Cell Systems Transparent Peer Review Record

Entropic Analysis of Antigen-Specific CDR3 Domains Identifies Essential Binding Motifs Shared by CDR3s with Different Antigen Specificities

Alexander M. Xu, William Chour, Diana C. DeLucia, Yapeng Su, Ana Jimena Pavlovitch-Bedzyk, Rachel Ng, Yusuf Rasheed, Mark M. Davis, John K. Lee, James R. Heath

Summary

Initial Submission:	Received November 15, 2021 Scientific editors: Quincey Justman, Ph.D. Ernesto Andrianantoandro, Ph.D.
First round of review:	Number of reviewers: Three *Three confidential, Zero signed* Revision invited April 27, 2022 *Major changes anticipated* Revision received September 01, 2022
Second round of review:	Number of reviewers: One *One confidential, Zero signed* Revision invited December 02, 2022 *Minor changes anticipated* Revision received December 08, 2022

Accepted March 01, 2023

This Transparent Peer Review Record is not systematically proofread, type-set, or edited. Special characters, formatting, and equations may fail to render properly. Standard procedural text within the editor's letters has been deleted for the sake of brevity, but all official correspondence specific to the manuscript has been preserved.

Editorial decision letter with reviewers' comments, first round of review

Dear Dr. Heath,

I'm enclosing the comments that reviewers made on your paper, which I hope you will find useful and constructive. As you'll see, they express interest in the study, but they also have a number of criticisms and suggestions. Based on these comments, it seems premature to proceed with the paper in its current form; however, if it's possible to address the concerns raised with additional experiments and/or analysis, we'd be interested in considering a revised version of the manuscript.

As a matter of principle, I usually only invite a revision when I'm reasonably certain that the authors' work will align with the reviewers' concerns and produce a publishable manuscript. In the case of this manuscript, the reviewers and I have make-or-break concerns that can be addressed by:

1. **Providing clearer justification of the working assumptions.**
2. **More clearly describing how SPANTCR works and what distinguishes it from competing approaches.**
3. **Ensuring fair comparison to competing approaches.**

Reviewer #3 provides a particularly lucid and insightful set of comments. Their "major concerns" section serves as an excellent guide for revision - if these can be addressed convincingly with additional analyses and changes to the text and figures, this will strengthen the manuscript. In addition to the concerns I've detailed above, I'd also like to be explicitly clear about an almost philosophical stance that we take at Cell Systems…

- We believe that understanding how approaches fail is fundamentally interesting: it provides critical insight into understanding how they work. We also believe that all approaches do fail and that it's unreasonable, even misleading, to expect otherwise. Accordingly, when papers are transparent and forthright about the limitations and crucial contingencies of their approaches, we consider that to be a great strength, not a weakness.

- We believe that the figures are the scientific backbone of the paper. Currently, it's not possible to understand the manuscript's conceptual advance from figures presented. Similarly, it's not possible to understand where your approach gets its analytical power. These things need to be demonstrated with data and analysis, in the form of figures with their legends or mathematical argumentation, and then supported with explanatory text.

As you address the concerns, it's important that you and I stay on the same page. I'm always happy to talk, either over email or by Zoom, if you'd like feedback about whether your efforts are moving the manuscript in a productive direction. Do note that we generally consider papers through only one major round of revision, so the revised manuscript would be either accepted or rejected based on the next round of comments we receive from the reviewers. If you have any questions or concerns, please let me know. More technical information and advice about resubmission can be found below my signature. Please read it carefully, as it can save substantial time and effort later.

I look forward to seeing your revised manuscript.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Reviewers' comments:

Reviewer #1: This paper presents a new computational tool to analyze TCRs with different length CDR3 variable regions. Entropy analysis is used to find k-mers that are essential for TCR:pMHC binding. This new approach can provide an overall landscapes of antigen-specific TCR.
I would suggest the authors to take following comments into account.
1) The authors should describe more details about how values of different parameters in their algorithm were adopted. For instance, the authors mentioned they applied exponential frequency functions for different databases (VDJDB: 5score, McPAS: 2id.method, PIRD: 3grade). It is not clear how these

parameters were determined.
2) The authors should provide more mechanistic interpretation on the structural biology or biophysical level about the meaning of the K-mers they identified. For instance, the authors mentioned that " Hydrophilic 2-mers appear more essential than neutral 2-mers (G, A-containing 2-mers). 2-mers with charged amino acids (N, K, Q, R) are most frequently associated with large entropy reductions". It would be helpful to dig into the origin of these discoveries.
3) Each figure of the paper contains too much information. When the authors describe their results, they skipped some technique details either in the main text or in the figure captions. As a result, some of these figures are difficult to understand especially for general audiences. Moreover, all the supplemental figures cannot be accessed. The authors should do better job during the revision to make their manuscript more easily acceptable for readers which background is not computational immunology.
4) One major assumption of the work is that TCR specificity emerges from hypervariable CDR3a and CDR3b loops, which is reasonable. However, the neighboring germline loops and the MHC protein also contribute remarkably to the binding between TCR and pMHC. One example is the complex formed by A6 TCR and the human T cell leukemia virus-1 Tax11-19 peptide presented by the class I MHC protein HLA-A*0201. The x-ray structure of the complex shows that CDR3a loop makes several strong electrostatic interactions with the HLA-A2 a1 helix. As a result, it was suggested that "when considering the determinants of TCR binding specificity we need to consider the interface in its entirety" in a recent review paper (J Immunol 2017; 199:2203-2213). I would appreciate if the authors could discuss this issue.
5) Finally, about the future application of the tool, in addition to find some global rules of the sequence patterns in CDR3 loops, how will it be practically used to predict the binding specificity of TCRs or design new TCRs, comparing other machine-learning based methods? For instance, in a recent study, antigen specificity of single T cells has been predicted by deep learning algorithm based on their TCR CDR3 regions (Fischer et al., 2020). Some perspectives need to be provided.

Reviewer #2: The manuscript by Xu et al. describes a novel tool (SPAN-TCR) that integrates length-agnostic CD3 sequence analysis for identification of k-mer sequences that determine antigen specificity of the TCR using entropy-based modelling.
I have read the manuscript with interest, and believe it is highly relevant. As the work is a little outside of my expertise, I hope they help the authors understand how their manuscript is perceived from a less specialized audience. I therefore have a number of questions and suggestions to increase clarity.

Authors analyse all 2mer sequences in publicly available CMV pp65-specific CDR3s (VDJDB database) and identify several strings that are common between TCRs of different MHC-peptide specificities within the CD3 region. The algorithm is then applied to a novel set of experimentally determined TCRs for CMV pp65 and they determine which sequences are most likely to have the antigen specificities of their published counterparts.
Authors are able to categorise the k-mers into different categories to classify their relevance to antigen specificity. Interestingly, authors then proceed to look at unrelated datasets (SARS-CoV-2) and make a series of interesting observations, one of which is that there are striking similarities between CD3 regions of TCRs that recognise the same antigen, and that the determining k-mers accumulate in the centre of the CD3 region.

My main comment is that it is not clear how the epitope specificity of the TCR was integrated into the workflow overall. Where all of the TCRs MHC-I specific (I assume so)? I first thought that in the validation experiment a pool of TCRs was pulled down using a diverse pool of MHC dextramers with a mixture of pp65 epitopes, but I take from the method section that a single peptide epitope, and a single HLA molecule (HLA-A2) was used. Does the original VDJDB data contain A2 TCR sequences specific to NLVPMVATV? Were those TCRs that were identified as close or even identical in sequence also A2 NLVPMVATV TCRs? What are the HLA associations of the SARS-CoV-2 epitopes, and did the TCRs that were closely related belong perhaps to a single 'HLA group'? Please indicate these details in the main text so they are immediately clear to the reader.

Regarding the interpretation of the results, is it not true that a TCRs within a specific epitope group also has an implied specificity to the presenting HLA molecule? Thus, the established relation between CDR3 motifs and antigens is more associative, not causal - unless authors consider and exemplify differences between epitope specificities of a single HLA group, in my option.

Reviewer #3: In 'Entropic Analysis of Antigen-Specific CDR3 Domains Identifies Essential Binding Motifs Shared by CDR3s with Different Antigen Specificities', the authors present a novel tool and accompanying R package to identify potentially relevant k-mers in epitope-specific T-cell receptor sequences (TCRs). The approach is novel and certainly interesting, however specific assumptions that it builds upon are questionable and not fully addressed in the manuscript. In addition, the advantage compared to existing methods is unclear at this time. Furthermore the results are presented as far more novel and relevant than they actually seem to be. The R package is well documented and seems easy enough to install.

Major comments:

1) The main assumption that SPANTCR builds upon is that the binding orientation of different TCRs with the same pMHC are constrained in their binding orientation, even if the TCR CDR3 length differs. However this does not match with the current literature surrounding TCR-pMHC complexes, where it has been shown that radically different complexes can involve the same pMHC. For example, reference 12 within the manuscript (Gras et al.), describes a TCR binding with a reversed conformation. This is further explored in more recent papers on TCR-pMHC binding, such as Knapp et al, (https://doi.org/10.1371/journal.pcbi.1007338). Thus the prevalent hypothesis is that different TCRs binding with the same pMHC, especially if they differ strongly in sequence (for example by length), have different constraints and are likely to use different amino acid interactions to mediate the recognition/binding. It is only very similar TCR sequences that can be expected to follow similar binding orientations.

2) SPANTCR uses an interpolation across 100 bins to find patterns that are length-invariant, from beginning to end. However it can be questioned if this approach is appropriate when considering the biochemical nature of amino acid interactions. A TCR can only interact with a pMHC with a limited

number of CDR residues. Through interpolation, SPANTCR seems to assume that the residues within the binding interface can be compressed if the TCR CDR sequence is longer. Or, in other words, that the relative positioning with regards to the full CDR length is what matters, and not the absolute positioning. For example, imagine a TCR with two interacting regions that are separated by two residues, SPANTCR assumes that a TCR with twice the length would position these interacting regions separated by four residues. This is an extreme example, but the manuscript currently does not address this concern.

3) The novelty of SPANTCR is presented as a tool 'that seeks to identify features shared between sets of TCRs sharing the same pMHC but with variable length CDR3s, or between TCR sets that bind different antigens'. With regards to the former, the length-agnostic nature, almost every common pMHC-specific TCR analysis is length agnostic: GLIPH (both 1 and 2) search for motifs within the variable CDR sequences. TCRdist (orginal and 3) use an alignment method to bridge different length CDRs. Similarly other methods that have been published to analyse (and often build models for) pMHC-specific TCR utilise solutions that transform the different lengths into something comparable, through either padding, alignment or motif detection.

