Analyzing the space of TCRs using optimal transport

Frederick “Erick” Matsen
Fred Hutchinson Cancer Research Center
http://matsen.group/
@ematsen

Thank you all for your work on AIRR!

Today's outline

- Main topic: optimal transport for TCRs (feedback please)
- An update on some other recent work 🍻

Goal: *define and learn from a rich space of TCRs*

Branden (Olson) Steele 🎓:

- Phil Bradley (Fred Hutch)
- Paul Thomas and Stefan Schattgen (St. Jude's)

I'll take these for granted at an AIRR meeting

- We have a lot of TCR data
- This data has rich structure
- We would like to learn about immunology from these data

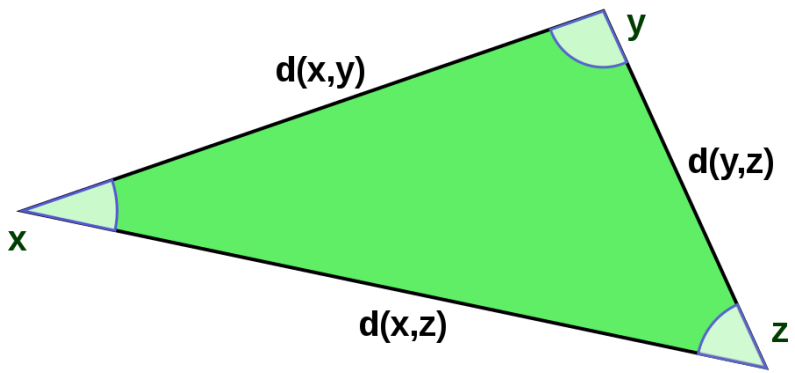
There are many ways of analyzing TCR sequence data, but yet…

How do we get “TCR goggles”?

1. Define a space in which the TCRs live
2. Do analysis / comparison in that space

Define a space in which the TCRs live

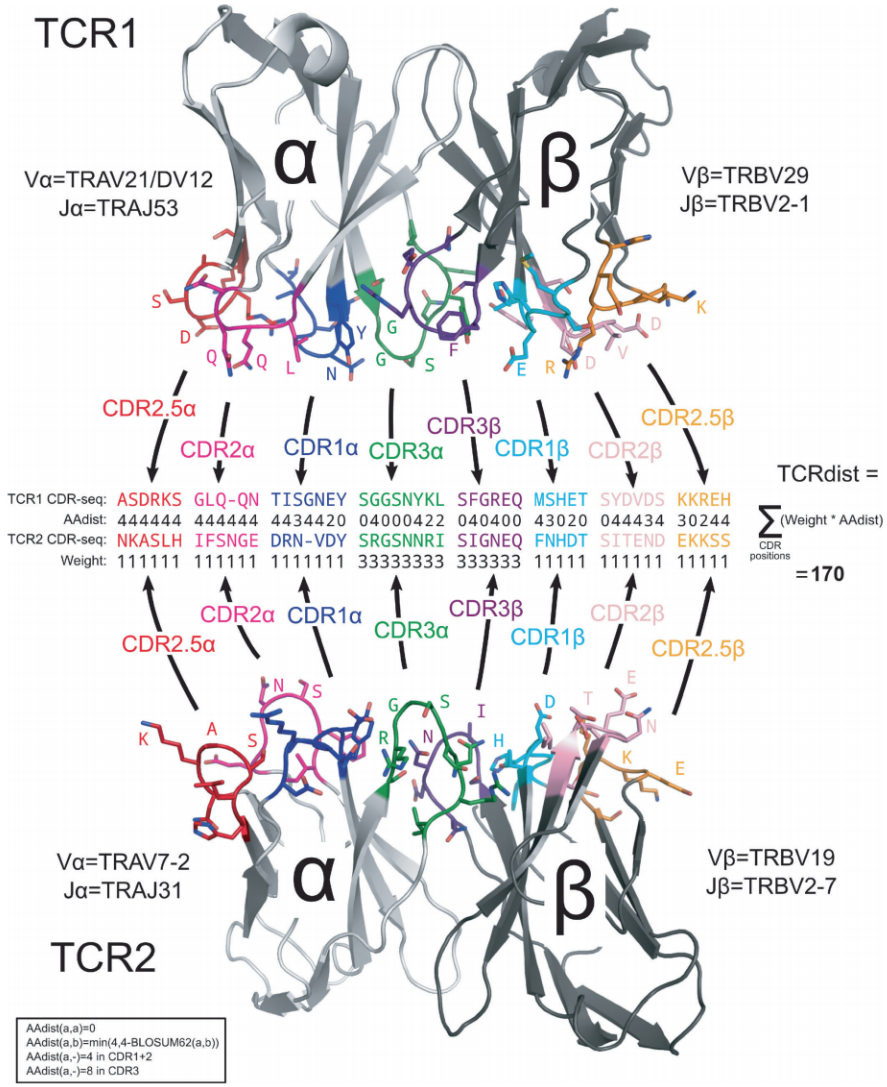
Define this space using a *distance*



Contrast to thinking of TCRs as a list of characteristics:

