nature portfolio

# Peer Review File

## Unraveling the spatial organization and development of human thymocytes through integration of spatial transcriptomics and single-cell multi-omics profiling

**REVIEWER COMMENTS**

**Reviewer #1 (Spatial imaging, T/thymocyte activation) (Remarks to the Author):**

The manuscript by Li et al characterizes the changes in human thymus over time by integrating single-cell multi-omics and spatial transcriptomics. To integrate and analyze their data, the authors developed algorithms such as TSO-his/hismap. This is a huge "atlas" which I am sure will be a great resource for many researchers.
This being said, I find it very difficult to follow and while I like the atlas and integration part, I am not personally convinced by some of their novel findings. The manuscript is extremely dense (could honestly have been multiple manuscripts) and is trying to deal with multiple very different questions. As a consequence, I feel some parts were not fully characterized and therefore overstated. In particular:
- In Figure 3, the authors describe the transient expression of CD99 and Elovl4 in DN/DP stages. The differences when they moved to a mouse system is not apparent (for CD99) or very mild. However, it is not clear whether/when those genes are actually expressed in mice. Checking protein expression in both mouse and human thymocytes at different stages would have increased the confidence in the conclusions. In addition,concerning Elovl4, it would be interesting to assess the Treg population since the phenotype is mainly apparent in CD4 cells.
- I must say that I am really confused about the b+d+ chain expressing thymocytes. The fact that they are more prominent than g+d+ cells is really surprising, especially for fetal thymus. Again, expression at the protein level would have helped. What is the hypothesis on the role of those cells? Is the d-chain important for thymic selection? Or did those cells wrongly recombined their TCR chains and will die? As it is, this is not really convincing and difficult to bring that back into perspective given current knowledge on thymocyte development.
- The authors described different DN and DP stages, but it is unclear how they relate to the current classification of thymocytes (DN1, DN2, and DN3 thymocytes for example). I then wonder whether the expression of the d-chain detected fits with the bifurcation between ab vs gd thymocytes and could be some kind of undetected doublets?
- The authors describe thymoma cTEC as tumor cells. What defines a tumor cell here, besides some transcriptional profile? Tumors are often characterized as hyper-proliferative and carrying specific oncogenes and mutations. Is it what is happening? I don't really see this statement as being fully backed-up. Is there anything in the literature that describes cTEC as behaving like tumors?
- There are lots of inferences in the text, where I think too much is concluded from transcriptional profile only. For example, Fig.7g shows correlation between low expression of MHC at the transcriptional level and abundance of mature CD4 and CD8 T cells, not causation. While I do agree that MHC protein expression level is important for T cell selection, it is unclear what the reported transcriptional differences mean for protein expression/stability. Similar assumptions are made for the chemokine network in thymoma.
- Finally, I wonder in what context the thymus was obtained. Did some patients have specific diseases?

**Reviewer #2 (Systems immunology, machine learning) (Remarks to the Author):**

I find it helpful to follow the questions raised by the guidelines,

1. What are the noteworthy results?
The is an excellent manuscript, that puts to use current techniques and is able to create an

interesting synergy between the techniques. Somewhat exaggerating, I would say that the work is able to recreate most of what has been discovered about thymic stages over the past 40 years, using a self-contained single set of experiments. While most of these discoveries have been obtained painstakingly using extremely complex model organisms and systems, here it is done mostly by following (not easy to obtain) available tissue and experimental technologies, compiled using computational technologies.

2. Will the work be of significance to the field and related fields?
Yes, I believe it would. It shows elegantly how the complexity of thymic maturation is captured using ST, histology, and their combination.

3. How does it compare to the established literature?
The work provides its own trajectory

4. Does the work support the conclusions and claims, or is additional evidence needed?
Most conclusions and claims are being supported by previously known facts about thymic maturation. No additional evidence is needed in my opinion.

5. Are there any flaws in the data analysis, interpretation and conclusions?
a. Much of the work is based on cell annotation. I was unable to find how this annotation was made. It could be somewhere in the methods, but I was unable to locate it. The information is definitely not in the main text, the figures, the extended figures or anywhere else in the manuscript. My question is, how sensitive is the whole analyses to the specific tool choice? Since it is not mentioned anywhere, I assume the authors used a Seurat function. What would have happened if they used one of the other methods (any of the other methods claims superior performance obviously).
b. The authors later claim that the accuracy of the cell annotation has been validated by measuring the purity of the cell clusters. As far as I understand this, it would not show accuracy, only consistency.
c. Figure 1f needs some sort of range information or any equivalence of error bars. It is not possibe to understand from the figure how different measurements really are.
d. What is the use of the word “engrave” of line 245?
e. Description of the classifier is insufficient. It would have been good to use one of the reporting standards such as DOME (https://doi.org/10.1038/s41592-021-01205-4)
f. Fig 4b contains many illegible characters
g. Extended data fig 7e contains a pseudo time order that for some reason follows proximity values in UMAP presentation. There is no reason for that to happen. See extensive literature by Lior Pachter (for example) about that.
h. Not sure I understand the results of 5k. All population types seem to be at 1 in DN, while the text refers to one specific population.
i. Also, for 5k, what are the error bars for these values?
j. Regarding the entire thymoma section. Although I very much appreciate the effort, it does not seem to be a needed section for the manuscript. It is not thorough enough to show significant findings, and in my humble opinion interrupts with the flow of the manuscript.

6. Do these prohibit publication or require revision?
They definitely do not prohibit publication, but they do require revision.

7. Is the methodology sound?
Yes

8. Does the work meet the expected standards in your field?
Yes

9. Is there enough detail provided in the methods for the work to be reproduced?
Yes, the methods are well detailed.

10. Regarding the figures. it is extrmely difficult to navigate between the figures as they do not have titles. I do not know if this is an editorial issue or a mistake the authors made, but the guesswork during the review process of which is which makes the process extra difficults, especiialy in a manuscript with so many figures.

**Reviewer #3 (Thymocyte development, immunotherapy) (Remarks to the Author):**

General Comments

This paper presents a comprehensive analysis of single cell gene expression (scRNA-Seq) and spatial transcriptomics (ST) in human thymus sampled from healthy subjects at different ages: prenatal, pediatric, adult and geriatric, as well as thymoma samples from older adults. To identify cell subsets based on gene expression profiles, they use relatively standard bioinformatic approaches (clustering, re-clustering, pseudotime trajectories) that many biologists who lack specialized bioinformatic/computational expertise can understand and use. However, these largely detected known T cell and non-T cell subsets and thus confirmed prior work in mice and humans (referenced in the manuscript), including how the relative abundance of certain subsets changes with age. They also present a bespoke bioinformatic approach (TSO-His) for analyzing the ST data and report that it performs better than existing tools to accurately predict the known localization of certain TEC subsets. They provide a lengthy and detailed methods section explaining these approaches, but understanding the computational logic and validity of the tool comparisons require highly specialized expertise in computational bioinformatics, and thus are best validated by publication in a peer-reviewed bioinformatics journal before being applied to novel datasets. The computationally predicted cellular localizations from the ST data fit the known patterns for many cell types and genes, but this could be a situation of “overfitting” based on prior knowledge. Two novel genes (ELOVL4, CD99) were discovered to be highly expressed in cortical DP thymocytes. Impressively, the authors devised a Crisper-Cas9 knocout approach to study their functions in murine thymopoiesis. Disruption of Elovl4 slightly decreased CD4 SP differentiation and function but mechanisms were not explored; Cd99 loss had no impact. The analysis of single cell TCR clonotypes confirmed the known decline in diversity of TCR recombination with age and also detected unusual cells with rearranged Trcd and/or Trcb and Trca, but their numbers were not reported. The existence of these unusual subsets and their inferred developmental trajectories were not independently validated using more standard methods, leaving the biological significance of these subsets unexplored. The thymoma analyses seem novel but unconnected to the main thrust of the study. Overall, the combined approaches yielded little new information, perhaps because 1) the cell type annotations and predicted localizations relied too heavily on extensive prior knowledge of both mouse and human thymus rather than a priori discovery and/or 2) insufficient cells per thymus were analyzed (see below). Studies of human tissues are understandably often descriptive but can generate new hypotheses to explore. Unfortunately, in this study most of the computationally inferred subsets and relationships and cell interactions are already known, and few validations were performed to validate the computational predictions of

novel subsets.