4) SPANTCR usually uses 2-mer motifs. However the motivation behind 2-mer motifs is unclear, especially as most other related methods use longer motifs. The first section of the results states 'Specifically, 2-mer motifs are common in TCR CDR3 data sets'. It is unclear what is meant by 'common' in this section, or how it relates to the selection of 2-mers as the primary analysis point. In addition, the usage of 2-mers is not consistent. The comparison with GLIPH uses 4-mers, and the GXG motif found for the SEHDY epitope group is a 3-mer.

5) Based on the description of the SPANTCR method, the entropy of 2-mer distributions at each position is calculated against the background that all amino acids would occur at equal frequency. However it is well known that amino acid usage in TCR CDR3 regions is not uniform, and that some amino acids occur more or less (and likely even specific 2-mers). It would make sense to calculate the entropy against the observed amino acid usage. Indeed, in several results, SPANTCR identifies as called by the authors "common motifs", i.e. 2-mers that occur frequently and are therefore uninteresting. Adopting an alternative entropy calculation might be more appropriate.

6) There is little comparison to existing methods with regards to results. While SPANTCR is clearly different in its approach, any advantages of its use compared to, for example, GLIPH or tcrdist are unclear.

7) The only comparison to GLIPH is a check of the overlap between found common 4-mers and the GLIPH output. As this derived from the same set of TCR sequences, one would indeed expect an overlap in two methods attempting to quantify the presents of specific amino acid sequences. However while it is claimed that the 'majority' supposedly overlap, a quick count shows that more than 80 out of 135 do not. Thus less than half overlap, and not a majority. In addition, the results mention TCRMatch and a potential comparison, but no results are shown.

8) A second validation comes in the next section when a single TCR pair from the dextramer library and the VDJdb is analysed with the Levenshtein/BLOSUM metric. This is only a single example, which could

have been cherry-picked. For validation purposes, the full repertoire should be considered. In addition, a more comprehensive comparison with a 'negative' data set would be appropriate to quantify the false positive rate.

9) A main selling point of the paper is that SPANTCR identifies 'motifs shared by CDR3s with different antigen specificities'. However this finding is entirely reliant on the results presented within the MIRA data, in particular in the [SEHDY] and [AFPFT] epitope groups. However these epitope groups were already listed within the MIRA data, and these groups are known as being linked to the same set of TCRs. Indeed for most, the database does not even distinguish individual preferences. Therefore these are sets of TCRs that all share the same somewhat-degenerate epitope specificity. Any other of the aforementioned TCR analysis tools would equally pick up the same common patterns, simply due to the heavy sharing in TCRs between these epitopes.

10) Figure 6 presents a comparison between CDR3 profiles and epitope distance where a very weak but reported significant relationship seems to have been found. However these this analysis compared all epitopes versus all other epitopes, so for VDJdb this would be 17 x 16 comparison. This 17x16 data set is then subjected to correlation analysis. But this correlation analysis is inflated, as the number of samples has been artificially extended from 17 to 17x16. Thus the observations within this analysis cannot be considered as independent. The resulting significant P-value is likely the result of this inflated number of observations, as the original data set would not have the power to establish this weak relationship. A similar issue exists within the same comparison for the MIRA data set, but this is enhanced by the shared TCRs in comment #9.

11) SPANTCR identifies essential 2-mers by identifying those that constrain the entropy throughout the CDR3 sequence (even across beta - alpha chains). However the CDR3 formation is in itself constrained by the V and J genes. Therefore one could assume that the essential 2-mers are simply representations of enriched V/J usage, especially in the case of alpha chains (where diversity beyond V/J usage is more limited) and subsequent superessential motifs.

Minor comments:

12) Many of the references with the introduction are somewhat outdated or out-of-place. For example, a discussion on the CDR3 diversity to the same pMHC targets in VDJdb refers to the review of Miho at al., where this is not explicitly discussed.

13) Given that SPANTCR can be used for extracting structural insights and is based on the hypothesis of similar structures, it may be appropriate to contrast the method theoretically with homology modelling, such as for example Milighetti et al. (https://doi.org/10.3389/fphys.2021.730908) or Lanzarotti et al (https://doi.org/10.3389/fimmu.2019.02080).

14) In addition, as SPANTCR aims to identify patterns across epitopes, it might be useful to also contrast with de novo TCR-epitope interaction models, such as Moris et al (https://doi.org/10.1093/bib/bbaa318), Weber et al (https://doi.org/10.1093/bioinformatics/btab294), and Lu et al (https://doi.org/10.1038/s42256-021-00383-2)

15) VDJdb is commonly written with 'db' as lower case.

16) The full name of McPAS is McPAS-TCR.

17) The text/figures are not consistent in their naming of TBAdb and PIRD, which are used interchangeably.

18) It is unclear what the stacked amino acids represent in figure 1B.

19) The first results sections includes the statement 'shows the degeneracy at the CDR3 N-terminus' to denote a fairly consistent amino acid motif in this region. The term 'degeneracy' here does not seem fully accurate, as it would commonly be assumed to be the opposite int he context of sequence motifs.

20) The results report that its 'general findings are consistent across all three databases'. This is not unexpected as there is large redundancy in these dataset as they are all derived from much of the same public resources.

21) The methods section reads that data is derived from 'three TCR databases, VDJDB, McPAS, and PIRD, and the MIRA database'. These are four TCR databases.

22) Supplemental figure 2E features an extremely trunctated TRA sequence, 'SSGNQFYF'.

23) Supplemental figure 5A shows poor clustering of motifs for the same epitope across different databases. As SPANTCR aims to identify epitope-specific motifs, especially given the overlap between databases, one would expect strong clustering for each epitope. These findings are currently not correctly discussed.

24) The authors mention compatibility between SPANTCR and ALICE in their discussion. Yet ALICE is meant to identify expanded TCR clusters within a single repertoire, and not epitope-specific TCRs, thus it is not immediately clear how these tools would be combined.

25) Figure 1D, 5A are missing a color legend.

Authors' response to the reviewers' first round comments

Attached.

Editorial decision letter with reviewers' comments, second round of review

Dear Dr. Heath,

I hope this email finds you well. The reviews are back on your manuscript and I've appended them below.

Please note that not all reviewers were available to re-review the manuscript so we asked a TCR repertoire expert to consult and determine whether the technical concerns of the remaining reviewers were fully addressed. While the expert did find the majority of concerns was met, they did feel that you did not fully address the concerns of Reviewer #3 about comparison to other approaches, and the TCR Specificity predictions from SPANTCR still required more comparison to the output of other approaches.

In short, it seems the rationale for situating SPANTCR as you have among other approaches is not clear enough. I believe this can remedied with some additional text changes to further clarify the actual purpose and conceptual advance of SPANTCR and put into context with other approaches. There needs to be a stronger argument for why a researcher would want to go beyond GLIPH and other options. You address this fragmentarily throughout, but the paper still lacks cohesion in this regard. While the title is spot-on, the Abstract and main text do not follow through in a convincing way. I wholeheartedly agree that the use of SPANTCR to determine "essential" and "superessential" k-mers is its value proposition, but how this relates to more conventional questions of TCR "specificity" and any biological insight that could be gleaned needs further elaboration. Some material in the Discussion does address this, but it would better be placed in the Introduction to frame the whole study.

1) Please clarify the driving problem/question in the Introduction.

You state in the introduction:

"Here we present Scanning PArametrized by Normalized TCR Length (SPAN-TCR) as a tool for extracting structural and chemical insights from groups of antigen-specific TCR sequences in a length-agnostic fashion"

It seem rather that the purpose is "to assess relationships between TCR sets that bind different antigens" as stated further down. It is, however, unclear why this would be a goal for developing a method. The biological driving question is not clearly stated. What is it about TCR sets that bind different antigens that discerning their relationship would answer? Again, material from the Discussion may be more appropriate here in the Introduction.

This should make it clear why you want to advocate a structurally based concept of specificity and avoid sequence-only approaches in favor of your method for identifying essential (and superessential) k-mers.

Beyond just being length agnostic, SPANTCR appears orthogonal to sequence alignment-based conceptualizations of specificity altogether.

2) Please clarify why you are making comparisons to existing techniques and elaborate on this rationale. Would a researcher really use GLIPH to do the same thing you would want to do with SPANTCR? Is this comparison a gut-check to see if SPANTCR is giving you answers you can believe? Or is this a demonstration of where GLIPH falls short and SPANTCR meets an unmet need? Just because both GLIPH and SPANTCR can both be length agnostic doesn't mean they are fit for the same purpose. It would be helpful to the reader to outline the structure of the study in the introduction so the rationale is easier to see.

I look forward to seeing your revised manuscript.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, *Cell Systems*

Reviewer comments:

Reviewer #1: The authors have made all of the changes that I have suggested.

Authors’ response to the editor’s comments

Attached.

Editorial decision letter with reviewers’ comments

Dear Dr. Heath,

I'm very pleased to let you know that the the peer-review process complete, and only a few minor, editorially-guided changes are needed to move forward towards publication.

In addition to the final comments from the reviewers, I've made some suggestions about your manuscript within the "Editorial Notes" section, below. Please consider my editorial suggestions carefully, ask any questions of me that you need, make all warranted changes, and then upload your final files into Editorial Manager.

I'm looking forward to going through these last steps with you. Although we ask that our editorially-guided changes be your primary focus for the moment, you may wish to consult our FAQ (final formatting checks tab) to make the final steps to publication go more smoothly. More technical information can be found below my signature, and please let me know if you have any questions.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Editorial Notes

Transparent Peer Review: Thank you for electing to make your manuscript's peer review process transparent. As part of our approach to Transparent Peer Review, we ask that you add the following sentence to the end of your abstract: "A record of this paper's Transparent Peer Review process is included in the Supplemental Information." Note that this ***doesn't*** count towards your 150 word total!

Also, if you've deposited your work on a preprint server, that's great! Please drop me a quick email with your preprint's DOI and I'll make sure it's properly credited within your Transparent Peer Review record.

Abstract:

Your abstract reads wonderfully but it is unfortunately too long. Please condense to 150 words or less.

Manuscript Text:

We do not support supplementary methods, results, or discussion – please incorporate this into the main text as appropriate. Much of the material looks like it could go into the STAR Methods.

Also:

- House style disallows editorializing within the text (e.g. strikingly, surprisingly, importantly, etc.), especially the Results section. These terms are a distraction and they aren't needed—your excellent observations are certainly impactful enough to stand on their own. Please remove these words and others like them. “Notably” is suitably neutral to use once or twice if absolutely necessary.
- Please only use the word "significantly" in the statistical sense.

Figures and Legends:

Please look over your figures keeping the following in mind:

- When color scales are used, please define them, noting units or indicating "arbitrary units," and specify whether the scale is linear or log.
- Please ensure that every time you have used a graph, you have defined "n's" specifically and listed statistical tests within your figure legend.
- Please ensure that all figures included in your point-by-point response to the reviewers' comments are present within the final version of the paper, either within the main text or within the Supplemental Information.