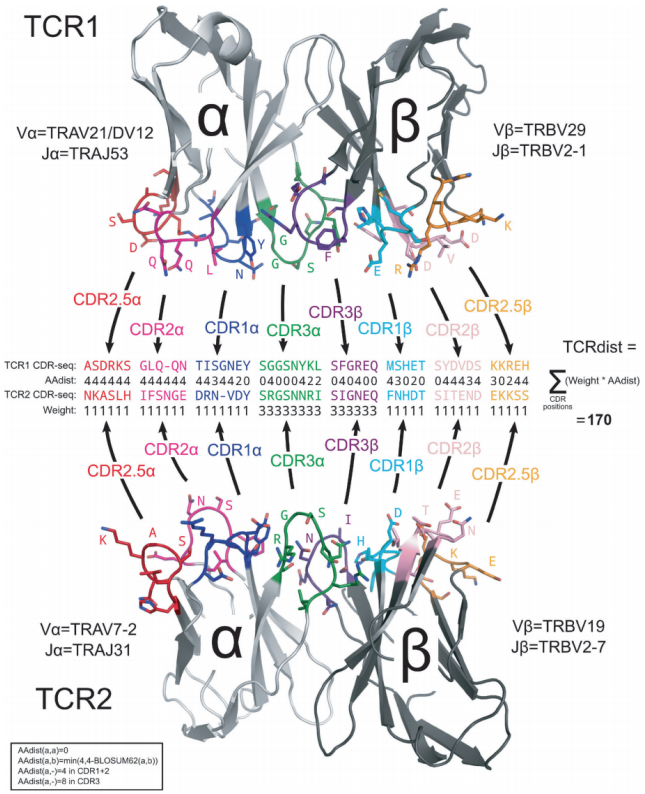
- V(D)J genes
- CDR3 length
- …

TCRDist: a structure-inspired distance between TCRs



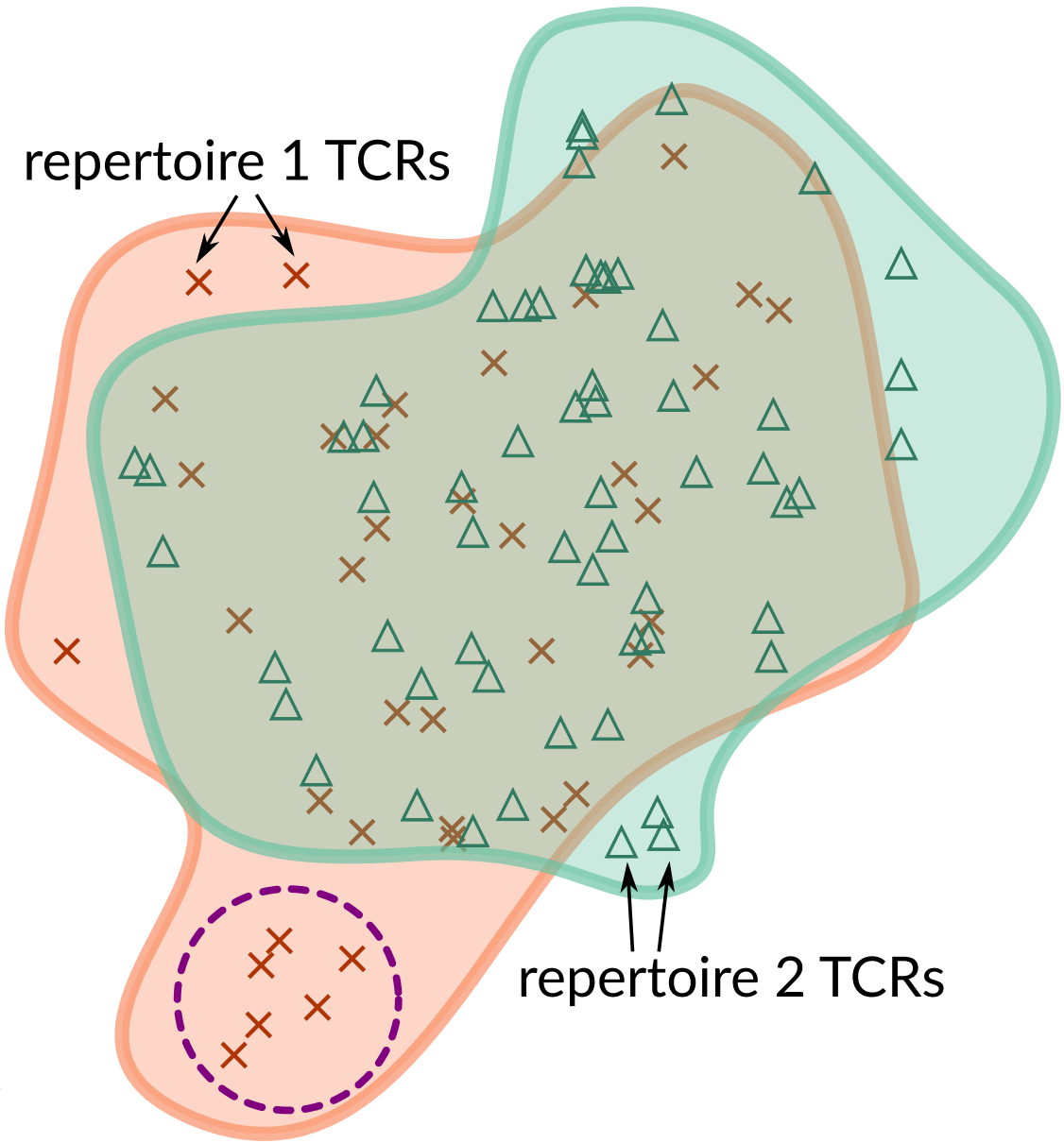
Dash,..., Bradley, Thomas (2017) Nature