Statisitical Comments:
1) The age-related subset comparisons are underpowered, due to sample sizes (N=2 for adult and geriatric groups).
2) This study also suffers from underpowered cell sampling. Only 5,000-12,000 cells/sample were analyzed by 5'GEX, which is not sufficient to characterize rare T cells and all of the non-T cell subsets (B cells, DC, Macs, Rb, TECs, Ery) which represent 0.01-1% of all thymocytes. Thus in the reported data, these subsets would represent <1-10 cells/sample. This means they could not be measured or analyzed with reasonable precision or accuracy, and thus the reported data on these subsets may not be reliable. The number of cells analyzed for each subset in each thymus should be reported as a Suppl. Table.
3) P values for DEG are reported - how was correction for multiple testing performed?
4) What statistic can be provided to assess the robustness of the inferred cTEC/thymocyte interactions predicted by CellChat? How does this method compare to CellphoneDB used by Park et al.?

Other Comments:
1) Rationales for certain subset annotations are murky. For many of the very rare subsets, only 1 or a few bona fide subset specfic genes are present (Suppl Table 2) and in some cases genes belonging to other subsets are included. This problem likely maps to the extremely low numbers of cells analyzed for these rare subsets. For example:
a. the Fb subset highly expresses Fgf7, but are any of the other most significant DEG (p-val) Fb-specific? Which highly significant DEG identify the Fb-cycling subset as fibroblasts?
b. The cTEC DEG list includes KRT5, an mTEC marker.
2) How was the CD8aa subset, which is extremely rare, computationally defined? Among the top DEG reported in Suppl. Table 2, only GNG4 was uniquely found in this subset by Park et al.
3) Methods for preparing thymus single cell suspensions, including enzymes used to release stromal cells, should be included.
4) It is not clear how enrichment/depletion of subsets across ages was calculated. Methods refers to Chi-Square of observed/expected – what is expected for each subset? Lenged for Fig. 1F says rato of observed/random was used. Which is it and what does random mean in this context?
5) How were batch effects for cell preparation, library construction and sequencing across experiments assessed or normalized?
6) The methods section mentions that different resolutions were used for clustering and re-clustering, depending on the subsets. Why not use the same resolution for all clustering steps?
7) Why are different tools used to assess purity of subsets in healthy thymus vs thymoma (ROGUE vs ESTIMATE)? How was subset purity validated?.

Tissue sampling: Partial thymectomy is commonly done in children <3Y for surgical repair of several cardiac malformations, but this is not needed once the thymus shrinks in adolescence. Why was thymus removed from healthy adult and geriatric subjects?

**We deeply appreciate the valuable feedback and insightful suggestions provided by the reviewers, as they will significantly enhance the quality of our manuscript. Presented below are point-to-point responses addressing the comments raised by each of the three reviewers. Notably, in response to the reviewers' feedback on the complexity of our manuscript and the inconsistency in the analysis between the thymoma section and normal thymus, we have removed the thymoma section from our revised manuscript. However, we remain committed to addressing the reviewers' concerns regarding this section.**

**Reviewer #1 (Spatial imaging, T/thymocyte activation, Remarks to the Author):**

The manuscript by Li et al characterizes the changes in human thymus over time by integrating single-cell multi-omics and spatial transcriptomics. To integrate and analyze their data, the authors developed algorithms such as TSO-his/hismap. This is a huge "atlas" which I am sure will be a great resource for many researchers. This being said, I find it very difficult to follow and while I like the atlas and integration part, I am not personally convinced by some of their novel findings. The manuscript is extremely dense (could honestly have been multiple manuscripts) and is trying to deal with multiple very different questions. As a consequence, I feel some parts were not fully characterized and therefore overstated.

**Major concerns:**

1. In Figure 3, the authors describe the transient expression of CD99 and Elovl4 in DN/DP stages. The differences when they moved to a mouse system is not apparent (for CD99) or very mild. However, it is not clear whether/when those genes are actually expressed in mice. Checking protein expression in both mouse and human thymocytes at different stages would have increased the confidence in the conclusions. In addition, concerning Elovl4, it would be interesting to assess the Treg population since the phenotype is mainly apparent in CD4 cells.

**Response:** We acknowledge the concern about the differences observed in the transient expression of *CD99* and *ELOVL4* when transitioning from human to mouse systems. To address this, we performed additional experiments to analyze the protein expression of *CD99* and *ELOVL4* in both mouse and human thymocytes at various stages. Our new data, presented in revised Fig. 4a-c, includes protein expression profiles for *CD99* and *ELOVL4* in both mouse and human thymocytes. This additional analysis enhances the comprehensiveness of our study and provides a more robust basis for our conclusions. Regarding the analysis of the Treg population in relation to *ELOVL4*, our current findings did not reveal significant differences between the wild-type (WT) and *ELOVL4*-KO groups in the percentages of Tregs. However, the reduced Treg numbers observed in the *ELOVL4*-KO group may be attributed to the loss of total CD4 T cells (Fig. 4k). We acknowledge the importance of this analysis and will conducting a more detailed examination of the Treg population in future studies. We hope for your understanding that this manuscript serves as a resource paper, with its primary focus on providing valuable resources for the academic community.

2. I must say that I am really confused about the b+d+ chain expressing thymocytes. The fact that they are more prominent than g+d+ cells is really surprising, especially for fetal thymus.

Again, expression at the protein level would have helped. What is the hypothesis on the role of those cells? Is the d-chain important for thymic selection? Or did those cells wrongly recombined their TCR chains and will die? As it is, this is not really convincing and difficult to bring that back into perspective given current knowledge on thymocyte development.

**Response:** Thanks for the insightful attention from the reviewer. In our analysis of thymic T cells, we employed a combination of scRNA-seq and *αβ*TCR-seq data. Specifically, T cells detected with *β* chain (*CDR3*) sequences were isolated through *αβ*TCR sequencing. This resulted in a more prominent presence of thymocytes expressing $\beta^+\delta^+$ chains compared to $\gamma^+\delta^+$ cells, which is in line with our exceptions.

In addressing the reviewer's inquiry regarding the protein-level detection of $\beta^+\delta^+$ TCR, we acknowledge the limitation posed by the absence of protein-level expression data in our study. Unfortunately, the lack of specific antibodies constrains our ability to directly assess protein-level expression and function. Nonetheless, our study endeavors to offer valuable insights into the potential roles and significance of $\beta^+\delta^+$ TCR-expressing cells in thymic selection and immune responses. By doing so, we aim to lay the groundwork for future investigations in this area. Furthermore, our findings are supported by previous research, such as the study published by Frans Hochstenbach & Michael B. Brenner in Nature (PMID: 2528071), which employed northern blots to detect $\beta^+\delta^+$ TCR, providing additional support for our observations.