STAR Methods:

Please convert the methods section to our STAR Methods format. See the STAR Methods guidelines for additional information.

Thank you!

Reviewer comments:

None.

Dear Dr. Heath,

I'm enclosing the comments that reviewers made on your paper, which I hope you will find useful and constructive. As you'll see, they express interest in the study, but they also have a number of criticisms and suggestions. Based on these comments, it seems premature to proceed with the paper in its current form; however, if it's possible to address the concerns raised with additional experiments and/or analysis, we'd be interested in considering a revised version of the manuscript.

As a matter of principle, I usually only invite a revision when I'm reasonably certain that the authors' work will align with the reviewers' concerns and produce a publishable manuscript. In the case of this manuscript, the reviewers and I have make-or-break concerns that can be addressed by:

1. **Providing clearer justification of the working assumptions.**

We have tried to address this concern in multiple ways. First, we have explored more thoroughly, and explained more deeply, the statistical models associated with SPAN-TCR. These revisions also address a concern of referee #3, comments #y7-10. Further, figure 1 has been modified for clarity. In particular, we emphasize that, following length normalization to enable comparisons of different CDR3 regions, we query for common n-mer motifs (here we emphasized 2-mers), and then we query for impact of that 2-mer motif on the sequence diversity of the same-change CDR3 and the other CDR3. This algorithm is a unique approach in the literature that also addresses comment 2 below.

A second justification of the working assumption that we now explore in detail is a demonstration that 2-mers that are identified using the above algorithm as 'essential' or 'super-essential' do, in fact, play significant roles in TCR-pMHC binding. Here we provide significant and compelling new data.

2. **More clearly describing how SPANTCR works and what distinguishes it from competing approaches.**

We have tried to address this concern also in multiple ways. First, we have addressed referee #3's concerns about our comparison with GLIPH – the referee pointed out (correctly) that SPAN-TCR and GLIPH only agreed on around 41% of the prediction. However, this low level of agreement was really an apples to oranges comparison, as the two algorithms do not focus on the same regions of the CDR3 chains. When we focus both algorithms on the same regions of CDR3 changes, the agreement increase to 70% and even 80%. We also emphasize the entropic analysis unique to our technique.

3. **Ensuring fair comparison to competing approaches.**

We are now included a more thorough discussion of competing approaches, in additional to making the comparison with GLIPH a better 'apples-to-apples' type comparison. We elaborate on our discussions of competing approaches, and when we use a concept in our algorithm that has drawn from a competing approach, we have been very careful to cross-reference.

Reviewer #3 provides a particularly lucid and insightful set of comments. Their "major concerns" section serves as an excellent guide for revision - if these can be addressed convincingly with additional analyses and changes to the text and figures, this will strengthen the manuscript. In addition to the concerns I've detailed above, I'd also like to be explicitly clear about an almost philosophical stance that we take at Cell Systems…

- We believe that understanding how approaches fail is fundamentally interesting: it provides critical insight into understanding how they work. We also believe that all approaches do fail and that it's unreasonable, even misleading, to expect otherwise. Accordingly, when papers are transparent and forthright about the limitations and crucial contingencies of their approaches, we consider that to be a great strength, not a weakness.

- We believe that the figures are the scientific backbone of the paper. Currently, it's not possible to understand the manuscript's conceptual advance from figures presented. Similarly, it's not possible to understand where your approach gets its analytical power. These things need to be demonstrated with data and analysis, in the form of figures with their legends or mathematical argumentation, and then supported with explanatory text.

As you address the concerns, it's important that you and I stay on the same page. I'm always happy to talk, either over email or by Zoom, if you'd like feedback about whether your efforts are moving the manuscript in a productive direction. Do note that we generally consider papers through only one major round of revision, so the revised manuscript would be either accepted or rejected based on the next round of comments we receive from the reviewers. If you have any questions or concerns, please let me know. More technical information and advice about resubmission can be found below my signature. Please read it carefully, as it can save substantial time and effort later.

I look forward to seeing your revised manuscript.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Manuscript formatting: Advice and pain points
Should we publish your paper, the next steps will go more smoothly if you keep this advice in

mind. It comes directly from the Cell Systems Editorial Team. We hope that our experience will minimize your frustration.

Transparent Peer Review
As part of the submission process, you were given the option to make a comprehensive record of your manuscript's Transparent Peer Review process public should we accept it for publication. We've recorded your choice, but if you believe you've made it in error, please email us at systems@cell.com. You're also welcome to read more about Transparent Peer Review here, read these FAQs, and email us with questions.

STAR Methods formatting
We ask that the methods section of your revised manuscript be substantively similar to what we'll ultimately publish, should we accept your paper. Please note that the final steps towards publication will be faster and smoother if your revised methods section adheres to our STAR Methods guidelines. Please contact me if you have any questions about STAR Methods.

Deposition of data and code
We require that code and datasets be deposited publicly. If you require an exception to this rule (e.g. your manuscript uses confidential patient data), please email your editor.

When your revisions are complete, submit your revised paper online at: http://cell-systems.edmgr.com/, and include a point-by-point list of the revisions made to address the reviewers' comments. For future reference, note that papers cannot be accepted until any necessary accession numbers are provided. Finally, please note that we generally consider papers through only one major round of revision, so the revised manuscript would be either accepted or rejected based on the next round of comments we receive from the reviewers.

Administrative correspondence:
Ashley Fortier
Editorial Operations Associate, Cell Systems
50 Hampshire Street, 5th Floor
Cambridge, MA 02139
781-663-2251
systems@cell.com

Reviewer #1, C1. : This paper presents a new computational tool to analyze TCRs with different length CDR3 variable regions. Entropy analysis is used to find k-mers that are essential for TCR:pMHC binding. This new approach can provide an overall landscapes of antigen-specific TCR. I would suggest the authors to take following comments into account.

1) The authors should describe more details about how values of different parameters in their algorithm were adopted. For instance, the authors mentioned they applied exponential frequency functions for different databases (VDJdb: 5score, McPAS: 2id.method, PIRD: 3grade). It is not clear how these parameters were determined.

Thank you for the recommendations. We have added text explaining this rationale in the main text

"VDJdb provides a confidence score from 0-3, with approximately 100x as many sequences reported with score=0 as score=3. Thus, we chose a sequence-specific weight of 5 score such that the net contribution of score=0 and score=3 sequences was similar. This score can be adapted for different scenarios."

and methods

"This function is malleable, the exponential terms here were chosen to allow high score values and low score values to collectively contribute similar weights, i.e. there were approximately 100x as many score=0 terms in VDJdb as score=3 terms. The choice of function depends on the nature of the input data, with the overall goal to emphasize the contribution from high quality CDR3s and limit noisy, low confidence data."

2) The authors should provide more mechanistic interpretation on the structural biology or biophysical level about the meaning of the K-mers they identified. For instance, the authors mentioned that " Hydrophilic 2-mers appear more essential than neutral 2-mers (G, A-containing 2-mers). 2-mers with charged amino acids (N, K, Q, R) are most frequently associated with large entropy reductions". It would be helpful to dig into the origin of these discoveries.

We have added significant new data and discussion to address this comment. We took an essential 2-mer ('NN') identified in the CDR3 chains of NLVPMVATV-specific TCRs, and carried out AlphaFold molecular dynamics simulations of 10 relevant TCR-pMHC complexes (in water). We first validated that AlphaFold (which starts from sequence information) could correctly reproduce the crystal structures of 3 published TCR-pMHC complexes to within around a 1Å resolution (Supplemental Fig 3A). For each of the 10 TCR-pMHC complexes, we identified that the NN 2-mer participates in hydrogen bonding (Supplementary Fig 3B, Figure 4D,E), although the details of that hydrogen bonding vary across the 10 systems. This suggests essential 2-mers are likely, in fact, essential to the non-covalent binding motifs that hold the complex together, but also suggests that far more data is needed before more specific chemical insights can be gained. We also have expanded our discussion around these findings.

3) Each figure of the paper contains too much information. When the authors describe their results, they skipped some technique details either in the main text or in the figure captions. As a result, some of these figures are difficult to understand especially for general audiences. Moreover, all the supplemental figures cannot be accessed. The authors should do better job during the revision to

make their manuscript more easily acceptable for readers which background is not computational immunology.

We have tried to simplify and streamline all the figures. Figure 1, in particular, is now intended to just introduce the reader to our strategy, with panel D now focusing on our objective to identify essential k-mers through entropic analysis. Figure 2 panels A and B have been altered to emphasize the connection between our k-mer landscape plots and the Logo plots used in the literature. The supplemental figures and tables have been condensed into a single PDF file.

4) One major assumption of the work is that TCR specificity emerges from hypervariable CDR3a and CDR3b loops, which is reasonable. However, the neighboring germline loops and the MHC protein also contribute remarkably to the binding between TCR and pMHC. One example is the complex formed by A6 TCR and the human T cell leukemia virus-1 Tax11-19 peptide presented by the class I MHC protein HLA-A*0201. The x-ray structure of the complex shows that CDR3a loop makes several strong electrostatic interactions with the HLA-A2 a1 helix. As a result, it was suggested that "when considering the determinants of TCR binding specificity we need to consider the interface in its entirety" in a recent review paper (J Immunol 2017; 199:2203-2213). I would appreciate if the authors could discuss this issue.

This is a difficult challenge in this field which neither our technique nor any others have managed to resolve. We have added the paper suggested (reference 46) in the discussion, we have included the caveat that the CDR3 forms bonds with the HLA outside of the epitope in our limitations, and we have made a point to discuss the role of CDR3-HLA interactions in our molecular dynamics simulation data (Figure 4 D, E, Supplementary Figure 3 B).

5) Finally, about the future application of the tool, in addition to find some global rules of the sequence patterns in CDR3 loops, how will it be practically used to predict the binding specificity of TCRs or design new TCRs, comparing other machine-learning based methods? For instance, in a recent study, antigen specificity of single T cells has been predicted by deep learning algorithm based on their TCR CDR3 regions (Fischer et al., 2020). Some perspectives need to be provided.

We have added references and discussion regarding machine learning tools used in this field (references 61-66). We have also emphasized our belief that the essential k-mers that we identify will be important for predicting TCR binding (“Unlike certain machine learning tools,61-66 we do not try to predict TCR specificity but instead identify k-mers that are likely important for TCR specificity.)