TCRDist: a structure-inspired distance between TCRs

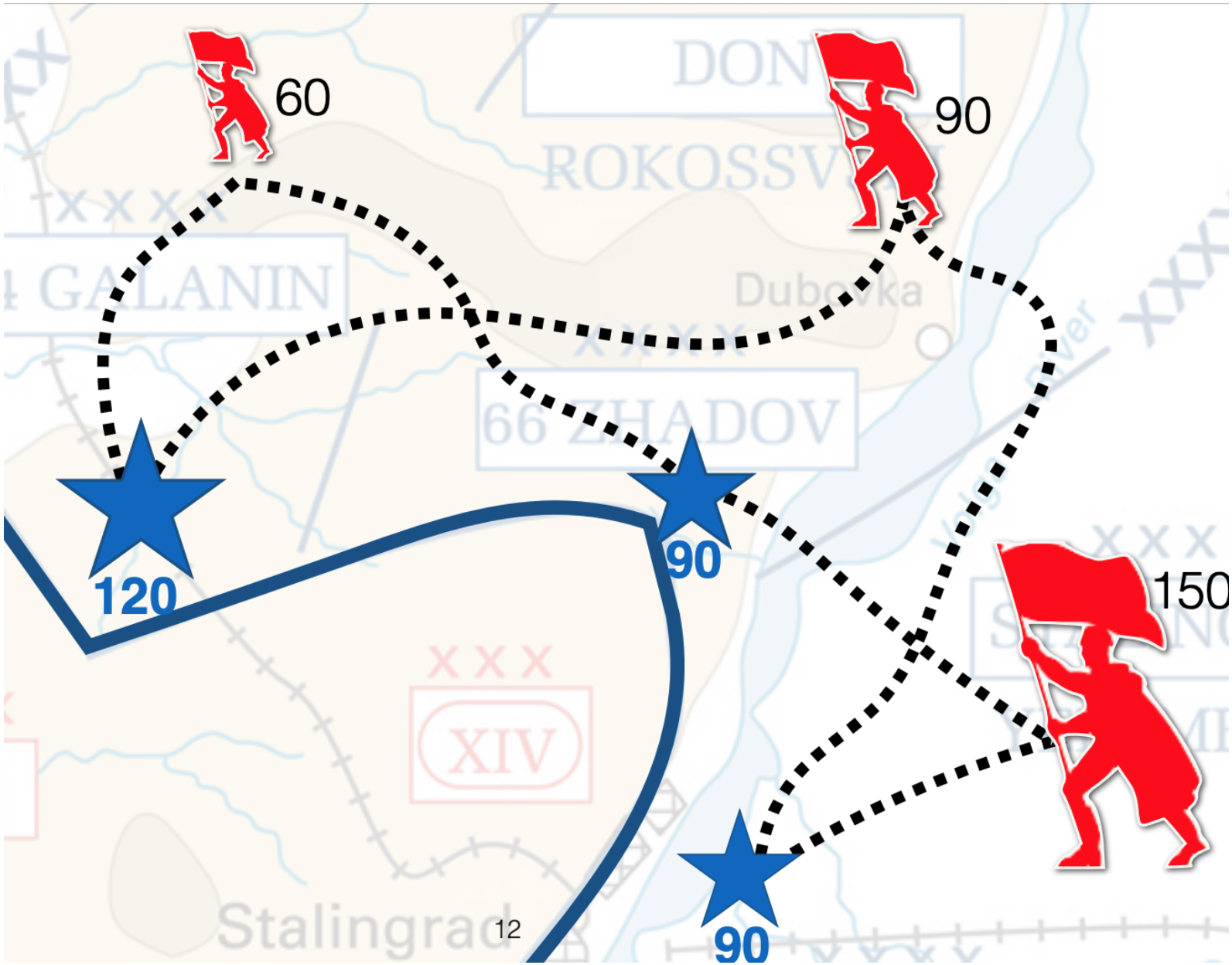


Dash,..., Bradley, Thomas (2017) Nature

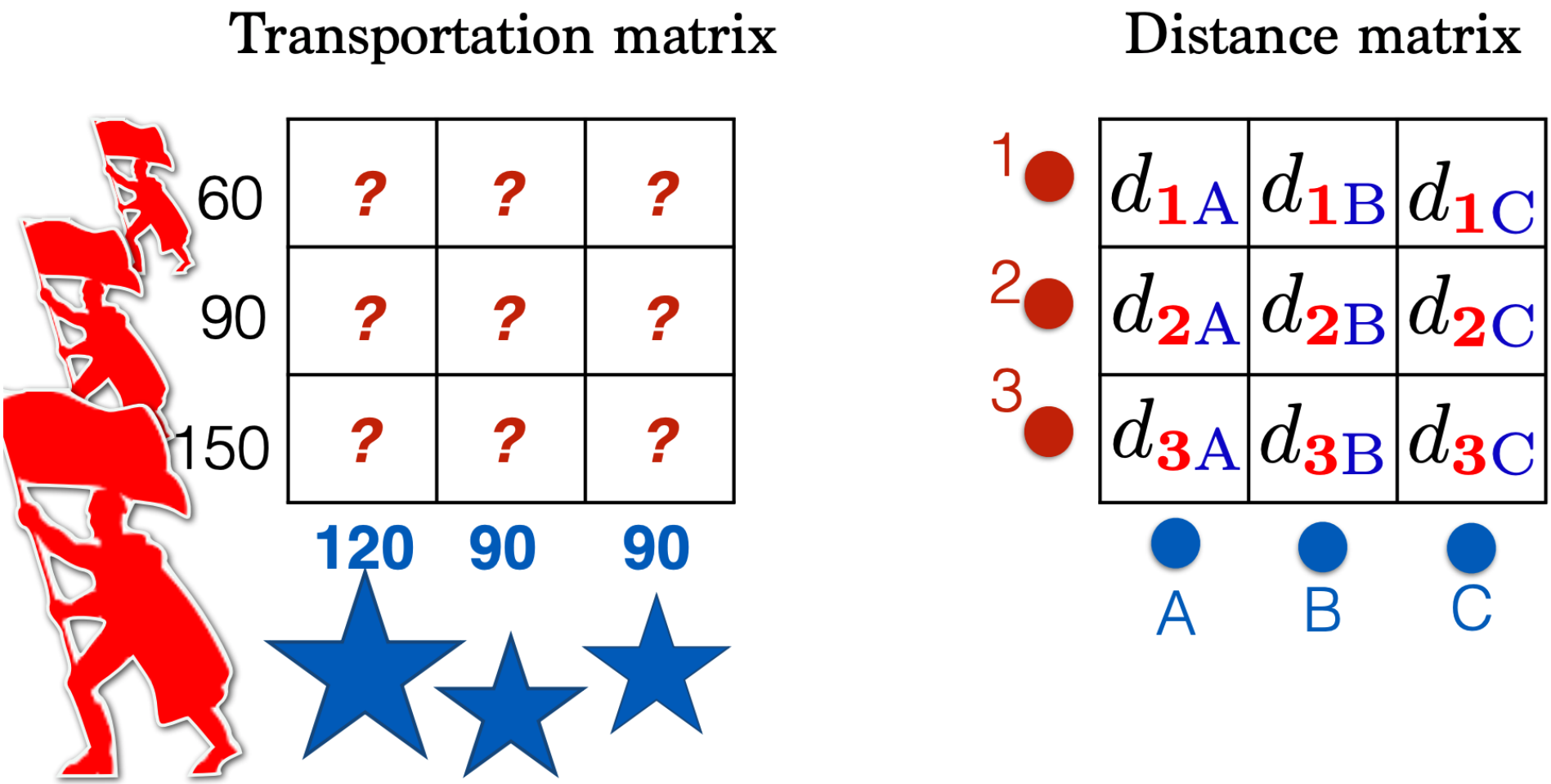
How can we learn a space defined by TCRdist?