To address concerns about cells with $\delta^+$ TCR chains being incorrectly recombined or undergoing cell death, we meticulously curated cells exhibiting identical *β* chains (CDR3 sequences) from the four T-cell subpopulations. Subsequently, we diligently tracked the fate of these clonal cells, with the sorting of T-cell subpopulations aligning with the chronological sequence outlined in Fig. 7g of our revised version. Drawing from the evidence presented in Fig.7h, our findings unveil a distinct and coherent developmental trajectory of these cells, progressing from the nascent TCR$\alpha^-$TCR$\beta^+$TCR$\delta^+$ subpopulation to the TCR$\alpha^+$TCR$\beta^+$TCR$\delta^-$ subpopulation. Notably, exemplified by *TRB:CASSDSNQPQHF*, this progression encompasses transitions from DP_blast1 (TCR$\alpha^-$TCR$\beta^+$TCR$\delta^+$), DP_blast2 (TCR$\alpha^+$TCR$\beta^+$TCR$\delta^+$), abT (entry) (TCR$\alpha$-TCR$\beta$+TCR$\delta$-) to the SP state (TCR$\alpha^+$TCR$\beta^+$TCR$\delta^-$). This implies that the four T-cell subpopulations exhibit a sequential developmental timeline, with $\beta^+\delta^+$ representing an intermediate state in the development of type ***αβ***T cells.

3. The authors described different DN and DP stages, but it is unclear how they relate to the current classification of thymocytes (DN1, DN2, and DN3 thymocytes for example). I then wonder whether the expression of the d-chain detected fits with the bifurcation between ab vs gd thymocytes and could be some kind of undetected doublets?

**Response:** Thanks for the reviewer's insightful feedback. In our study, we systematically categorized DN cells into distinct subtypes: DN_early, DN_trans, DN_blast, DN_re, and ISP. In response to the reviewer's concerns, we conducted a comparative analysis with current established classifications, specifically DN1 (CD34+CD38-CD1A-, representing the earliest stage of T cell development), DN2 (CD34+CD38+CD1A-), and DN3 (CD34+CD38+CD1A+, indicative of *β* chain rearrangement) stages

(Fig. 7c). Our findings indicate that DN_early predominantly corresponds to the DN1 stage, DN_trans exhibits overlap primarily with DN1 and DN2, DN_blast shows enrichment in DN2 with a transitional profile toward DN3, while DN_re and ISP are primarily clustered within the DN3 stage (Fig. 7c). Furthermore, in the context of T cell DP stage, Park et al. (Science, 2020), Cordes et al. (Science Advances, 2022), and Li et al. (Genome Medicine, 2021) delineated it into proliferative and rearrangement phases. Consistent with their classification scheme, our study categorized DP ($CD4^+CD8^+$) cells into DP_blast (characterized by *MKI67* and *TOP2A*) and DP_re (identified by *RAG1* and *RAG2*, rearrangement phase) (Fig. 1c,d). Additionally, for DP_blast cells, refined annotation (DP_blast1 to DP_blast5) was conducted using gene sets associated with cell cycle phases (G1/S, S, G2/M, M, M/G1), drawing inspiration from Li et al. (Genome Medicine, 2021) (Fig. 7d; Extended Data Fig. 9g,h). In the revised version, analyses revealed consistent dynamic changes across DP_blast1 to DP_blast5 in various cell cycle stage scores, also supported by RNA velocity analysis (Fig. 7d; Extended Data Fig. 9g,h). As for DP_re subtypes (DP_re1 to DP_re4), given the lack of established classical classification strategies, the analysis involved the usage frequency of *TRAV** genes and RNA velocity methods to elucidate the overall differentiation trajectory from DP_re1 to DP_re4 (Fig. 7e; Extended Data Fig. 9j,k).

We fully recognize the importance of ensuring the reliability of analyzing the intricate interplay between T cell development and the T cell receptor chain. Through integrating single-cell thymic data with TCR-seq data, we observed stable existence of four T-cell subgroups in our thymic dataset, as well as datasets from Park et al. and Cordes et al., collectively constituting over 99% of total T cells, with each subgroup representing no less than 4% (Fig. 6d). To address concerns about potential doublets, we employed two additional doublet detection tools, scrublet (Wolock et al., Cell Systems, 2019) and DoubletDetection (Gayoso et al., 2020). Results from scrublet indicated relatively low and comparable doublet scores for the $TCR\alpha^-TR\beta^+TR\delta^+TCR\gamma^-$, $TCR\alpha^+TR\beta^+TR\delta^+TCR\gamma^-$, $TCR\alpha^-TR\beta^+TR\delta^-TCR\gamma^-$, and $TCR\alpha^+TR\beta^+TR\delta^-TCR\gamma^-$ subpopulations, with only 6 cells identified as doublets (Fig. 6e; Extended Data Fig. 8f). DoubletDetection consistently revealed doublet rates below 5% across all four T cell subtypes (Fig. 6e; Extended Data Fig. 8g). These additional analyses strongly support the notion that the four T cell subtypes are predominantly composed of reliable singlet cells.

4. The authors describe thymoma cTEC as tumor cells. What defines a tumor cell here, besides some transcriptional profile? Tumors are often characterized as hyper-proliferative and carrying specific oncogenes and mutations. Is it what is happening? I don't really see this statement as being fully backed-up. Is there anything in the literature that describes cTEC as behaving like tumors?

**Response:** Thanks for the reviewer's insightful feedback. To define cortical thymic epithelial cells (cTECs) as tumor cells, we employed multiple computational tools to delve into the characteristics of cortical thymic epithelial cells (cTECs) in thymomas. Firstly, ESTIMATE (Yoshihara et al., Nature Communication, 2013), a tool for tumor purity estimation, unveiled a notably higher tumour purity score for cTECs compared to other cell types within the thymoma microenvironment (Previous version: Fig. 6h). Secondly, CopyKat (Gao et al., Nature Biotechnology, 2021) was utilized to distinguish malignant and normal cells in the tumor microenvironment by estimating genomic copy number profiles. It highlighted a significant proportion (83.2%) of cTECs as tumor cells (Previous version: Fig.

6l). Thirdly, *InferCNV*, primarily used for inferring copy number variations, detected markedly higher levels of copy number variation in thymoma cTECs compared to normal cTECs (Previous version: Fig. 6m). Moreover, our analysis unveiled an enrichment of pathways associated with TP53 activation specifically in thymoma cTECs, in contrast to normal cTECs (Previous version: Fig. 6n). This further strengthens the argument for their tumorigenic nature and suggests their potential involvement in the abnormal development of T cells within thymomas.

To address concerns regarding hyper-proliferation and abnormal expression of oncogenes in tumors, we conducted analyses from the following aspects: 1) Cell proliferation activity: A gene set related to cell proliferation was obtained from the MSigDB database (CELL_PROLIFERATION_GO_0008283), comprising 511 genes, of which 486 were detected in the single-cell dataset. To evaluate the overall expression activity of cell proliferation-related genes in normal cortical thymic epithelial cells (cTECs) and thymic tumor cTECs, the *AddModuleScore* function in Seurat was utilized to calculate activity scores of cell proliferation-related genes in single cells. Results indicated significantly higher proliferation activity in thymic tumor cTECs compared to normal thymus cTECs **(Figure. R1; left panel)**. 2) Oncogene expression level: A list of oncogenes was obtained from the Cancer Gene List (https://www.oncokb.org/cancer-genes) by filtering for genes categorized as Is.Oncogene == 'Yes' & Is.Tumor.Suppressor.Gene == 'Yes', resulting in 54 genes. Similarly, the *AddModuleScore* function was applied to assess the expression activity of oncogenes at the single-cell level. Consistent with the cell proliferation analysis, the expression level of oncogenes in thymoma cTECs was significantly higher compared to the normal thymus group **(Figure. R1; right panel)**. These results collectively support the characterization of tumors as hyper-proliferative and carrying specific oncogenes.

Addtionally, recent studies have shed light on the heterogeneity of TECs and their potential involvement in thymic epithelial tumors (TETs). In Xin et al.'s research (Nature Communications, 2022), they identified a subset of tumor cells in type 2 thymomas primarily composed of *CCL25*+ cTEC-like cells. Additionally, Yasumizu et al. (Nature Communications, 2022) characterized a subgroup of neuromuscular mTECs within TETs. These findings suggest that certain subtypes of TECs, including cTECs, may contribute to the cellular composition of thymic tumors. While the exact mechanisms underlying the transformation of TECs into tumor cells remain to be fully elucidated, these studies provide compelling evidence that supports the notion of cTEC involvement in thymic tumorigenesis.

