. There is a consistent trend of decreasing performance across all metrics when the dataset size is reduced, see main text and Supplementary Table A6.

5. In Figure 5b, when showing example genes, I suggest that the authors annotate the PCC or EMD values in the figure to help readers understand the actual performance of

the tool.

We thank the reviewer for this suggestion. We have included these values now.

6. The authors developed a web application to display the model's predictions on TCGA data. However, the current web page can only read and display slices from the TCGA database sequentially. I suggest the authors add a feature to select or search for slice IDs, which would make the website more user-friendly.

We thank the reviewer for this suggestion. We are working to deploy the new website, and will have it deployed in 1-2 weeks.

The revised code is now operational, but users need to install `openslide` (>v3.4.0) in addition to the dependencies listed in `requirements.txt`. This requirement should be specified in the GitHub repository.

We appreciate the effort invested by this reviewer in testing our codes. We have now added the requirement for the `openslide` library in our GitHub repository.

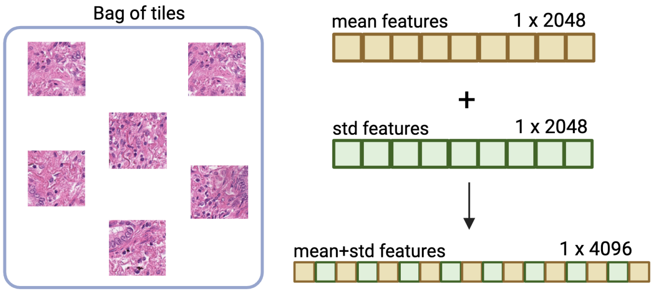
Reviewer #2 (Remarks to the Author):

I would like to thank the authors for addressing this reviewer's comments. There are two additional suggestions that this reviewer believes will further strengthen the study.

1- It is understandable that adding learnable attention mechanism will increase the complexity of the model (#of parameters, memory requirements). However, a simple mean, although can remove noise, can still obscure variation in features. I would recommend adding std alongside the mean as this strategy will not add too much complexity and may in fact improve results.

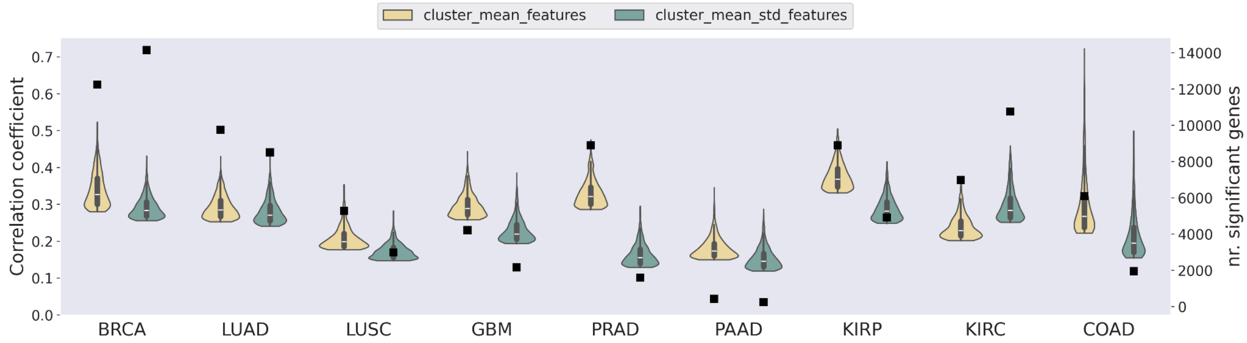
We appreciate the suggestion made by this reviewer. We performed this analysis in parallel with the experiments on page 1 for benchmarking the new model components, so we performed this experiment on the original SEQUOIA model (ResNet+transformer), on the original 9 cancer types.

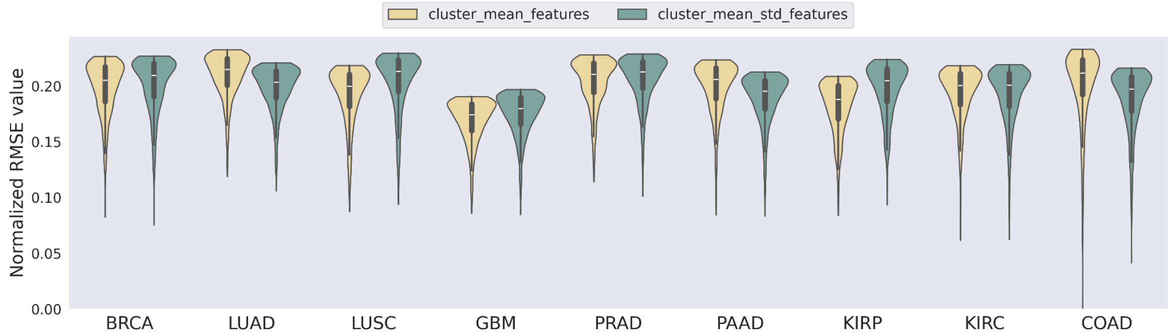
After adding the std, the feature dimension of each tile increased from 1 X 2048 to 1 X 4096. As detailed in our manuscript, we utilized both Pearson correlation analysis and root-mean-squared error (RMSE) to compare the ground truth versus the predicted gene expression values for evaluating the models.



However, adding the std resulted in an overall decreased predictive performance compared to using the mean features alone (see figures below). In terms of the correlation coefficient, adding std decreased the predictive performance in 8 out of 9 evaluated cancer types (Mann Whitney U test, P < 3E-24). The only exception is KIRC, where adding the std significantly increased the correlation coefficient (P = 9E-204). In terms of RMSE, adding the std resulted in an increase in RMSE values in 5 out of 9 evaluated cancer types (i.e., BRCA, LUSC, GBM, PRAD, KIRP) (Mann Whitney U test, P < 0.0008). Adding the std decreased the RMSE in only 3 cancer types (LUAD, PAAD, COAD). In terms of the number of well-predicted genes that passed our significant thresholds, adding the std decreased the number of the well-predicted genes in 7 out of 9 evaluated cancer types.

The decrease in the model's overall performance by adding std is likely due to feature redundancy, which resulted in overfitting of the models. Therefore, we did not further pursue this strategy.





2- With respect to reviewer #1 comment related to the number of datasets utilized in this study, it is true that there is lack of existing data resources that contain both H&E images and matched RNA-seq data. However, TCGA contains other large cohorts that could be utilized in this study. I understand that there won't be independent datasets to validate findings but a cross-validation strategy is something to consider. The reason that I suggest the above analysis is that I see a lot of merit in this study and would like to see how the model generalizes to other cancers. It is perhaps not possible to perform the suggested analysis on all TCGA cohorts, but the authors can select the cohorts in TCGA with large # of cases to do this extra analysis.

We thank this reviewer for phrasing the merit of our study. In response to this comment, we have now added 7 additional cancer types (SKCM, THCA, UCEC, HNSC, STAD, BLCA, LIHC) in our revised manuscript, which brings the total number of tumor specimens to 7,584 from 16 cancer types. It is important to note that the dataset has now included all TCGA cancer types which have more than 300 tumor samples available for both H&E and RNA-seq data, allowing us to faithfully train and evaluate our model for cancer-type specific gene expression prediction. We have updated the Figures 1 and 3 and the corresponding Results section to demonstrate the performance of our model in all evaluated cancer types.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We appreciate the time and effort invested by this reviewer in evaluating our manuscript. Please see our point-to-point response above.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

I have no more questions.

Reviewer #2 (Remarks to the Author):

Thank you for addressing this reviewer's comments. I have no further suggestions.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.