Find groups of sequences ➔
that are *typical of their repertoire* but are *atypical of a second repertoire*

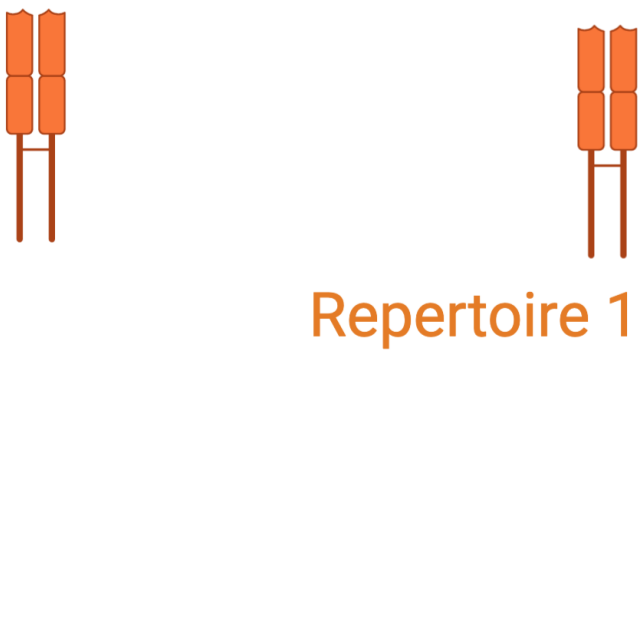


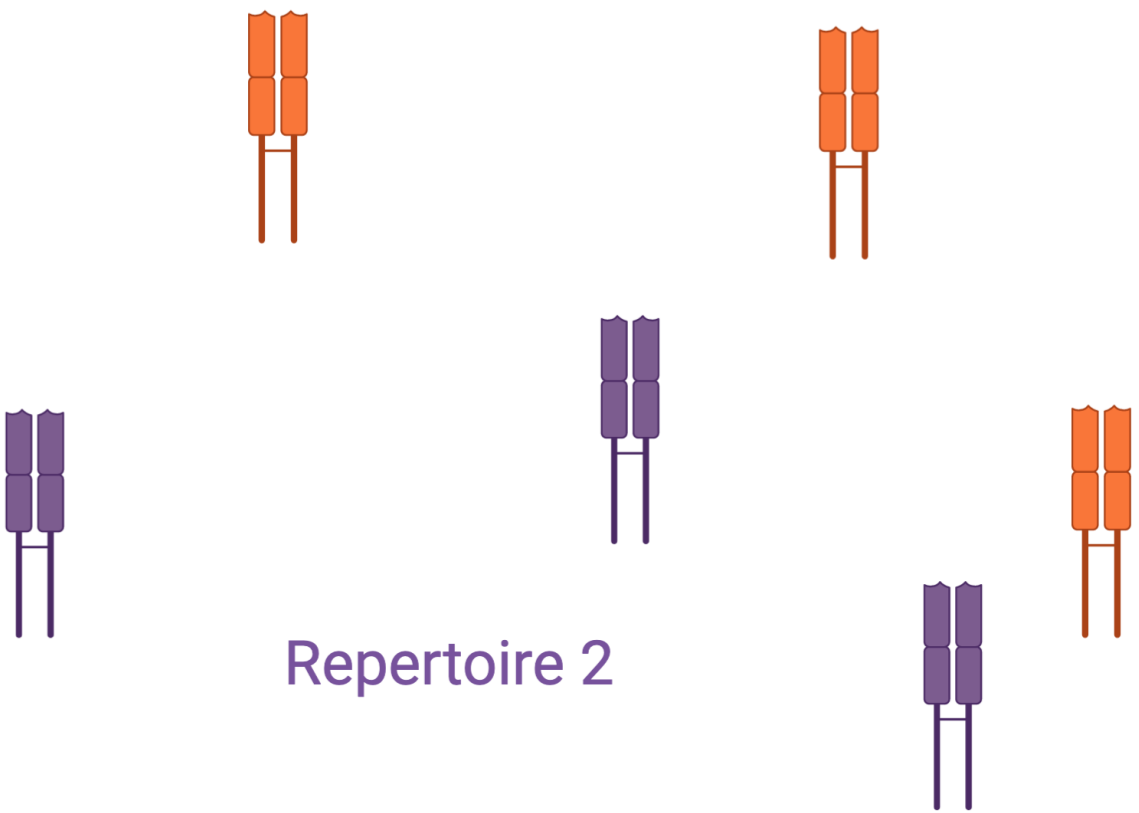
Marco Cuturi

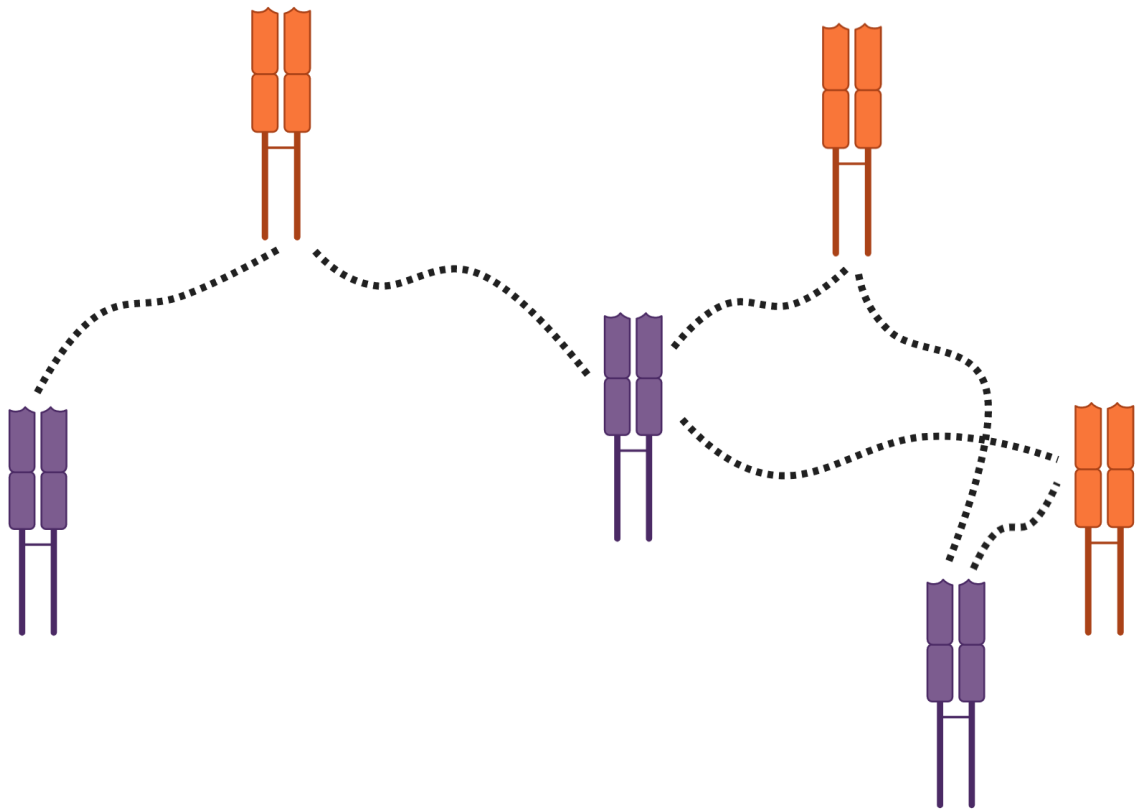


Marco Cuturi