Special note: In consideration of the recommendation from Reviewer 2, we have removed the section on thymoma to enhance the logical coherence of the study.

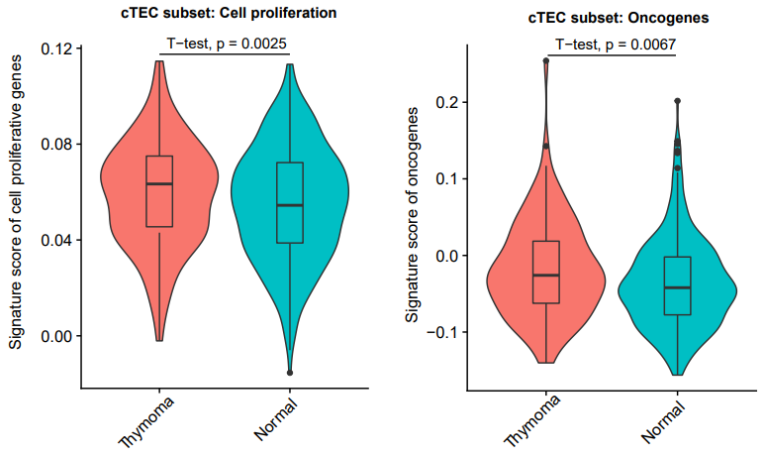


**Figure R1.** Comparison of cell proliferation activation and oncogene expression level between thymoma cTEC and normal thymus cTECs. (Left) Cell proliferation; (Right) Oncogenes.

5. There are lots of inferences in the text, where I think too much is concluded from transcriptional profile only. For example, Fig.7g shows correlation between low expression of MHC at the transcriptional level and abundance of mature CD4 and CD8 T cells, not causation. While I do agree that MHC protein expression level is important for T cell selection, it is unclear what the reported transcriptional differences mean for protein expression/stability. Similar assumptions are made for the chemokine network in thymoma.

**Response:** We extend our sincere gratitude to the reviewers for their invaluable feedback. Regarding the observed correlation between low MHC expression at the transcriptional level and the abundance of mature CD4 and CD8 T cells, we concur that further investigation into protein expression and stability is necessary to elucidate the underlying mechanisms. Similarly, the concerns raised regarding the assumptions made for the chemokine network in thymoma underscore the need for additional experimental validation to ascertain the functional relevance of the observed transcriptional differences in the context of protein expression and stability.

In response to reviewers' concerns about complexity and coherence, we've removed thymoma sections to focus solely on thymic spatial transcriptomics. Due to time constraints, protein experiments for thymoma won't be conducted in this revision. We appreciate the reviewers' feedback and remain committed to addressing the concerns in future research.

6. I wonder in what context the thymus was obtained. Did some patients have specific diseases?

**Response:** Thank you for raising this important point regarding the context of thymus procurement for our study. **Fetal Thymus:** Thymus samples from fetuses were obtained from miscarriages or stillbirths. **Pediatric Thymus:** Thymus tissues from pediatric patients were primarily sourced from children undergoing surgical interventions for congenital heart disease. Surgeons, in some cases, removed portions of the thymus to facilitate access to the operative field. **Adult and Elderly Thymus:** Thymus specimens from adult and elderly individuals were acquired posthumously from organ donors who had generously consented to multiple organ donations, and their thymus glands were donated specifically

for research purposes. All tissue procurement procedures strictly adhered to the highest ethical standards and were conducted with full informed consent from the appropriate parties, including patients or their families.

These details have been integrated into the "Thymic Tissue Acquisition" section of the Materials and Methods in our revised version.

**Reviewer #2 (Systems immunology, machine learning) (Remarks to the Author)**

The is an excellent manuscript, that puts to use current techniques and is able to create an interesting synergy between the techniques. Somewhat exaggerating, I would say that the work is able to recreate most of what has been discovered about thymic stages over the past 40 years, using a self-contained single set of experiments. While most of these discoveries have been obtained painstakingly using extremely complex model organisms and systems, here it is done mostly by following (not easy to obtain) available tissue and experimental technologies, compiled using computational technologies.

**Major concerns:**

1. Much of the work is based on cell annotation. I was unable to find how this annotation was made. It could be somewhere in the methods, but I was unable to locate it. The information is definitely not in the main text, the figures, the extended figures or anywhere else in the manuscript. My question is, how sensitive is the whole analyses to the specific tool choice? Since it is not mentioned anywhere, I assume the authors used a Seurat function. What would have happened if they used one of the other methods (any of the other methods claims superior performance obviously).

**Response:** We appreciate the reviewer's attention to detail and apologize for any confusion regarding the cell annotation process in our manuscript. In our Materials and Methods section titled "Single-cell RNA-seq data processing, integration, and dimensionality reduction," we provided a detailed description of the preprocessing, integration, and dimensionality reduction processes. Additionally, in the section "Cell type annotation and visualization for normal thymus," we outlined the strategy for the iterative refinement annotation processes, along with the key marker genes used. The full list of marker genes for cell types is provided in Extended Data Table 2.

Regarding the sensitivity of our analyses to the specific tool choice, we acknowledge the significance of this consideration. While not explicitly stated in our manuscript, our decision to utilize the Seurat package was motivated by two primary factors: First, Seurat's extensive adoption within the field and its documented efficacy across various single-cell RNA sequencing investigations (Michelson et al., 2022; Fawkner-Corbett et al., 2021; Ji et al., 2022; Chen et al., 2022). Second, Seurat offers a user-friendly data structure and comprehensive analytical workflow, specifically tailored to single-cell and spatial transcriptomic data, distinguishing it from tools focused solely on individual facets.

From preprocessing to annotating single-cell data, to the best of our understanding, the most critical determinant affecting the trustworthiness of annotation outcomes is the strategy employed for sample integration, with the aim of minimizing batch effects. In response to the reviewer's concerns more comprehensively, we conducted an analysis encompassing three integration strategies: Seurat's CCA, Scanpy's BBKNN (Polański, et al., 2020, Bioinformatics ), and Harmony (Korsunsky et al., 2019, Nature Methods). Our findings indicate that Seurat and Harmony outperform BBKNN in integrating single-cell data from various age groups (Extended Data Fig. 1d). Additionally, while the post-integration annotation results showed high concordance overall, we also observed nuances due to algorithmic differences **(Figure. R2)**. Notably, in our analysis of thymus single-cell data, we implemented an iterative refinement annotation strategy based on the Seurat tool. This approach

facilitated the identification of heterogeneous cell subpopulations within major clusters and helped mitigate biases that may arise from the use of a specific integration tool.

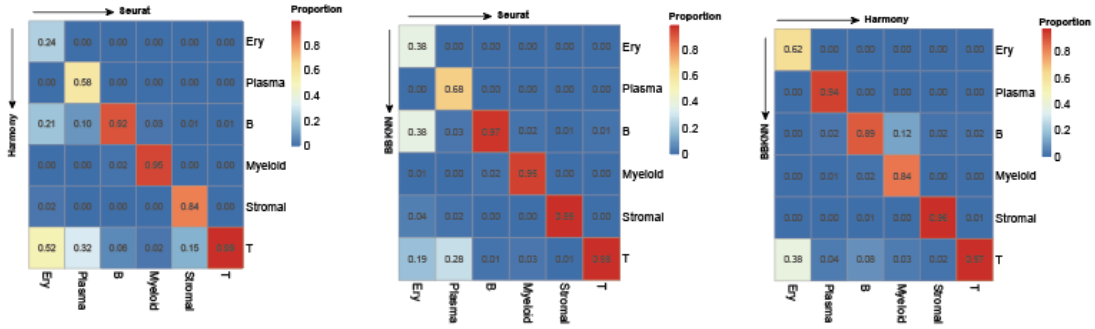


**Figure R2.** Comparison of the consistency between annotated cell types based on various integration tools.

2. The authors later claim that the accuracy of the cell annotation has been validated by measuring the purity of the cell clusters. As far as I understand this, it would not show accuracy, only consistency.