Reviewer #2: The manuscript by Xu et al. describes a novel tool (SPAN-TCR) that integrates length-agnostic CD3 sequence analysis for identification of k-mer sequences that determine antigen specificity of the TCR using entropy-based modelling.
I have read the manuscript with interest, and believe it is highly relevant. As the work is a little outside of my expertise, I hope they help the authors understand how their manuscript is perceived from a less specialized audience. I therefore have a number of questions and suggestions to increase clarity.
Authors analyse all 2mer sequences in publicly available CMV pp65-specific CDR3s (VDJdb database) and identify several strings that are common between TCRs of different MHC-peptide

specificities within the CD3 region. The algorithm is then applied to a novel set of experimentally determined TCRs for CMV pp65 and they determine which sequences are most likely to have the antigen specificities of their published counterparts.
Authors are able to categorise the k-mers into different categories to classify their relevance to antigen specificity. Interestingly, authors then proceed to look at unrelated datasets (SARS-CoV-2) and make a series of interesting observations, one of which is that there are striking similarities between CD3 regions of TCRs that recognise the same antigen, and that the determining k-mers accumulate in the centre of the CD3 region.

My main comment is that it is not clear how the epitope specificity of the TCR was integrated into the workflow overall. Where all of the TCRs MHC-I specific (I assume so)?

Yes, all MHCs analyzed were MHC-I, and this has been noted in the methods ("VDJdb chains were all MHC Class I specific.").

I first thought that in the validation experiment a pool of TCRs was pulled down using a diverse pool of MHC dextramers with a mixture of pp65 epitopes, but I take from the method section that a single peptide epitope, and a single HLA molecule (HLA-A2) was used. Does the original VDJdb data contain A2 TCR sequences specific to NLVPMVATV? Were those TCRs that were identified as close or even identical in sequence also A2 NLVPMVATV TCRs? What are the HLA associations of the SARS-CoV-2 epitopes, and did the TCRs that were closely related belong perhaps to a single 'HLA group'? Please indicate these details in the main text so they are immediately clear to the reader.

Yes, the reviewer has drawn attention to one of the weaknesses of many TCR studies, which is the limited diversity of HLA alleles used. In fact, every NLVPMVATV-specific TCR available in VDJdb was associated with HLA-A02, as NLVPMVATV is likely an HLA-A02-specific peptide. So each TCR that was identified as close or identical was also HLA-matched. This detail has been added to the main text

"… we carried out this analysis only to validate our core algorithm by analyzing VDJdb-reported CMV pp65 CDR3s specific to the epitope NLVPMVATV presented on HLA-A02,…"

"After calculating the difference between these CDR3s and CMV-specific, HLA-A02-specific VDJdb TCRs,…"

and the discussion of this limitation has been expanded in the "limitations" section of the discussion.

"Another biological limitation that affects the entire field of TCR sequence analysis is HLA specificity. Epitopes do not present universally across HLA alleles…"

Regarding the interpretation of the results, is it not true that a TCRs within a specific epitope group also has an implied specificity to the presenting HLA molecule? Thus, the established relation between CDR3 motifs and antigens is more associative, not causal - unless authors consider and exemplify differences between epitope specificities of a single HLA group, in my option.

Yes, HLA specificity is a recurring concern for this study and we have increased discussion of this issue. We have added references 67 and 68 discussing this issue,

“Epitopes do not present universally across HLA alleles and measuring the extent of shared TCR specificity across HLAs is a complementary field of research67, 68…”

Indeed, any relationship between CDR3 motifs and antigens is associative. Causality is difficult if not impossible to establish in such a complex system, and we avoid any claims to causality.

Reviewer #3: In 'Entropic Analysis of Antigen-Specific CDR3 Domains Identifies Essential Binding Motifs Shared by CDR3s with Different Antigen Specificities', the authors present a novel tool and accompanying R package to identify potentially relevant k-mers in epitope-specific T-cell receptor sequences (TCRs). The approach is novel and certainly interesting, however specific assumptions that it builds upon are questionable and not fully addressed in the manuscript. In addition, the advantage compared to existing methods is unclear at this time. Furthermore the results are presented as far more novel and relevant than they actually seem to be. The R package is well documented and seems easy enough to install.

Major comments:

1) The main assumption that SPANTCR builds upon is that the binding orientation of different TCRs with the same pMHC are constrained in their binding orientation, even if the TCR CDR3 length differs. However this does not match with the current literature surrounding TCR-pMHC complexes, where it has been shown that radically different complexes can involve the same pMHC. For example, reference 12 within the manuscript (Gras et al.), describes a TCR binding with a reversed conformation. This is further explored in more recent papers on TCR-pMHC binding, such as Knapp et al, (https://doi.org/10.1371/journal.pcbi.1007338). Thus the prevalent hypothesis is that different TCRs binding with the same pMHC, especially if they differ strongly in sequence (for example by length), have different constraints and are likely to use different amino acid interactions to mediate the recognition/binding. It is only very similar TCR sequences that can be expected to follow similar binding orientations.

The reviewer is correct on this point. We are making an assumption here by necessity – we are developing an informatic approach in the absence of detailed molecular structures of the TCR-pMHC complexes that we are interested in. This is similar, of course, to other approaches, since sequence information is always the most abundant, by far. We have now qualified this statement, to make it clear that we understand the limitations of this assumption, as the referee has pointed out (1st paragraph of discussion).

We have spent a significant amount of time and effort exploring this further. We performed MD simulations to explore the role of our essential k-mers in binding. We found that the same k-mer in the same position of different TCRs has important, context-dependent binding interactions such as hydrogen bonds. We have added this discussion to Figure 4 D, E, and Supplementary Figure 3B.

We also agree that for very similar TCR sequences, the binding orientations are likely to be similar, for which we have added text to the discussion:

> “While these steps do not eliminate the variability of TCR-pMHC binding orientation from the model, essential k-mers discovered through SPAN-TCR appear to reflect groups of TCRs where the binding orientation is most similar.”

One of our objectives with this work was to identify what we believe are ‘effectively’ similar TCR sequences. It is also relevant to note that in Gras et al, the reversed binding conformation did not exhibit effective TCR signaling. Thus, such exotic angles may be less likely to be validated and reported in literature such as VDJdb, which was our primary source of data.

2) SPANTCR uses an interpolation across 100 bins to find patterns that are length-invariant, from beginning to end. However it can be questioned if this approach is appropriate when considering the biochemical nature of amino acid interactions. A TCR can only interact with a pMHC with a limited number of CDR residues. Through interpolation, SPANTCR seems to assume that the residues within the binding interface can be compressed if the TCR CDR sequence is longer. Or, in other words, that the relative positioning with regards to the full CDR length is what matters, and not the absolute positioning. For example, imagine a TCR with two interacting regions that are separated by two residues, SPANTCR assumes that a TCR with twice the length would position these interacting regions separated by four residues. This is an extreme example, but the manuscript currently does not address this concern.

We have added our motivation for using a 100-bin interpolation rather than a sequence-matching method in the introduction

> “The common thread between these methods is a sequence-based framework where the single residue is the basic unit. However, structural analysis and molecular simulation demonstrates variability and “jitter” between amino acid residue positions in TCR/pMHC binding. While sequence alignment tools can be useful, they also employ a rigid representation of peptides in sequence.”

And discussion sections.

> “SPAN-TCR’s niche lies between CDR3 sequence alignment strategies54, which can oversimplify a complex binding interface that includes CDR3 interactions with the HLA but not the epitope as well as non-CDR3 interactions55, 56, and full molecular simulations9, 57, 58, which remain prohibitively expensive59.”

For the counterexample raised, SPAN-TCR would attempt to capture the possibility that due to variance in bond angles and intermolecular interaction lengths (H-bond, hydrophobic, etc.), it would be possible for two interacting regions to be separated by different numbers of residues and still recognize the same epitope. It is likely that our method fails for extreme cases as the reviewer proposes.

We have searched the TCR-pMHC structural literature to find instances where different TCRs are reported to bind the same epitope to explore this idea further. There is very limited structural data compared to sequence data, but in one instance a set of TCRs (structures 3HG1, 3QDG, 3QDM, binding epitope ELAGIGLTV presented on HLA-A*02:01) form hydrogen bonds with the same residue on the epitope using different residues of the CDR3 (6th/13, 8th/14, 8th/15). The CDR3s are of different lengths, and the hydrogen-forming residue of the longest CDR3 is 2 spots further in the sequence than the shortest. The 6th out of 13 and 8th out of 15th residue are more likely to have the same role when considering the overall length of the CDR3, which represents the type of interaction we hope to capture using SPAN-TCR. Unfortunately, the exact amino acids are different and the literature was too limited to find a direct comparison of the same k-mer in different length CDR3s. Given the limited structural data available, we performed our MD simulations described above to address this shortcoming in the field.

3) The novelty of SPANTCR is presented as a tool 'that seeks to identify features shared between sets of TCRs sharing the same pMHC but with variable length CDR3s, or between TCR sets that bind different antigens'. With regards to the former, the length-agnostic nature, almost every common pMHC-specific TCR analysis is length agnostic: GLIPH (both 1 and 2) search for motifs within the variable CDR sequences. TCRdist (orginal and 3) use an alignment method to bridge different length CDRs. Similarly other methods that have been published to analyse (and often build models for) pMHC-specific TCR utilise solutions that transform the different lengths into something comparable, through either padding, alignment or motif detection.

We value these methods and agree that this phrase was insufficient to distinguish the methods. We attempted to describe all the aforementioned methods in the previous sentence in the introduction and this has been clarified.

“These tools utilize protein sequence alignment (TCRdist),27, 28 or incorporate features such as amino-acid chemical similarity,29-32 shared-motif identification (GLIPH),33, 34 and machine learning techniques35.”

We believe our method is distinct in that it is not restricted to a residue-by-residue comparison. We allow for k-mers to be offset, to accommodate concepts like binding angle and bond length variability. We have changed the wording to reflect this.

“However, structural analysis and molecular simulation demonstrates variability and “jitter” between amino acid residue positions in TCR/pMHC binding. While sequence alignment tools can be useful, they also employ a rigid representation of peptides in sequence.”

One area where our “length agnosticism” is important is in our graphical presentation of TCR landscapes, where Logo plots are still the norm. For CDR3s of the same length, our strategy exactly reproduces Logo plots, but we are able to incorporate more of the diversity of CDR3 sequences with mismatched lengths on the same plot. We have attempted to clarify Figure 2A to reflect this application.

4) SPANTCR usually uses 2-mer motifs. However the motivation behind 2-mer motifs is unclear, especially as most other related methods use longer motifs. The first section of the results states

'Specifically, 2-mer motifs are common in TCR CDR3 data sets'. It is unclear what is meant by 'common' in this section, or how it relates to the selection of 2-mers as the primary analysis point. In addition, the usage of 2-mers is not consistent. The comparison with GLIPH uses 4-mers, and the GXG motif found for the SEHDY epitope group is a 3-mer.

Apologies for the confusion, we have clarified this further in the introduction

> "Although k-mers can be of variable length, we mostly focus on 2-mers, since longer motifs are much less common. For example, unique 3-mers are found at ~1:20 the frequency of 2-mers."