Minimize *transportation cost* $\sum_{ij} m_{ij} d_{ij}$ over transportation matrices m_{ij} with the correct row and column sums.


Repertoire 1


Repertoire 2



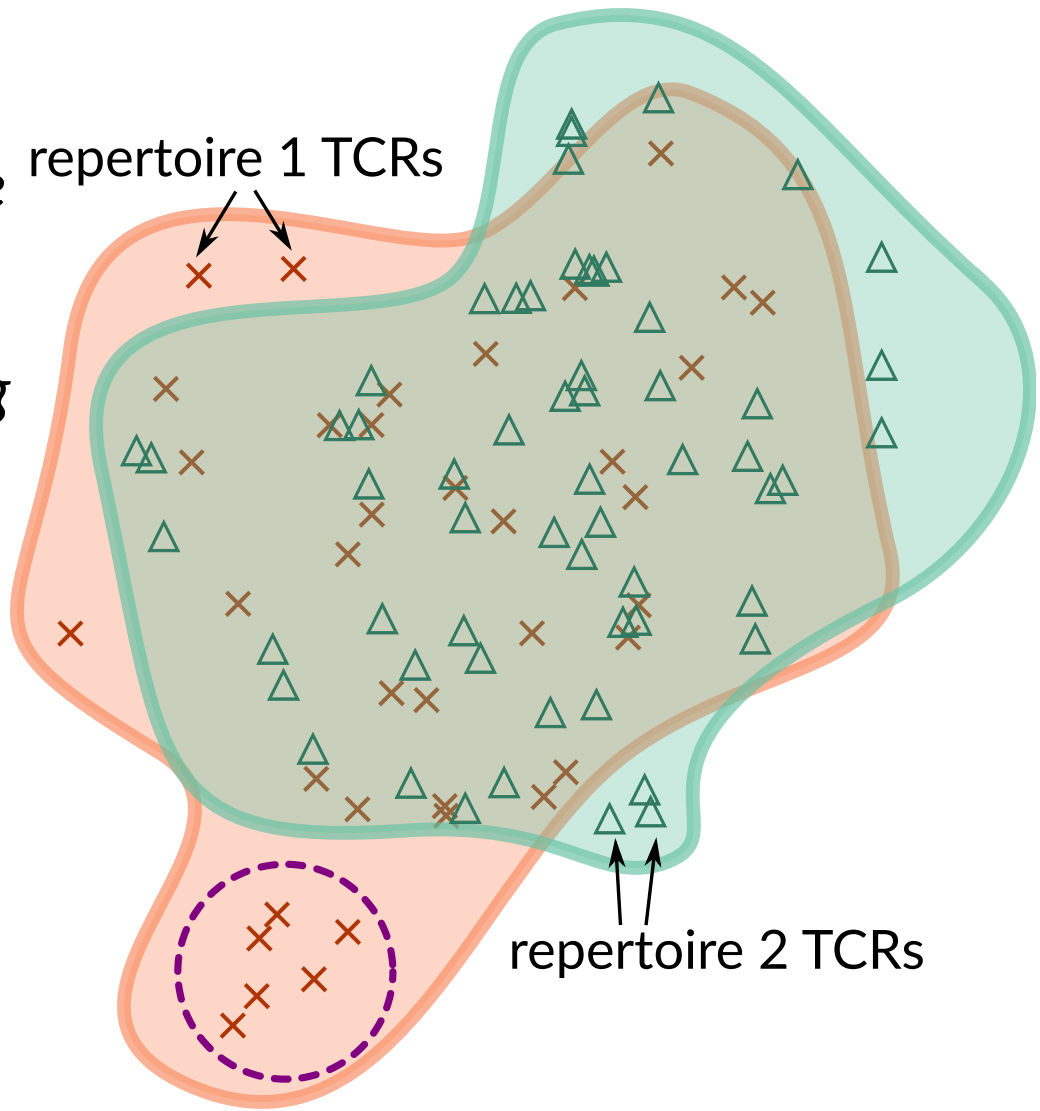
Minimize the amount of “transport” required in the space defined by TCRdist in order to “move” one repertoire into another.