**Response:** We concur with the reviewer that accurately assessing annotations poses challenges due to the absence of a definitive gold standard and the intrinsic diversity of cell types. Upon thorough review of the previous manuscript, we found inaccuracies in the use of the term "accuracy" and subsequently replaced it with "reliable" in the revised version. In our study, we employed several strategies to evaluate the reliability of annotations. Firstly, we utilized the ROUGE tool (an entropy-based approach) developed by Liu et al. (Nature Communication, 2020) to quantify the purity of each annotated cell cluster (Extended Data Fig. 2a). Clusters with high purity are more likely to represent a single cell type, whereas those with a low purity index may contain more heterogeneous cell subsets. Secondly, we assessed whether the expression patterns of high-confidence marker genes for each cell type were specifically and prominently expressed in the single-cell clusters annotated as that cell type (Fig. 1d). Lastly, we validated our annotations using an external dataset (Extended Data Fig. 2d). By comparing our annotation with the comprehensive single-cell atlas of the thymus published by Parker et al. (Science, 2020), we affirmed the consistency of our annotations with Parker et al.'s (Extended Data Fig. 2d). Through these assessment steps, we firmly believe that the annotation of thymic single-cell data in this study is reliable.

3. Figure 1f needs some sort of range information or any equivalence of error bars. It is not possibe to understand from the figure how different measurements really are.

**Response:** Thanks for the reviewer's suggestion. In the revised version, we have added barplots with error bars (Extended Data Fig. 2f) corresponding to Fig. 1f and Extended Data Fig. 2e to clarify the variability of the presented measurements.

4. What is the use of the word "engrave" of line 245?

**Response:** Thanks for the reviewer's attention to the usage of the term "engrave". In this context, "engrave" metaphorically represents highlighting the top five most prominently expressed genes in the cortical and medullary regions, respectively. In our revised version, we have adjusted the description to provide clearer clarification for selecting over-expressed genes.

5. Description of the classifier is insufficient. It would have been good to use one of the reporting standards such as DOME (https://doi.org/10.1038/s41592-021-01205-4).

**Response:** Thank you for bringing this matter to our attention. We recognize the importance of providing standardized information regarding the classifier for the sake of reproducibility and comparability. DOME (Walsh et al., Nature Methods, 2021) offers a framework for reporting critical aspects of machine learning models, encompassing dataset, model architecture, training/testing methods, evaluation metrics, and implementation specifics. The objective of DOME is to establish a unified and dependable approach, thereby facilitating better understanding and comparison of various models. In our revised version, we have included the description of the XGBoost model based on DOME reporting standard (Extended Data Table 6).

6. Fig 4b contains many illegible characters.

**Response:** We sincerely appreciate the reviewer's feedback. In the previous version, excessive compression of the PDF resulted in pronounced distortion of certain images and text elements in Fig. 5 (latest Figure number in our revised version). Therefore, in this revised version, we have meticulously readjusted the compression ratio to effectively rectify this issue. We extend our gratitude for bringing this matter to our attention.

7. Extended data fig 7e contains a pseudo time order that for some reason follows proximity values in UMAP presentation. There is no reason for that to happen. See extensive literature by Lior Pachter (for example) about that.

**Response:** Thanks for your insightful feedback. Following your suggestion, we consulted the researches of Lior Pachter and colleagues (Fang et al., 2024, bioRxiv; Gorin et al., 2022, Plos computational biology) and fully acknowledge their viewpoints regarding the limitations of pseudo-time methods for estimating cell development sequences. As highlighted by Pachter et al., pseudo-time lacks a direct biological explanation and may be susceptible to biases arising from data noise and cell variability, potentially leading to uncertain results and overfitting issues.

To address these concerns, we have integrated RNA velocity analysis into our study. RNA velocity utilizes the ratio of spliced to unspliced mRNA abundances to infer gene dynamics and predict changes in cellular states. By combining trajectory inference with mRNA dynamic modeling, we can more accurately determine the developmental trajectories between different cell subpopulations. Our analysis indicates that development initiates at DP_blast1, progresses through DP_blast2 and DP_blast3 to DP_blast4, with an alternative pathway leading from DP_blast1 to DP_blast5 (Fig. 7d). Notably, there appears to be limited flow between DP_blast4 and DP_blast5. Nevertheless, considering the higher expression of *RAG1* recombination genes in DP_blast5, we designate DP_blast5 as the terminal stage

of DP_blast, marking the junction between DP_blast and DP_re (Extended Data Fig. 9g,h). Moreover, for DP_re subsets, RNA velocity indicates clear directional trends from DP_re1 to DP_re4 (Extended Data Fig. 9k).

8. Not sure I understand the results of 5k. All population types seem to be at 1 in DN, while the text refers to one specific population. Also, for 5k, what are the error bars for these values?

**Response:** Thanks for the reviewer's concern on Fig. 7g **(in our revised version)** in our manuscript. To investigate the temporal differentiation of four T-cell subpopulations (TCR$\alpha^-$TCR$\beta^+$TCR$\delta^+$, TCR$\alpha^+$TCR$\beta^+$TCR$\delta^+$, TCR$\alpha^-$TCR$\beta^+$TCR$\delta^-$, and TCR$\alpha^+$TCR$\beta^+$TCR$\delta^-$), we intentionally tracked T cells sharing the same Beta chain (CDR3 sequence) across these four subpopulations and tabulated their frequencies in different T-cell types (from DN_early to SP). This resulted in a statistical table, with T-cell types transitioning from DN_early to SP as rows and the four T-cell subpopulations (TCR$\alpha^-$TCR$\beta^+$TCR$\delta^+$, TCR$\alpha^+$TCR$\beta^+$TCR$\delta^+$, TCR$\alpha^-$TCR$\beta^+$TCR$\delta^-$, and TCR$\alpha^+$TCR$\beta^+$TCR$\delta^-$) as columns. After column normalization, we plotted cumulative frequency distribution curves from SP as the starting point to DN_early (i.e., steady point with corresponding cumulative value of 1) as the end point for all four T-cell subpopulations. In Fig. 7g, the region with the greatest variation in each curve represents the enriched area for a specific T-cell subtype (i.e., the main developmental time segment).

Additionally, to address the reviewer's concern regarding error bars, we have incorporated confidence intervals into the cumulative distribution curves to better capture the reliability of the temporal differentiation of the four T-cell subpopulations in our revised version (Fig. 7g).

9. Regarding the entire thymoma section. Although I very much appreciate the effort, it does not seem to be a needed section for the manuscript. It is not thorough enough to show significant findings, and in my humble opinion interrupts with the flow of the manuscript.

**Response:** We sincerely appreciate the reviewer's suggestions. After careful consideration, in our revised version, the thymoma section has been removed from the manuscript to maintain the coherence of the narrative.

10. Regarding the figures. it is extrmely difficult to navigate between the figures as they do not have titles. I do not know if this is an editorial issue or a mistake the authors made, but the guesswork during the review process of which is which makes the process extra difficults, especiialy in a manuscript with so many figures.

**Response:** Thank you for bringing this issue to our attention. We apologize for any inconvenience caused by the lack of titles on the figures. In the revised manuscript, we have addressed this concern by adding titles to most of the figures to improve clarity and ease of navigation for the reader.