We analyzed 3-mers and 4-mers in the supplement to demonstrate the possibility and to align with GLIPH motifs, which are typically longer. This has been emphasized in the results

> "GLIPH2 is a powerful method to extract previously unknown antigen-specific TCRs by identifying long, variable-length strings of residues shared by many TCRs. SPAN-TCR utilizes a complementary approach, whereby putative or known antigen-specific TCRs are analyzed to identify smaller, fixed strings of residues important for TCR binding. We compared the two methods by first using GLIPH2 to identify amino acid subsequences of note. We then applied SPAN-TCR using longer 4-mers to determine the locations and compositions of these subsequences (Fig. 2C, Sup. Table 1)."

The GXG motif mentioned in the SEHDY epitope group can be visualized by a 3-mer as the reviewer says, but since SPAN-TCR specifies k-mers by their location, the frequent observation of GX followed by XG 2-mers identified the pattern, not the 3-mer GXG. This has been emphasized in the text

"Finally, by plotting the composition of all TCRs containing YE in the position range from 0.7-0.8 (Fig. 5F), we identified the frequent appearance of GXG motifs at the center as GX 2-mers followed by XG 2-mers"

5) Based on the description of the SPANTCR method, the entropy of 2-mer distributions at each position is calculated against the background that all amino acids would occur at equal frequency. However it is well known that amino acid usage in TCR CDR3 regions is not uniform, and that some amino acids occur more or less (and likely even specific 2-mers). It would make sense to calculate the entropy against the observed amino acid usage. Indeed, in several results, SPANTCR identifies as called by the authors "common motifs", i.e. 2-mers that occur frequently and are therefore uninteresting. Adopting an alternative entropy calculation might be more appropriate.

The primary constraint to k-mer analysis was the absolute number of available sequences. We agree that for rare combinations of amino acids, it might be descriptively interesting to note that they appear more or less than expected, but without sufficient numbers of sequences the entropy calculations were not comparable. Since the data sets available were not large enough to make that comparison, we considered amino acid usage as a variable (Sup. Fig. 1C).

We also apologize if our wording implied that some amino acids are uninteresting due to their "common-ness". We believe that even common amino acids such as Glycine play important roles in binding, and that the best way to determine the “interesting-ness” of a k-mer is by measuring their apparent role in binding, which we did using our entropy metrics. We have added this to the Discussion section (paragraph 2).

We will likely add alternative entropy metrics as options in the package in the future, but for the moment we prefer to use only a single entropy metric for simplicity, and have added discussion of other entropy calculation options with reference 47.

> “and other types of entropy measurements may reveal further insights47.”

6) There is little comparison to existing methods with regards to results. While SPANTCR is clearly different in its approach, any advantages of its use compared to, for example, GLIPH or tcrdist are unclear.

We have added discussion about our method and others, and clarified the comparison between GLIPH and SPANTCR (see comment 7). We have also further expanded on the advantages and limitations of our strategy, especially to sequence-based strategies in general. We believe our primary advantage is the way we have performed entropic analysis using our framework, and since there are not equivalent methods in the literature we have expanded our discussion of the value generated in the discussion (Discussion paragraph 2)

> “A distinguishing feature of this work is the use of informational entropy to distinguish between the distinct concepts of enriched k-mers and essential (or super-essential) k-mers.”

7) The only comparison to GLIPH is a check of the overlap between found common 4-mers and the GLIPH output. As this derived from the same set of TCR sequences, one would indeed expect an overlap in two methods attempting to quantify the presents of specific amino acid sequences. However while it is claimed that the 'majority' supposedly overlap, a quick count shows that more than 80 out of 135 do not (41% overlap). Thus less than half overlap, and not a majority. In addition, the results mention TCRMatch and a potential comparison, but no results are shown.

The reviewer is correct. GLIPH does not incorporate the beginning and end of the CDR3 into its analysis, so the comparison we described was inaccurate. We apologize for the confusion, to fairly compare with GLIPH, we only considered k-mers that were found in the middle 50% of the CDR3.

> “After the most frequent 4-mers (>0.5% abundance) were identified using SPAN-TCR between relative positions 0.25 and 0.75,…”

Within this range, 46/66 (70% overlap) significant k-mers overlapped with GLIPH motifs. We briefly mentioned this constraint but our text was too ambiguous – which we have now rectified. This proportion increases to 42/51 (80%) when the middle 40% of the CDR3 is considered, which is more aligned with GLIPH’s output.

We mentioned TCRMatch as another valuable tool available in this field, but did not mention a potential comparison as comparing TCRs is not the primary goal of SPAN-TCR. To our knowledge, the closest analogue or “competing approach” was GLIPH2 to identify motifs as we identify essential k-mers, and we chose to initiate a collaboration between our two methods as coauthors rather than “compete”. Specifically, one of the primary differences is that GLIPH2 identifies its motifs based on frequency, whereas SPAN-TCR performs an additional entropic analysis on the k-mers to determine if they are “essential”. This difference has been expanded on in the discussion (paragraphs 1, 2).

8) A second validation comes in the next section when a single TCR pair from the dextramer library and the VDJdb is analysed with the Levenshtein/BLOSUM metric. This is only a single example, which could have been cherry-picked. For validation purposes, the full repertoire should be considered. In addition, a more comprehensive comparison with a 'negative' data set would be appropriate to quantify the false positive rate.

We have clarified the explanation of this data with additional comparisons as requested and more detailed descriptions. Our intent is to clarify the types of differences that this method resolves by plotting the contrast between Levenshtein and SPAN-TCR methods. We also expect that many clones will not be too similar to the public clones reported in VDJdb, as they will be private patient-specific clones. We are using the full repertoire of this experiment, which contains many data points in the upper right quadrant that are pulled down in the experiment but are not similar to any sequences reported in VDJdb, which are likely the patient’s private repertoire. We add text to explain this data further:

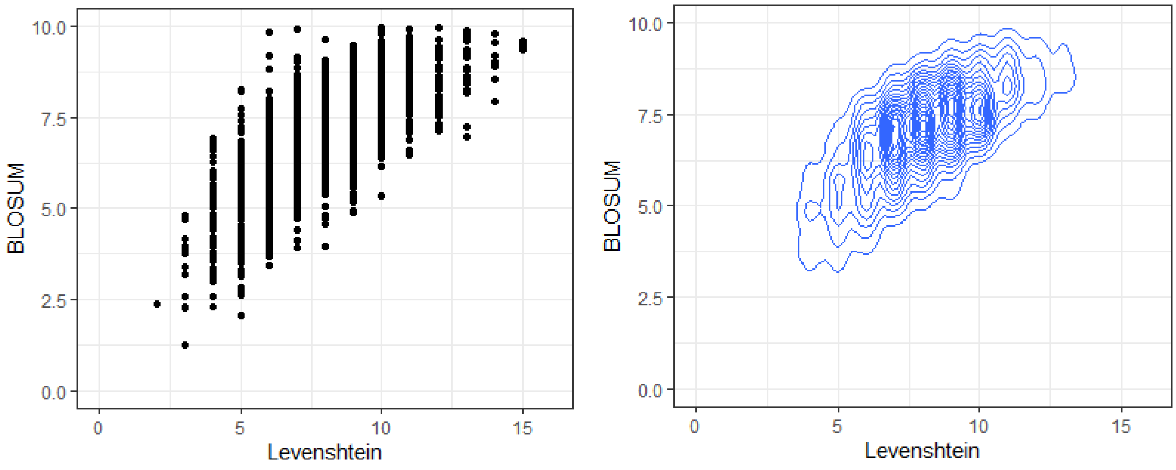
> “Of the entire TCR repertoire, most sequences were not similar to any sequences found in VDJdb, which indicates that they were either nonspecifically captured, or they are patient-specific TCRs unreported in VDJdb”

We draw attention to the single TCR pair example to show the difference between Levenshtein distance and our metrics, where there are quite a few close matches including an exact match (at 0,0) that do not exist in the EBV comparison, which validates the sequenced cells as enriched in CMV-specific TCRs. We clarify that the single example chosen is meant as an example only:

> “As an example, for two highlighted putative pp65 TCRs,”

As an additional “negative” data set, we have performed the analysis using available 10X data and are including it in the supplement. We found that the 10X dataset had no direct overlaps with the VDJdb CMV TCRs and a lower proportion of closer matches. Here we have plotted the 10X comparison for all sequences (point and density to show the infrequency of the closest matches observed). We will include this in Sup. Fig. 2 and have added text:

> “when comparing the experimental results to known TCRs specific to a different antigen or a negative control set of TCRs”



9) A main selling point of the paper is that SPANTCR identifies 'motifs shared by CDR3s with different antigen specificities'. However this finding is entirely reliant on the results presented within the MIRA data, in particular in the [SEHDY] and [AFPFT] epitope groups. However these epitope groups were already listed within the MIRA data, and these groups are known as being linked to the same set of TCRs. Indeed for most, the database does not even distinguish individual preferences. Therefore these are sets of TCRs that all share the same somewhat-degenerate epitope specificity. Any other of the aforementioned TCR analysis tools would equally pick up the same common patterns, simply due to the heavy sharing in TCRs between these epitopes.

We apologize for the confusion, we used SEHDY and AFPFT as our comparisons, but as observed in Fig. 5D, there are many other epitope groups which have YE k-mers at a significant frequency (>100), including ~7 with >500 frequency. We apologize if we gave the impression that our data was entirely reliant on this single result, it was merely a case study. We have clarified the language around this example.