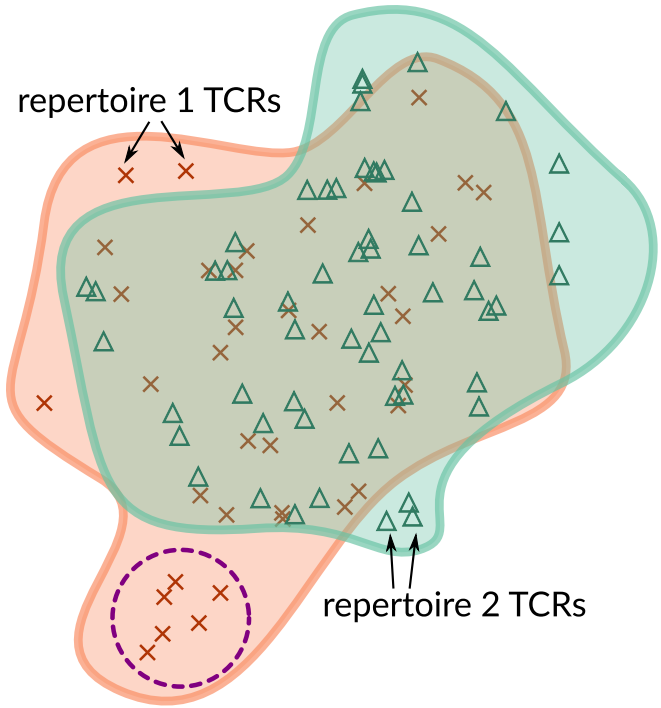
This gives us a correspondence between sequences of different repertoires even if none are identical.

Find groups of sequences that are *typical of their repertoire* but are *atypical of a second repertoire*

Find groups of sequences that are *close to one another according to TCRdist* but *have to be transported a long way to their corresponding sequence in a second repertoire*

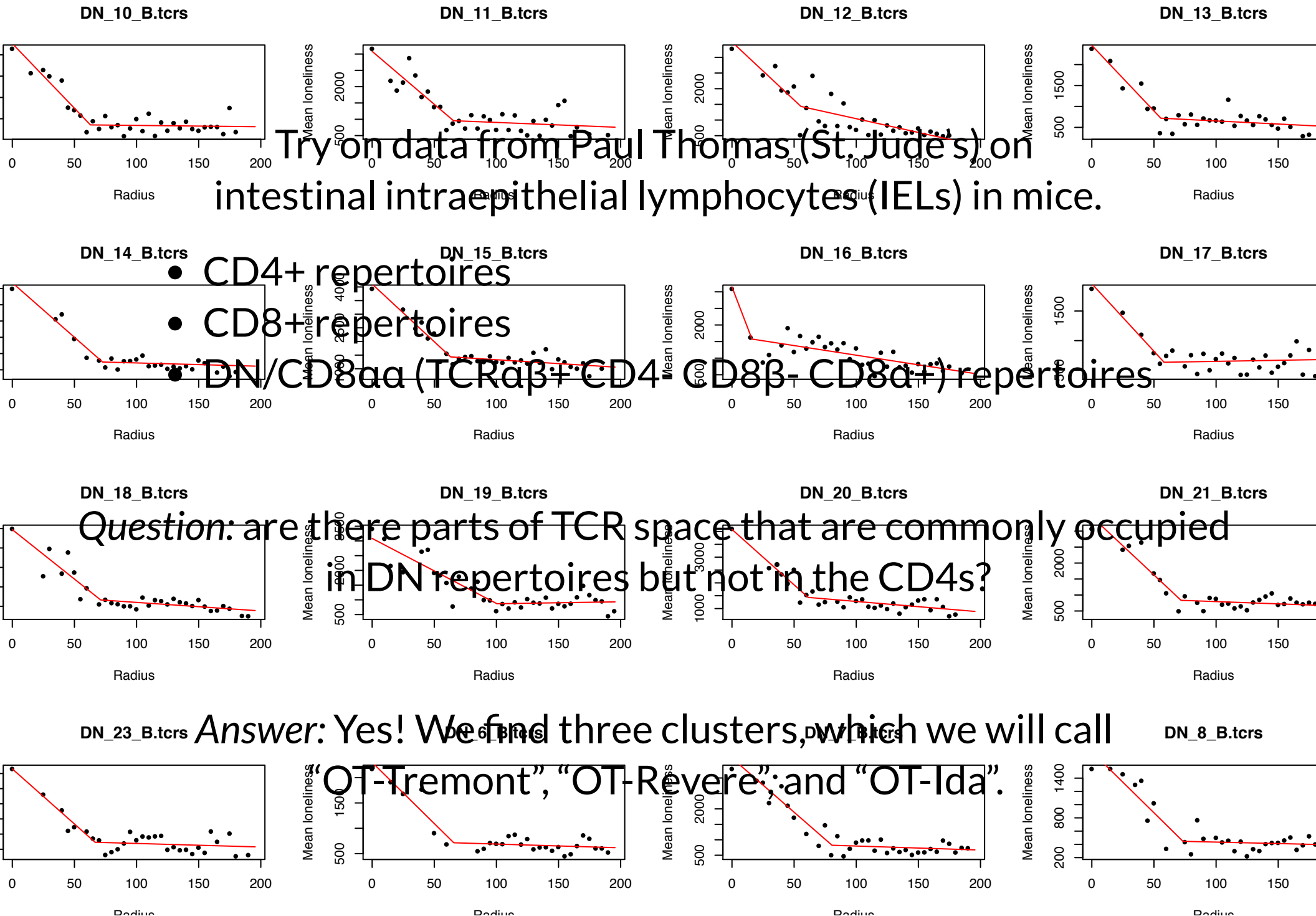
We call groups of sequences **lonely** that have to be transported a long way to their corresponding sequences in a second repertoire.