**Reviewer #3 (Thymocyte development, immunotherapy) (Remarks to the Author)**

This paper presents a comprehensive analysis of single cell gene expression (scRNA-Seq) and spatial transcriptomics (ST) in human thymus sampled from healthy subjects at different ages: prenatal, pediatric, adult and geriatric, as well as thymoma samples from older adults. To identify cell subsets based on gene expression profiles, they use relatively standard bioinformatic approaches (clustering, re-clustering, pseudotime trajectories) that many biologists who lack specialized bioinformatic/computational expertise can understand and use. However, these largely detected known T cell and non-T cell subsets and thus confirmed prior work in mice and humans (referenced in the manuscript), including how the relative abundance of certain subsets changes with age. They also present a bespoke bioinformatic approach (TSO-His) for analyzing the ST data and report that it performs better than existing tools to accurately predict the known localization of certain TEC subsets. They provide a lengthy and detailed methods section explaining these approaches, but understanding the computational logic and validity of the tool comparisons require highly specialized expertise in computational bioinformatics, and thus are best validated by publication in a peer-reviewed bioinformatics journal before being applied to novel datasets. The computationally predicted cellular localizations from the ST data fit the known patterns for many cell types and genes, but this could be a situation of "overfitting" based on prior knowledge. Two novel genes (ELOVL4, CD99) were discovered to be highly expressed in cortical DP thymocytes. Impressively, the authors devised a Crisper-Cas9 knocout approach to study their functions in murine thymopoiesis. Disruption of Elovl4 slightly decreased CD4 SP differentiation and function but mechanisms were not explored; Cd99 loss had no impact. The analysis of single cell TCR clonotypes confirmed the known decline in diversity of TCR recombination with age and also detected unusual cells with rearranged Trcd and/or Trcb and Trca, but their numbers were not reported. The existence of these unusual subsets and their inferred developmental trajectories were not independently validated using more standard methods, leaving the biological significance of these subsets unexplored. The thymoma analyses seem novel but unconnected to the main thrust of the study. Overall, the combined approaches yielded little new information, perhaps because 1) the cell type annotations and predicted localizations relied too heavily on extensive prior knowledge of both mouse and human thymus rather than a priori discovery and/or 2) insufficient cells per thymus were analyzed (see below). Studies of human tissues are understandably often descriptive but can generate new hypotheses to explore. Unfortunately, in this study most of the computationally inferred subsets and relationships and cell interactions are already known, and few validations were performed to validate the computational predictions of novel subsets.

**Statisitical Comments:**

1. The age-related subset comparisons are underpowered, due to sample sizes (N=2 for adult and geriatric groups).

**Response:** We appreciate the reviewer's insightful concern. Indeed, our study exhibits an imbalance in sample sizes across different age groups: fetal (N=4), children (N=8), adult (N=2), and geriatric (N=2) stages. Initially, we attempted to combine the adult and geriatric stage samples into a single category, considering them collectively as adulthood. However, upon discovering notable differences in thymic microenvironments between these two stages, we recognized the necessity of treating them as distinct

age groups to ensure the reliability of subsequent analyses.

To address the issue of insufficient statistical power specifically concerning the adult thymic samples (N=2), as illustrated in Fig. 1f (Extended Data Fig. 2e,f), Fig. 3g (Fig. 3h), and Fig. 3i (Fig. 3j) of the revised version, we incorporated external thymic single-cell datasets for validation in our comparative analyses. One of the most crucial datasets is provided by Park et al. (Sceince, 2020), comprising substantial samples from fetal (N=37), children (N=17), and adult (N=20) stages, thereby providing a solid foundation for validating our findings and interpretations across various age groups. Unfortunately, with regard to thymic single-cell studies pertaining to the geriatric stage, besides the two normal thymus samples included in our study, no reported single-cell datasets have been found in the literature. This scarcity poses a significant challenge to our investigation during the geriatric stage. Nonetheless, our geriatric single-cell dataset still serves as an important data resource for thymic research.

2. This study also suffers from underpowered cell sampling. Only 5,000-12,000 cells/sample were analyzed by 5'GEX, which is not sufficient to characterize rare T cells and all of the non-T cell subsets (B cells, DC, Macs, Rb, TECs, Ery) which represent 0.01-1% of all thymocytes. Thus in the reported data, these subsets would represent <1-10 cells/sample. This means they could not be measured or analyzed with reasonable precision or accuracy, and thus the reported data on these subsets may not be reliable. The number of cells analyzed for each subset in each thymus should be reported as a Suppl. Table.

**Response:** Thanks for the reviewer to bring up the important concern. As per your suggestion, we have included a table into Extended Data Table 1 detailing the cell counts for each cell type across all thymus samples. Additionally, the numbers of cells containing different cell types in each thymic sample can also be observed in Extended Data Fig. 1h,i of the revised version.

Indeed, as speculated by the reviewer, rare T cells such as DN_early and T_proliferating, as well as some non-T cell subsets like Fb_cycling, Lymph, cTEC, mTEC, B_naive, and Plasma, were found in fewer than 10 cells in over 60% of samples (Extended Data Fig. 1h,i). Because of our thymus single-cell data were not sorted using flow cytometry for specific cell types, this ensured a consistent background distribution of cellular compositions within the thymic microenvironment across different samples. Consequently, even if certain cell types were represented by a limited number of cells within a sample, they remained statistically comparable across age groups defined by distinct individuals. Additionally, to mitigate potential analysis biases, we introduced data from Park et al. for concurrent validation in our comparative analyses. We believe this strategy can enhance the reliability and validity of our research findings.

3. P values for DEG are reported - how was correction for multiple testing performed?

**Response:** The correction for multiple testing was conducted using the Bonferroni method. Specifically, we utilized the *FindAllMarkers* function from the "Seurat" package, applying the Wilcoxon test strategy to identify differentially expressed genes (DEGs). These *p*-values were then adjusted using the Bonferroni method to account for multiple comparisons. We have incorporated

this detail in the Materials and Methods section of the revised version to elucidate the primary criteria utilized for the selection of DEGs.

4. What statistic can be provided to assess the robustness of the inferred cTEC/thymocyte interactions predicted by CellChat? How does this method compare to CellphoneDB used by Park et al.?

**Response:** Thanks for the reviewer's insightful feedback and concerns regarding the robustness of the inferred cell-cell interactions (CCIs) predicted by CellChat. In response to reviewer's concerns, we have conducted a thorough evaluation utilizing a downsampling strategy to assess the stability of CellChat's predictions regarding interactions between fibroblast (Fb) cells and thymic cells *(It's noteworthy that we did not investigate cTEC interactions due to the removal of the thymoma section as per the suggestion of reviewer 2, ensuring the coherence and integrity of our manuscript)*. The implemented approach is outlined as follows: 1) Downsampling was performed for each cell type based on given cell proportions. Starting with a 100% initial sampling ratio, and iteratively decreased the ratio by 5% until reaching 50% of the original cell type count, resulting in 11 downsampled datasets. 2) Inference of Fb-Thymocyte interactions using CellChat for each downsampled dataset, respectively. 3) Measurement of interaction strength consistency using the Coefficient of Variation (CV).

Through this method, we observed a generally stable interaction strength between Fb and thymic cells across different downsampling datasets, particularly notable between Fb and DP_blast to CD8T_mem, with a CV less than 5% (Extended Data Fig. 4b,c). This analysis underscores the reliability of CellChat's predictions even under varying sampling percentages.

Furthermore, to provide a comparative perspective with CellPhoneDB, another widely used tool for cell-cell interaction analysis, we conducted an investigation. Our findings revealed a high degree of consistency between CellChat and CellPhoneDB, with over 80% of the significant receptor-ligand pairs inferred by CellChat being identified in CellPhoneDB, particularly for interactions involving Fb->DP_blast and Fb->DP_re, where the consistency is almost 100% (Fig. 2i). This finding demonstrates the robustness and reliability of CellChat in inferring receptor-ligand interactions, aligning closely with the outcomes obtained through CellPhoneDB.

To further enhance the rigor of our analysis, only receptor-ligand pairs as significant by both CellChat and CellPhoneDB were retained for visualization (Fig. 2j).