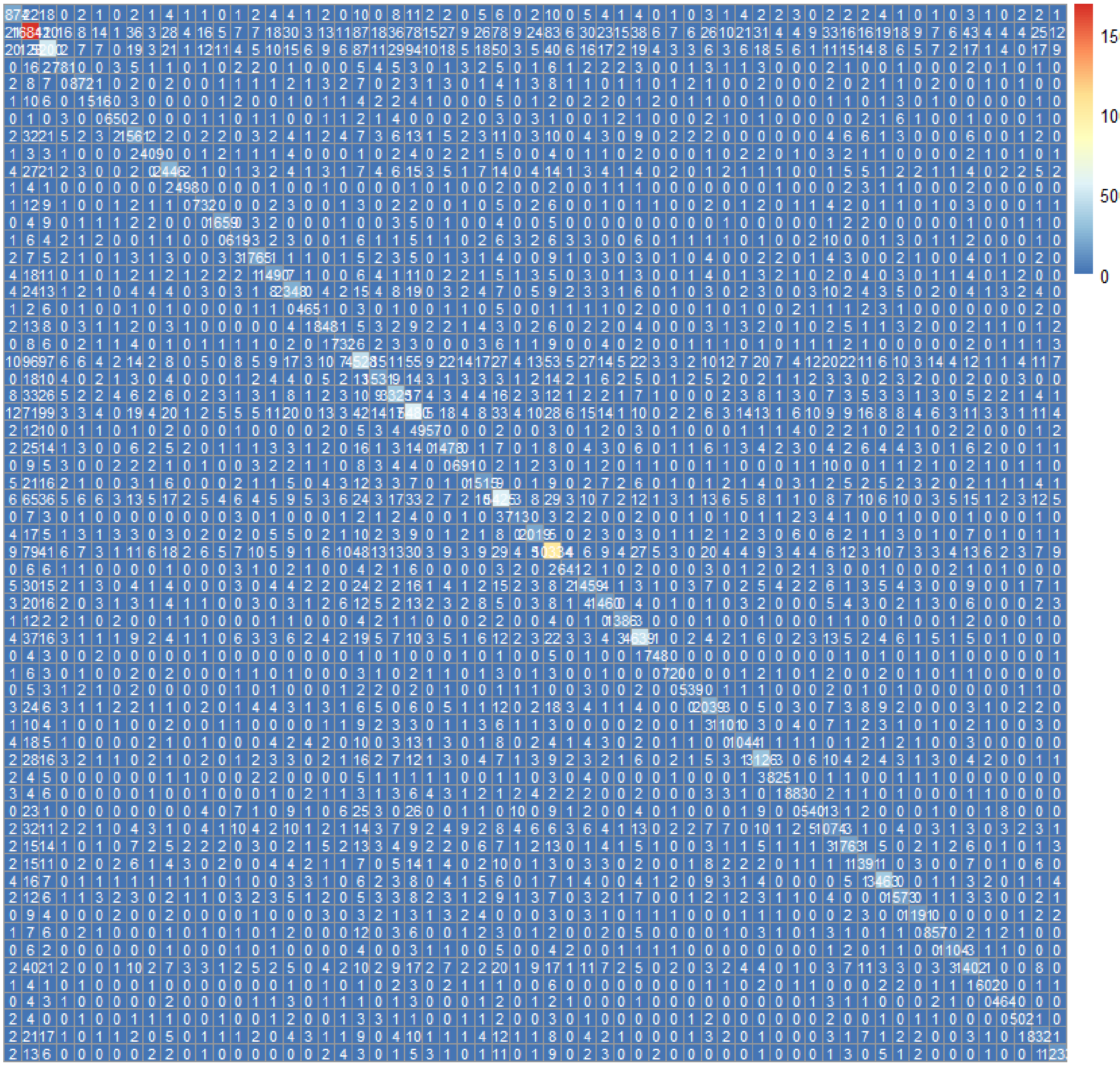
“As a case study, we explored the 2-mer YE in Fig 5C because it has both high frequency and a large entropic effect”

Additionally, we have checked the number of overlapping TCR CDR3 sequences between SEHDY and AFPFT and there are only 7 overlaps out of 3463 and 5200 sequences, respectively. We have added text discussing the very low degree of CDR3 overlap in the MIRA dataset:

“However, in MIRA, there was very little sharing of TCRs between epitope groups (Sup. Fig. 6D, E).”

There appears to be a miscommunication, and there are multiple MIRA data sets which may be leading to the confusion. To avoid confusion around this topic, we have added an additional supplementary figure 6 addressing this point. Below we have attached the overlap matrix of CDR3 sequences between all of the major MIRA groups we studied, the largest value of overlapping sequences was 128 out of 16841/5200 sequences. The diagonal is the self-comparison, and almost all off-diagonal boxes are near zero. The average number of overlapping sequences between any two

MIRA groups was only 3. There was very little overlap or sharing between the MIRA group TCRs. This is critical to establish to address further comments below. A figure expressing the overlap on a log scale is included in Sup. Fig. 6.

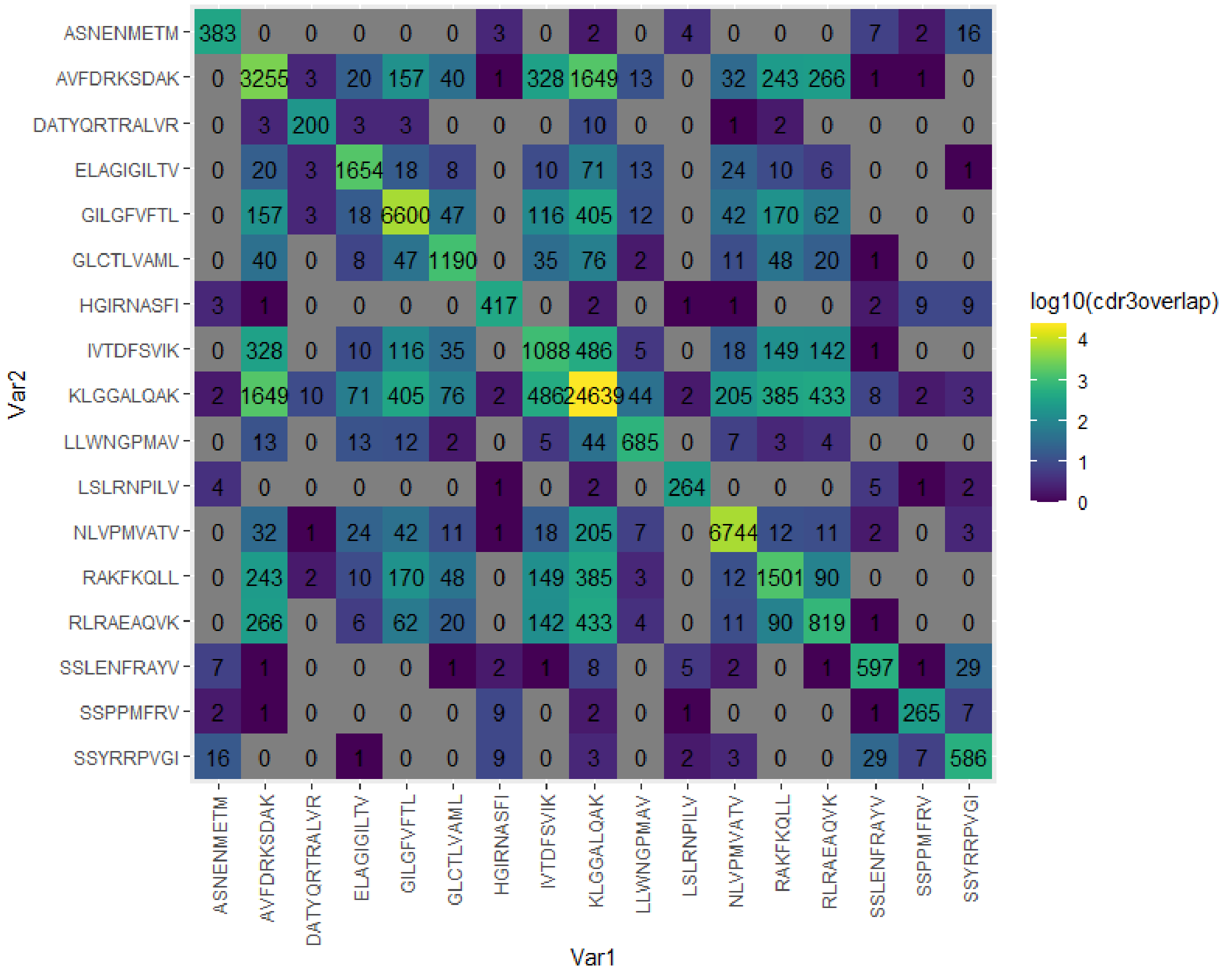


10) Figure 6 presents a comparison between CDR3 profiles and epitope distance where a very weak but reported significant relationship seems to have been found. However these this analysis compared all epitopes versus all other epitopes, so for VDJdb this would be 17 x 16 comparison. This 17x16 data set is then subjected to correlation analysis. But this correlation analysis is inflated, as the number of samples has been artificially extended from 17 to 17x16. Thus the observations within this analysis cannot be considered as independent. The resulting significant P-value is likely the result of this inflated number of observations, as the original data set would not have the power to establish this weak relationship.

The reviewer raises interesting points. The short response is that the entropy profile contains up to 400*100=40,000 values (number of k-mers * number of positions). Therefore, knowing the entropy reduction profile difference between two epitope groups A+B, and B+C, gives almost no information about the difference between A+C, and the 17x16 values are effectively independent. We will go into deep detail here, as we wished to be thorough in addressing this comment and it is quite interesting to explore the data this deeply.

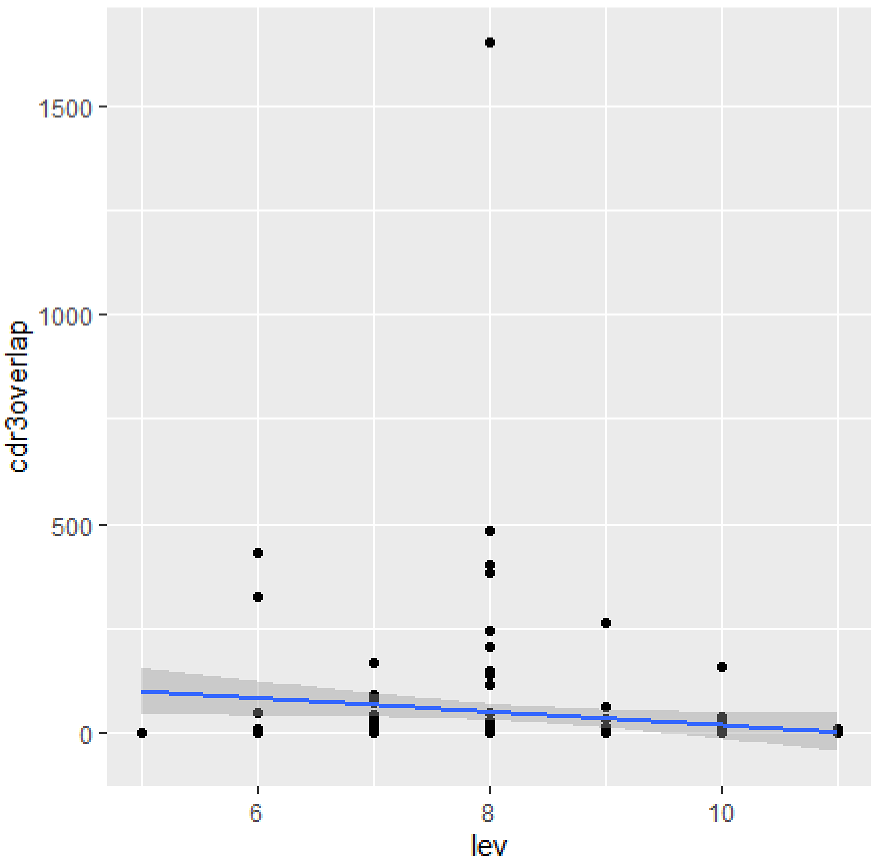
A similar issue exists within the same comparison for the MIRA data set, but this is enhanced by the shared TCRs in comment #9.