To find lonely clusters of sequences:

1. Calculate pairwise TCRdists
2. Use fast “Sinkhorn” entropy-regularized optimal transport
3. Find very lonely TCRs x.
4. Use segmented regression to find a cluster of sequences around x with high total loneliness.



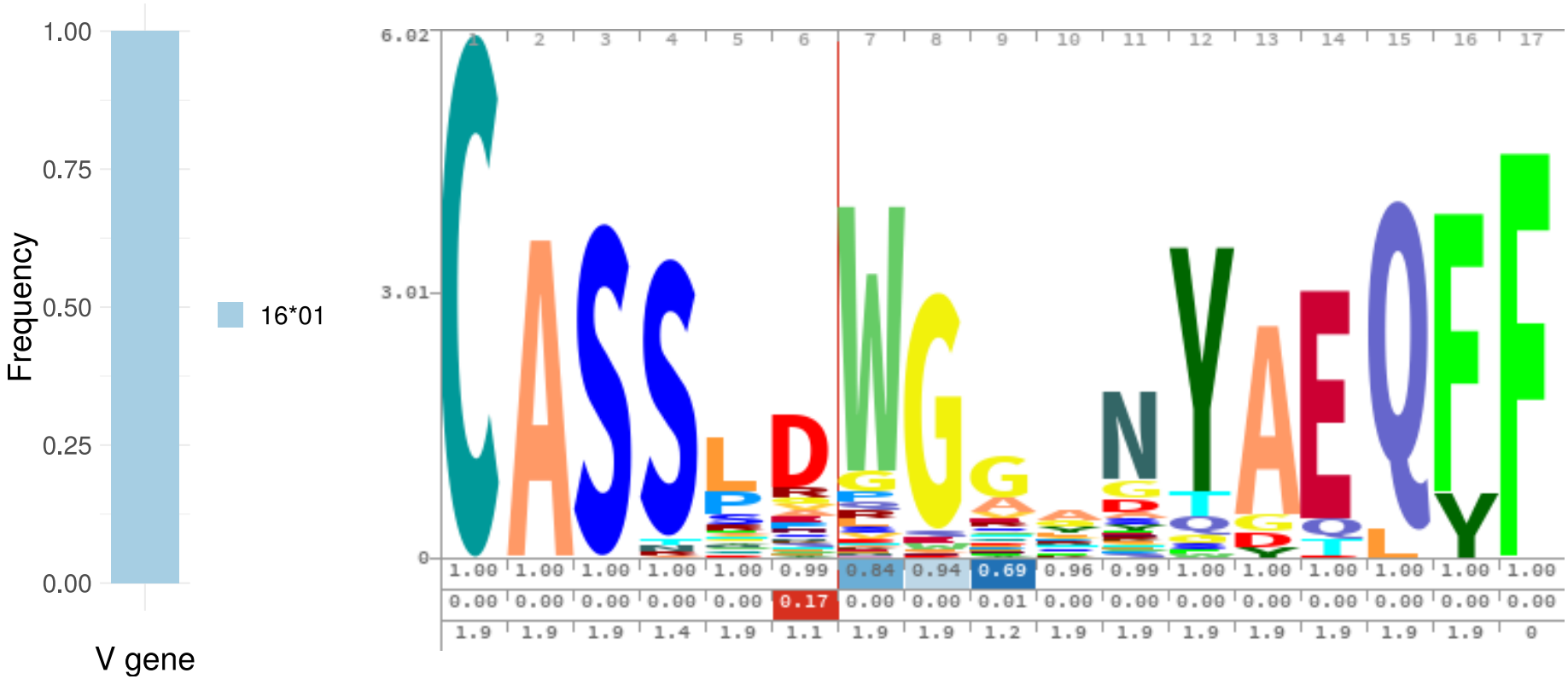
Try on data from Paul Thomas (St. Jude's) on intestinal intraepithelial lymphocytes (IELs) in mice.

- CD4+ repertoires
- CD8+repertoires
- DN/CD8αα (TCRαβ+ CD4- CD8β- CD8α+) repertoires

Question: are there parts of TCR space that are commonly occupied in DN repertoires but not in the CD4s?

Answer: Yes! We find three clusters, which we will call "OT-Tremont", "OT-Revere", and "OT-Ida".

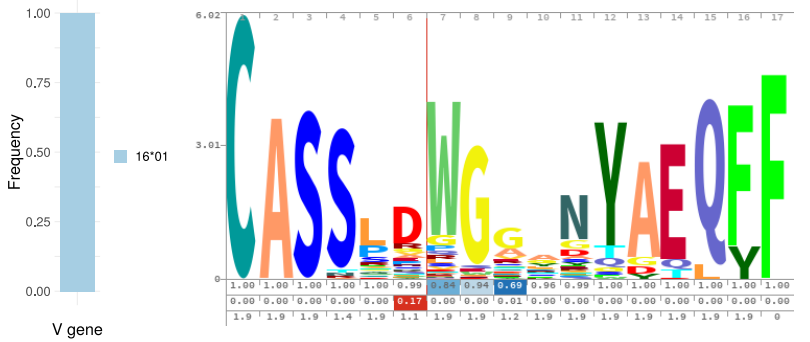
We describe these clusters using profile HMMs