**Other Comments:**

1. Rationales for certain subset annotations are murky. For many of the very rare subsets, only 1 or a few bona fide subset specfic genes are present (Suppl Table 2) and in some cases genes belonging to other subsets are included. This problem likely maps to the extremely low numbers of cells analyzed for these rare subsets. For example:

a. the Fb subset highly expresses Fgf7, but are any of the other most significant DEG (p-val) Fb-specific? Which highly significant DEG identify the Fb-cycling subset as fibroblasts?

**Response:** Thanks for the reviewer's valuable feedback. We have carefully examined the list of differential genes in Extended Data Table 2. We noticed that our previous compilation of differential genes in the Fb_cycling subset was derived from the comparison between Fb and Fb_cycling, aiming to highlight the cell cycle process occurring in Fb_cycling subset. This might have caused the marker genes of Fb to not manifest in Fb_cycling. However, through violin plots, we observed that the Fb-associated marker gene *FGF7* also exhibited high expression in Fb_cycling **(Figure. R3; top panel)**. Additionally, in Fb_cycling, cycle-related genes such as *TYMS* showed specific high expression **(Figure. R3; bottom panel)**. In our revised version, we have updated the list of Fb_cycling marker genes to Extended Data Table 2.

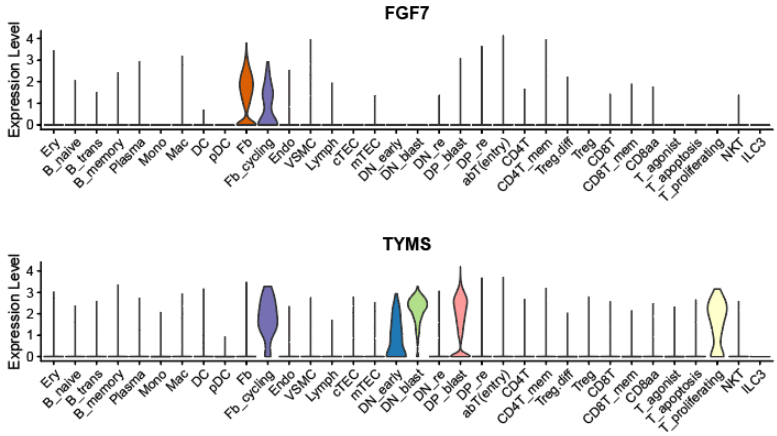


**Figure R3.** The expression of genes *FGF7* and *TYMS* across different cell types in thymus single-cell data.

b. The cTEC DEG list includes KRT5, an mTEC marker.

**Response:** We appreciate the reviewer's valuable feedback. During our annotation process, we categorized thymic epithelial cells (TECs) into cortical TECs (cTECs) expressing *CCL25* and medullary TECs (mTECs) expressing *CCL19* as the primary groups (Fig. 1d). To address the reviewer's concerns, we included violin plots illustrating the expression patterns of *CCL25*, *CCL19*, and *KRT5* across various cell types in the thymus. Specifically, we observed specific high expression of *CCL25* in cTECs and *CCL19* in mTECs, confirming the accuracy of our annotations **(Figure. R4A)**. However, *KRT5* expression showed similar intensity in both cTECs and mTECs, without clear specificity **(Figure. R4A)**. Additionally, we subdivided cTECs into *KRT5*+ and *KRT5*- clusters, finding that *KRT5*+ cells represented 76.4% of cTECs, while in mTECs, they accounted for 95.8%.

To explore whether the *KRT5*+ cTEC subset in our data might reflect misclassified mTEC clusters, we compared expression profiles with those (mTEC subset as external independent data) from Park et al. (Science, 2020). We found a strong correlation (0.96) between KRT5+ cTECs and KRT5- cTECs, while correlations with mTECs were consistently low (below 0.15), suggesting a distinct KRT5+ cTEC subtype **(Figure. R4B)**. Indeed, Park et al. also identified such a subtype within cTECs, termed "mcTECs," characterized by high *KRT5* expression. Given the unsorted nature of our data, leading to fewer cTEC cells, we opted not to further refine cTEC clustering to ensure robustness in subsequent analyses.

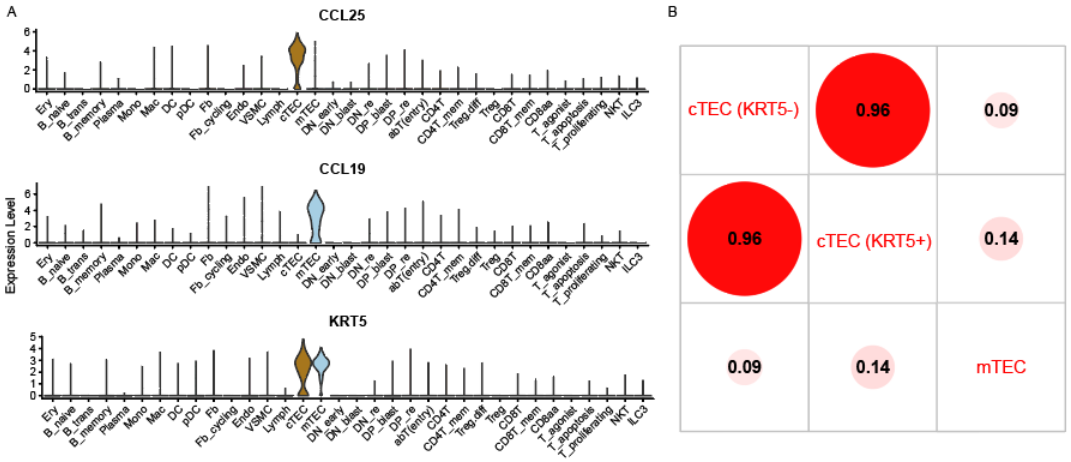


**Figure R4.** (A) The expression of genes *CCL25*, *CCL19* and *KRT5* across various cell types in thymus single-cell data. (B) Correlation matrix among *KRT5*+ cTEC, *KRT5*- cTEC, and mTEC (data source from by Park et al.)

2. How was the CD8aa subset, which is extremely rare, computationally defined? Among the top DEG reported in Suppl. Table 2, only GNG4 was uniquely found in this subset by Park et al.

**Response:** Thank you for raising this important point. CD8aa represents a distinct subset of T cells, distinguished from traditional CD8+ T cells by expressing the *CD8A* molecule while suppressing *CD8B*. Functionally, CD8aa potentially acts as a core inhibitor of TCR activation by reducing antigen sensitivity, thus negatively regulating T cell activation (Cheroutre et al., Immunity, 2008). This subset has been characterized in single-cell studies of the thymus by Park et al. (2020, Science) and Suo et al. (2022, Science), with *GNG4* serving as a pivotal marker gene. In our analysis of thymic T cell clusters, we computationally defined the CD8aa subset adhering to specific criteria: $CD8A^{hi}CD8B^{lo}CD4^{-}$, $GNG4^{hi}$, as depicted in the revised violin plots (Extended Data Fig. 7m,n). CD8aa cells compise approximately 2.32% of total T cells.

To address the reviewer's concern regarding CD8aa marker genes in Table S2, we cross-referenced the top 40 genes (ordered by *log2FC* values) significantly upregulated in CD8aa with the CD8aa.I (n=20) and CD8aa.II (n=20) marker genes provided by Park et al. (Science, 2020). Apart from *GNG4*, we found 11 overlapping genes (Extended Data Fig. 7o). This suggests a robust concordance between the CD8aa subset determined through our computational approach and previous study.

3. Methods for preparing thymus single cell suspensions, including enzymes used to release stromal cells, should be included.