We established above that the MIRA data analyzed here has no significant shared TCRs (comment 9). However, VDJdb does have significant overlap. The overlap plot is shown below with a logarithmic color scheme to show the overlap, and the average overlap between epitopes was 44.85 TCRs (compared to 3.12 for MIRA). In VDJdb, 18/136 pairs of epitopes had at least 50 shared TCRs, and in MIRA only 8/1830 pairs had at least 50 shared TCRs. Shared TCRs are clearly a greater issue in VDJdb than MIRA, and it is beyond the scope of this study to make claims to whether shared TCRs are truly cross-reactive, or merely non-specific and reported as specific to multiple epitopes by VDJdb contributors. We are bounded by the quality of the database (VDJdb) for this, and we raise this point as a limitation in the Discussion. Below, we will make the case that the entropy profile correlation shown in Figure 6 holds for data with either shared or non-shared TCRs. We first establish that the shared TCRs are not responsible for this correlation.



	ASNENMETM	AVFDRKSDAK	DATYQRTRALVR	ELAGIGILTV	GILGFVFTL	GLCTLVAML	HGIRNASFI	IVTDFSVIK	KLGGALQAK	LLWNGPMAV	LSLRNPILV	NLVPMVATV	RAKFKQLL	RLRAEAQVK	SSLENFRAYV	SSPPMFRV	SSYRRPVGI
ASNENMETM	383	0	0	0	0	0	3	0	2	0	4	0	0	0	7	2	16
AVFDRKSDAK	0	3255	3	20	157	40	1	328	1649	13	0	32	243	266	1	1	0
DATYQRTRALVR	0	3	200	3	3	0	0	0	10	0	0	1	2	0	0	0	0
ELAGIGILTV	0	20	3	1654	18	8	0	10	71	13	0	24	10	6	0	0	1
GILGFVFTL	0	157	3	18	6600	47	0	116	405	12	0	42	170	62	0	0	0
GLCTLVAML	0	40	0	8	47	1190	0	35	76	2	0	11	48	20	1	0	0
HGIRNASFI	3	1	0	0	0	0	417	0	2	0	1	1	0	0	2	9	9
IVTDFSVIK	0	328	0	10	116	35	0	1088	486	5	0	18	149	142	1	0	0
KLGGALQAK	2	1649	10	71	405	76	2	486	24639	44	2	205	385	433	8	2	3
LLWNGPMAV	0	13	0	13	12	2	0	5	44	685	0	7	3	4	0	0	0
LSLRNPILV	4	0	0	0	0	0	1	0	2	0	264	0	0	0	5	1	2
NLVPMVATV	0	32	1	24	42	11	1	18	205	7	0	6744	12	11	2	0	3
RAKFKQLL	0	243	2	10	170	48	0	149	385	3	0	12	1501	90	0	0	0
RLRAEAQVK	0	266	0	6	62	20	0	142	433	4	0	11	90	819	1	0	0
SSLENFRAYV	7	1	0	0	0	1	2	1	8	0	5	2	0	1	597	1	29
SSPPMFRV	2	1	0	0	0	0	9	0	2	0	1	0	0	0	1	265	7
SSYRRPVGI	16	0	0	1	0	0	9	0	3	0	2	3	0	0	29	7	586

We found that the Levenshtein distance between epitopes in VDJdb was not significantly correlated with the amount of shared TCRs (plot below, R^2=0.015, p=0.151). Two epitopes that were more similar by Levenshtein distance did not share more TCRs, so this effect is not contributing to the Levenshtein/entropy reduction profile association we report.

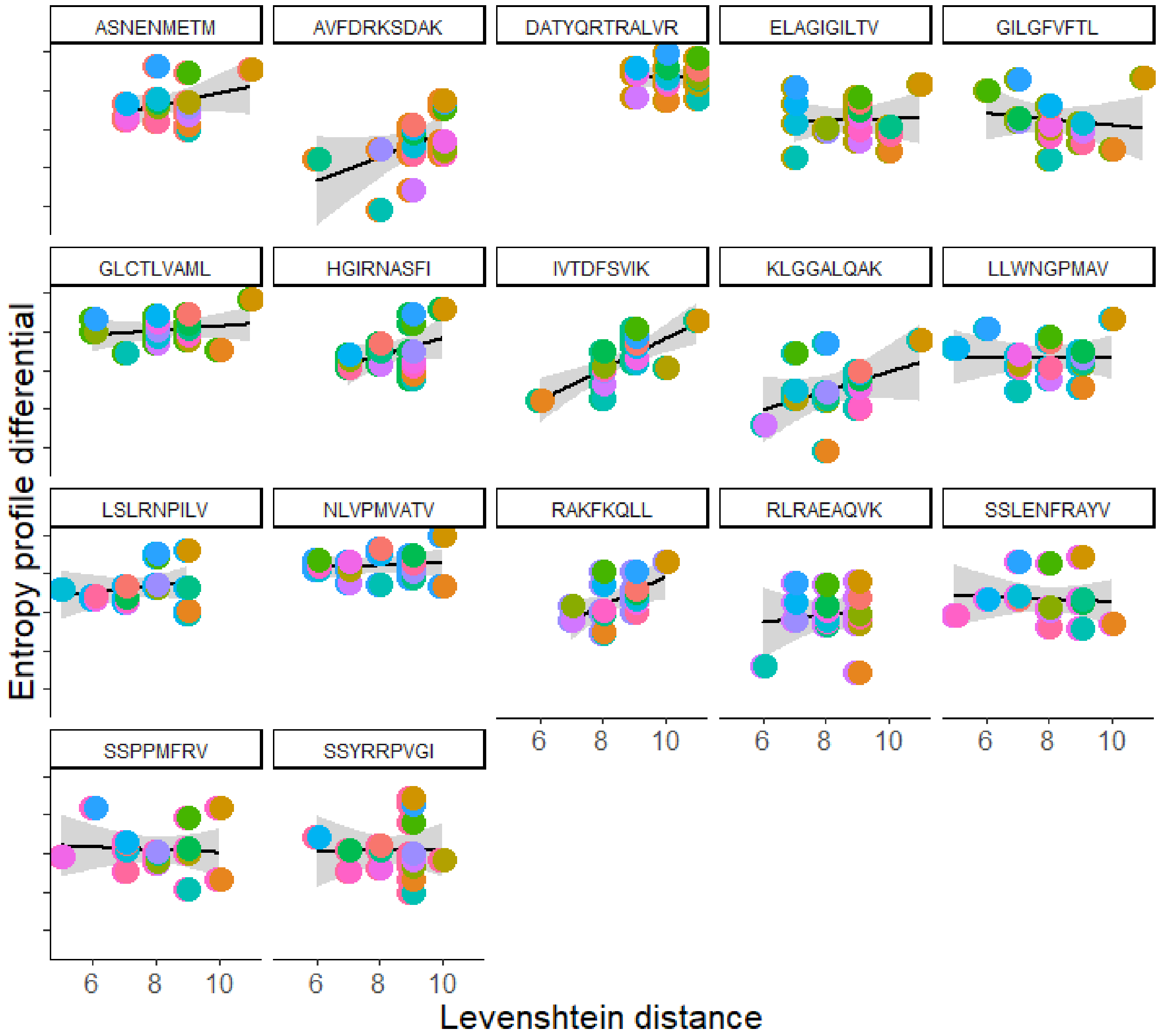


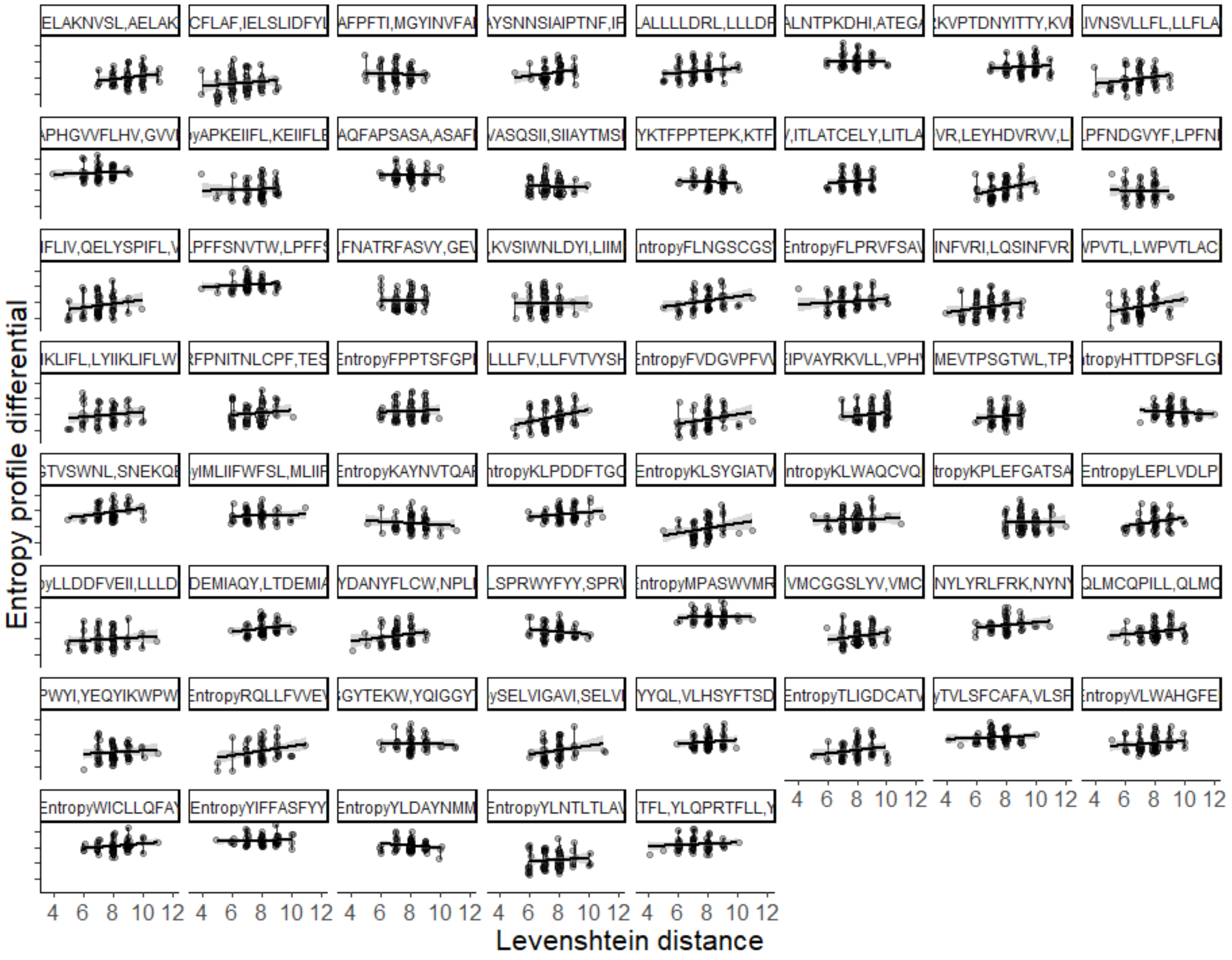
We add text discussing these points:

"This relationship was maintained despite significant TCR sharing between epitopes, which was not correlated with the Levenshtein distance between epitopes (Sup. Fig. 6B, C)."