**Response:** We have included the methods for preparing thymus single cell suspensions in the revised manuscript ("Tissue dissociation and preparation of single-cell suspensions" part in ***Materails and Methods*** section). Briefly, tissue samples were processed by washing in precooled PBS, cutting into small pieces, and incubating with the Miltenyi MACS Human Tumor Dissociation kit at 37°C. Subsequently, the cell suspension was filtered through a 40μm cell strainer, followed by centrifugation and treatment with Red Blood Cell Lysis Solution. Cell viability was assessed using trypan blue

staining, and the concentration was adjusted accordingly. The final cell suspension was used for downstream experiments, such as single cell immune profiling using the 10X Genomics chip. For more details, please refer to our revised manuscript.

4. It is not clear how enrichment/depletion of subsets across ages was calculated. Methods refers to Chi-Square of observed/expected – what is expected for each subset? Lenged for Fig. 1F says rato of observed/random was used. Which is it and what does random mean in this context?

**Response:** Thank you for raising this concern. To clarify, in the revised version of our manuscript, we have provided detailed descriptions of how we calculated the enrichment or depletion of subsets across ages. This was determined using a chi-square test comparing observed versus expected frequencies. In our study, "observed" refers to the actual cell number of subsets across various age groups, while "expected" signifies the frequencies anticipated if there were no association between subsets and ages. These expected values were derived by multiplying the row and column margins of each subset and dividing by the total cell number. Finally, we calculated the ratio of observed to expected frequencies (*R*o/e) to evaluate the extent of enrichment or depletion.

Regarding Fig. 1f, the "expected number of cells at random" represents the scenario where there is no association between subsets and ages, allowing us to observe the expected number of cells. In the revised version, we have provided additional explanatory notes in the legend of Fig. 1f.

5. How were batch effects for cell preparation, library construction and sequencing across experiments assessed or normalized?

**Response:** Thanks for the reviewer's important concern. In our study, we addressed the mitigation of batch effects based on two aspects.

**Experimental Consistency:** To minimize batch effects during experimentation, two critical aspects were ensured: 1) Utilization of uniform equipment platform throughout the experiments. 2) Usage of commercially available enzymes from the same brand and batch across all experiments.

**Data Integration Strategy:** On the data analysis front, the CCA (Canonical Correlation Analysis) algorithm embedded within Seurat was employed to integrate multiple single-cell datasets, thereby mitigating batch effects. Before integration, significant batch effects were observed, particularly concerning different age groups in the samples (Extended Data Fig. 1c). However, post-integration, the single-cell distributions from various samples were effectively homogenized (Extended Data Fig. 1d; top right panel). This rigorous integration ensured the reliability of thymic single-cell clustering and annotations.

6. The methods section mentions that different resolutions were used for clustering and re-clustering, depending on the subsets. Why not use the same resolution for all clustering steps?

**Response:** In the process of unsupervised clustering in single-cell analysis, the resolution size emerges as the primary determinant influencing cluster numbers. Throughout the iterative annotation of thymic

single cells, maintaining uniform resolution may yield two undesirable outcomes: Firstly, an excess of cell subclusters poses significant challenges in annotation and biological interpretation. Moreover, numerous small cell clusters may lack statistical power for analysis across diverse age groups. Secondly, inadequate cell subclusters fail to distinguish closely related cell subtypes along differentiation lineages, thus constraining the comprehensive portrayal of cell types within the thymic microenvironment. Consequently, tailoring specific resolution parameters to individual cell clusters ensures the generation of more dependable and meaningful results.

7. Why are different tools used to assess purity of subsets in healthy thymus vs thymoma (ROGUE vs ESTIMATE)? How was subset purity validated?

**Response:** The utilization of tools *ROGUE* (Liu et al., 2020, Nature communications) and *ESTIMATE* (Yoshihara et al., 2013, Nature communications) proved applicable in specific biological contexts. *ROGUE* (a Shannon entropy-based method) was specifically chosen for its capacity to evaluate the purity of cell clusters. Higher ROGUE values denote purer cell clusters, whereas lower values signify higher heterogeneity, as depicted in Extended Data Fig. 2a. Conversely, *ESTIMATE* was employed to estimate tumor purity by generating gene signatures for stroma and immune cells, with higher values representing more tumor cells in solid tissue, and its effectiveness demonstrated in previous studies such as Chung et al.'s (2017, Nature Communications) single-cell research on breast cancer.

In terms of validation, for *ROGUE*, we trained a logistic regression model to define the purity index, corroborating the reliability of the *ROGUE* index (Extended Data Fig. 2a,c). As for *ESTIMATE*, *CopyKat* (Gao et al., 2021, Nature Biotechnology) was employed to estimate tumor cells. Our results demonstrate that in thymomas, over 80% of cells in the cTEC subset exhibited characteristics of tumor cells, consistent with the high tumor purity score estimated by *ESTIMATE* for the cTEC subset in thymomas (Pervious version: 6h,i).

Notably, due to the unavailability of ST data for thymomas, a research gap exists between the study of thymomas and the normal thymus. Following the suggestion of Reviewer 2 and our thorough consideration, we have opted to exclude the thymoma sections in the revised version to maintain the logical coherence of our study.

8. Tissue sampling: Partial thymectomy is commonly done in children <3Y for surgical repair of several cardiac malformations, but this is not needed once the thymus shrinks in adolescence. Why was thymus removed from healthy adult and geriatric subjects?

**Response:** Thanks to the reviewer for raising this concern. Thymic samples from fetuses are obtained from miscarriages or deaths, while those from children mainly come from patients undergoing congenital heart disease surgery, as part of the surgery involves the removal of some thymic tissue. Thymic specimens from healthy adults and elderly individuals are obtained from deceased organ donors. Although these individuals have not undergone thymectomy for medical reasons, their donated thymus glands provide valuable opportunities for studying thymic T-cell development across different age groups.

All tissue acquisition processes strictly adhere to the highest ethical standards and are conducted with full informed consent, including consent from patients or their relatives. We have revised the manuscript with more detailed explanations regarding the acquisition of sample materials.

**REVIEWERS' COMMENTS**

**Reviewer #1 (Remarks to the Author):**

The authors have simplified their narrative and clarified their conclusions, either in the text or with new experimental data. I have no further comment.

**Reviewer #2 (Remarks to the Author):**

The authors have made a real effort to respond to all of my queries. This is a good manuscript that has, in my opinion, imporoved significantly over the review process.

**Reviewer #3 (Remarks to the Author):**

The authors have provided comprehensives responses to the critique and made substantive changes to the manuscript that address the many of the specific concerns. But the fact remains that this highly technical study reports little new information.

**Dear Editor,**

The enclosed manuscript, titled "Unraveling the spatial organization and development of human thymocytes through integration of spatial transcriptomics and single-cell multi-omics profiling" (NCOMMS-23-50841A), has been revised in accordance with the insightful suggestions of the reviewers and editors. It is hereby re-submitted for possible publication in Nature Communications.

**A point-by-point reply to the reviewers' comments follows.**

**Reviewer #1 (Remarks to the Author):**

The authors have simplified their narrative and clarified their conclusions, either in the text or with new experimental data. I have no further comment.

**Response:** Thanks Reviewer #1 for your kindly feedback and acknowledgment. We appreciate your review and all insightful comments.

**Reviewer #2 (Remarks to the Author):**

The authors have made a real effort to respond to all of my queries. This is a good manuscript that has, in my opinion, improved significantly over the review process.

**Response:** Thanks Reviewer #2 for your kindly feedback and all insightful comments.

**Reviewer #3 (Remarks to the Author):**

The authors have provided comprehensives responses to the critique and made substantive changes to the manuscript that address the many of the specific concerns. But the fact remains that this highly technical study reports little new information.

**Response:** Thanks Reviewer#3 for your detailed review and the recognition of the comprehensive responses and substantive changes we have made to the manuscript. We appreciate the time and effort you have invested in evaluating our work. Our study integrates spatial transcriptomics and single-cell multi-omics to reveal novel co-localization relationships and dynamic changes in thymic cells, providing an unprecedented high-resolution map of thymic architecture. These findings significantly enhance our understanding of T cell development and will serve as a valuable resource for future research on thymic development.