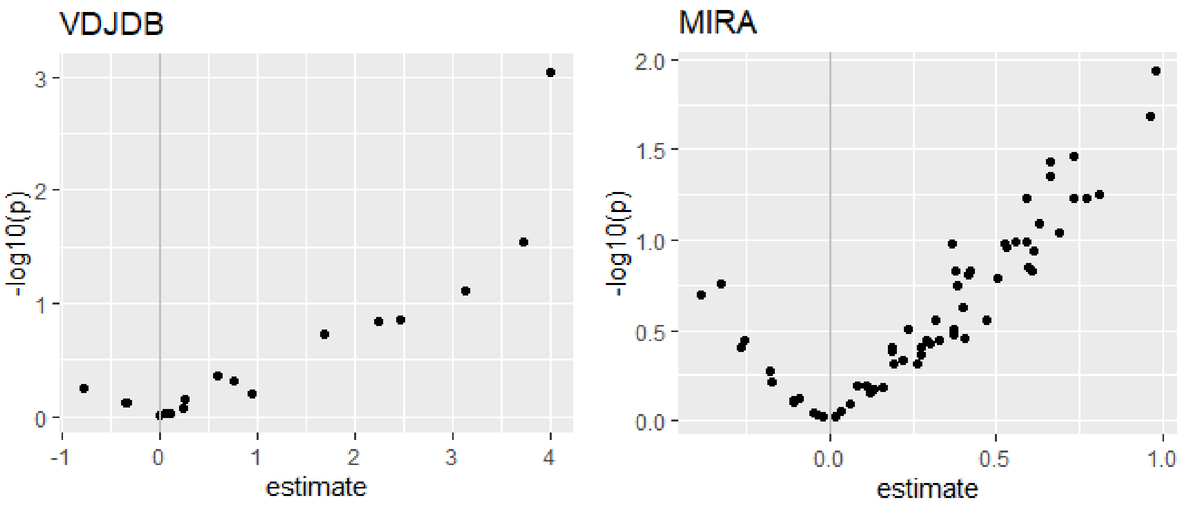
Next, we address the concern that statistical significance is arising due to the expansion of data points from 17 to 17x16 for VDJdb (and 61 to 61x60 for MIRA). First, as a small point, the total number of data points on the plot is not 17x16, but 17_c2=17x16/2=136, as we are not double-counting pairs of comparisons. Next, we analyzed the entropy reduction profile similarity between each epitope group and all others separately. This resulted in 17 (VDJdb) and 61 (MIRA) separate plots showing the individual correlations of Levenshtein distance and entropy profile reduction for each epitope (shown below). We observe significant heterogeneity between the Levenshtein distance to other epitopes and the entropy profile similarity between the TCR sets. We include text supporting this figure:

"The correlation of epitope Levenshtein distance and entropy profile difference was also observed at the individual epitope group level (Sup. Fig. 6F)."


ASNENMETM
AVFDRKSDAK
DATYQRTRALVR
ELAGIGILTV
GILGFVFTL
GLCTLVAML
HGIRNASFI
IVTDFSVIK
KLGGALQAK
LLWNGPMAV
LSLRNPILV
NLVPMVATV
RAKFKQLL
RLRAEAQVK
SSLENFRAYV
SSPPMFRV
SSYRRPVGI
Entropy profile differential
6
8
10
Levenshtein distance



However, the majority of the correlation plots have positive correlation between Levenshtein distance and entropy profile reduction, with volcano plots for VDJdb and MIRA shown below plotting the coefficient of correlation (slope) vs the p value of the linear model. We see that a majority of the coefficients are positive-signed. Thus, after calculating the relationship between Levenshtein distance and entropy reduction profile similarity for each epitope separately, we find a similar correlation as reported in the main text. T-tests performed on the coefficient of correlation for each plot showed a statistically significant association between Levenshtein distance and entropy profile slope without artificially increasing the sample numbers, addressing the comment raised by the reviewer (VDJdb: n=17, p=0.007. MIRA: n=61, p=5.6e-10). This analysis is added to the Supplementary Figure 6.



11) SPANTCR identifies essential 2-mers by identifying those that constrain the entropy throughout the CDR3 sequence (even across beta - alpha chains). However the CDR3 formation is in itself constrained by the V and J genes. Therefore one could assume that the essential 2-mers are simply representations of enriched V/J usage, especially in the case of alpha chains (where diversity beyond V/J usage is more limited) and subsequent superessential motifs.

This is an interesting point, as many researchers have found that V/J genes alone provide much of the needed information. For example, Nanostring provides a tool for spatial TCR sequencing that is focused on V/J gene usage, rather than full CDR3. Here we take full advantage of the CDR3 sequence data available to study the TCRs, giving us the ability to capture recombination events that produce essential k-mers within V/J combinations that do not typically carry the k-mer. We agree with the idea that often researchers can compare TCR clones with V/J gene alone to a degree of acuracy, but nonetheless believe it is important to establish technology and methods that considers full CDR3 sequences, as many problems may require subtle changes in CDR3 to be resolved, such as neoantigen recognition, or TCR recognition of virus mutational variants. We have added a statement in the introduction touching on the diversity of V/J gene versus full CDR3 sequences.

“In addition to considerable V- and J-gene diversity, the insertion and deletion of nucleotides results in the remarkable heterogeneity of CDR3 length and peptide composition.”

Minor comments:

12) Many of the references with the introduction are somewhat outdated or out-of-place. For example, a discussion on the CDR3 diversity to the same pMHC targets in VDJdb refers to the review of Miho at al., where this is not explicitly discussed.

Agreed, the Miho reference is a thorough review paper discussing diversity in TCR, the reference pertained to measuring CDR3 diversity in general, not the specific diversity of TCRs binding the same target, which is instead demonstrated by Fig. 1B showing the wide range of TCR CDR3 lengths observed, including CMV-specific CDR3s. The reference has been moved to the appropriate location (Reference 27).

13) Given that SPANTCR can be used for extracting structural insights and is based on the hypothesis of similar structures, it may be appropriate to contrast the method theoretically with homology modelling, such as for example Milighetti et al. (https://doi.org/10.3389/fphys.2021.730908) or Lanzarotti et al (https://doi.org/10.3389/fimmu.2019.02080).

Thank you for these references, we will include these in the limitations sections of the discussion (References 58/59), as the structural approach is a viable alternative to SPAN-TCR and other sequence-based methods.

14) In addition, as SPANTCR aims to identify patterns across epitopes, it might be useful to also contrast with de novo TCR-epitope interaction models, such as Moris et al (https://doi.org/10.1093/bib/bbaa318), Weber et al (https://doi.org/10.1093/bioinformatics/btab294), and Lu et al (https://doi.org/10.1038/s42256-021-00383-2)

Thanks, these references will be added in the machine learning/pattern recognition discussion section (References 64-66).

15) VDJdb is commonly written with 'db' as lower case.

Corrected.

16) The full name of McPAS is McPAS-TCR.

Corrected.

17) The text/figures are not consistent in their naming of TBAdb and PIRD, which are used interchangeably.

Consolidated to TBAdb.

18) It is unclear what the stacked amino acids represent in figure 1B.

They represent the frequency of amino acids used throughout the entire CDR3, equivalent to compressing a Logo plot across all positions. Figure caption is edited for clarity.

“The relative frequency of amino acids across the entire CDR3 is shown on the Logo plot (right).”

19) The first results sections includes the statement 'shows the degeneracy at the CDR3 N-terminus'

to denote a fairly consistent amino acid motif in this region. The term 'degeneracy' here does not seem fully accurate, as it would commonly be assumed to be the opposite int he context of sequence motifs.

Changed to “more conserved”.

“The blue highlighted region shows the more conserved sequences at the CDR3 N-terminus, consisting primarily of hydrophobic 2-mers (CA, CI).”

20) The results report that its 'general findings are consistent across all three databases'. This is not unexpected as there is large redundancy in these dataset as they are all derived from much of the same public resources.
Agreed, we treat this as a “sanity check”.

21) The methods section reads that data is derived from 'three TCR databases, VDJdb, McPAS, and PIRD, and the MIRA database'. These are four TCR databases.
Thanks for catching, MIRA was of course added upon internal revisions.

“Publicly available data from four TCR databases, VDJdb, McPAS-TCR, and TBAdb, and the MIRA database44 were used”

22) Supplemental figure 2E features an extremely tructated TRA sequence, 'SSGNQFYF'.
Apologies, we rely on QC from our databases and didn’t make any assumptions about data found there, but this is clearly not a standard sequence and it will be removed.

23) Supplemental figure 5A shows poor clustering of motifs for the same epitope across different databases. As SPANTCR aims to identify epitope-specific motifs, especially given the overlap between databases, one would expect strong clustering for each epitope. These findings are currently not correctly discussed.
These discrepancies are the result of the mismatch between these large data sets. For instance, VDJdb had approximately 1000 unique, distinct CDR3s specific to CMV compared to TBAdb at the time of analysis, which causes the drift between data set comparisons. We describe this difference in more detail in the results.

“These general findings were generally consistent across all three databases (VDJdb, McPAS-TCR, TBAdb) (Sup. Fig. 5) with differences emerging due to distinct CDR3s contained in one data set but not the others.”

24) The authors mention compatibility between SPANTCR and ALICE in their discussion. Yet ALICE is meant to identify expanded TCR clusters within a single repertoire, and not epitope-specific TCRs, thus it is not immediately clear how these tools would be combined.
ALICE can be applied to propose putative epitope specific TCRs, based on the selective expansion of certain TCR clusters. TCRs belonging to these clusters can be labeled and tracked through analysis by SPAN-TCR in a complementary fashion.

25) Figure 1D, 5A are missing a color legend.

Dec 7, 2022

Cell Systems

Re: Second revision of Manuscript Number: CELL-SYSTEMS-D-21-00612. *Entropic Analysis of Antigen-Specific CDR3 Domains Identifies Essential Binding Motifs Shared by CDR3s with Different Antigen Specificities*

Dear Dr. Andrianantoandro,

Thank you for your efforts with our revision, especially in finding the supplemental reviewer. We are pleased to submit our revised manuscript, where we have significantly changed the summary and introduction of the manuscript to highlight the added value of SPAN-TCR via the entropic analysis. We have also added passages to clarify the differences and use cases between our technique and related methods such as GLIPH, and have included an outline of the structure of our study in the introduction.

Sincerely,

Jim Heath, President Institute for Systems Biology, jheath@isbscience.org 310 383 8199

Alex Xu alex.m.xu@gmail.com