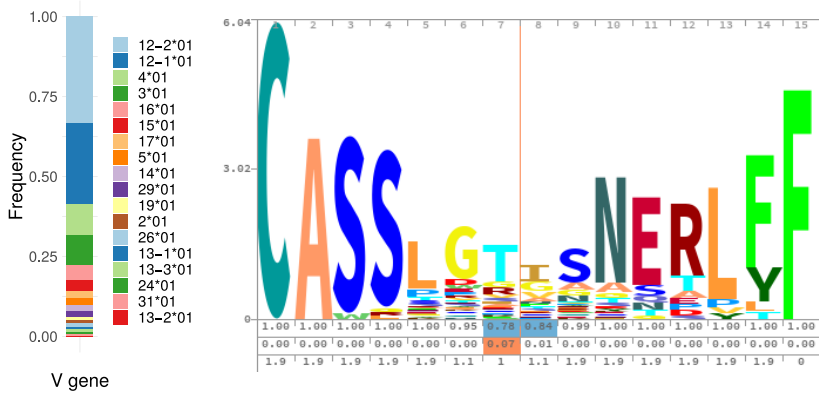
http://hmmer.org/ https://skylign.org/

Three “regions” of DN repertoires not in CD4+ repertoires

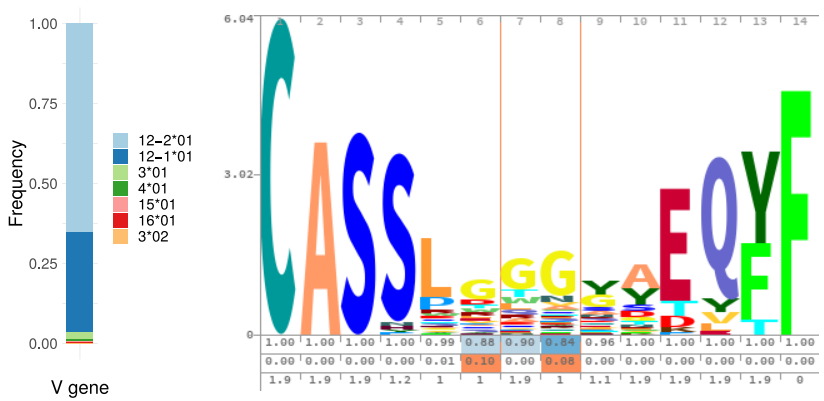
OT-Tremont



OT-Revere



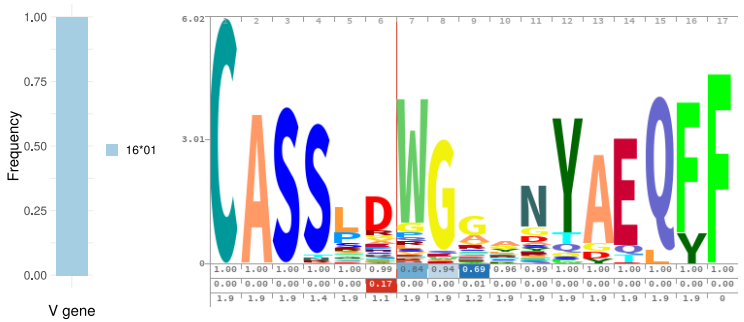
OT-Ida

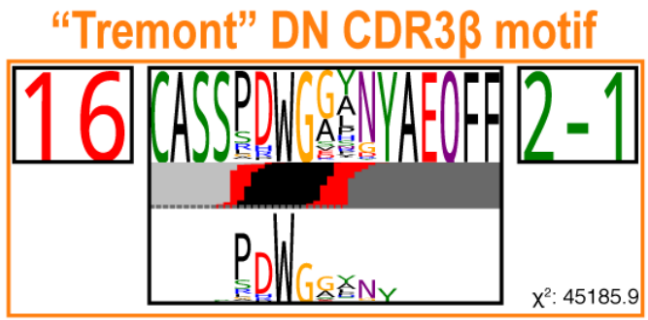


Our automated results

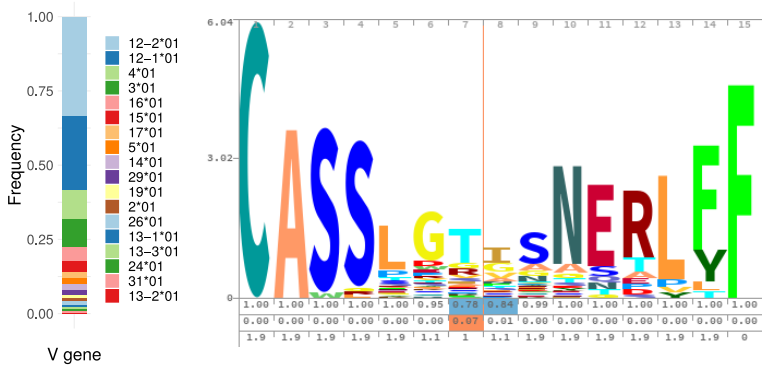
Original results

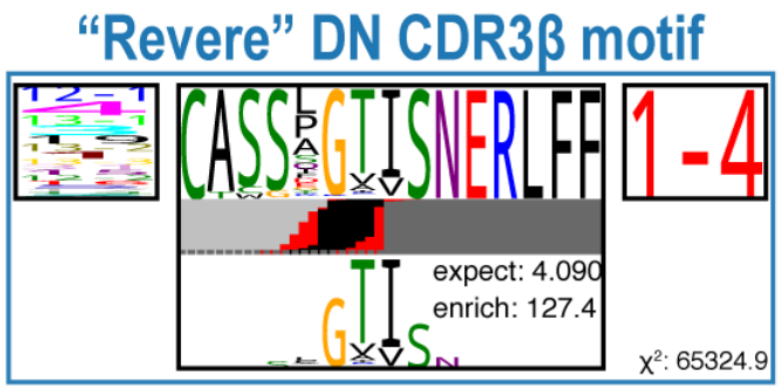
OT-Tremont



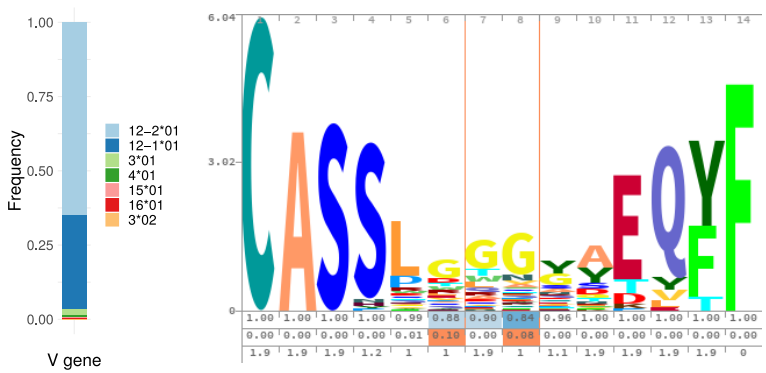


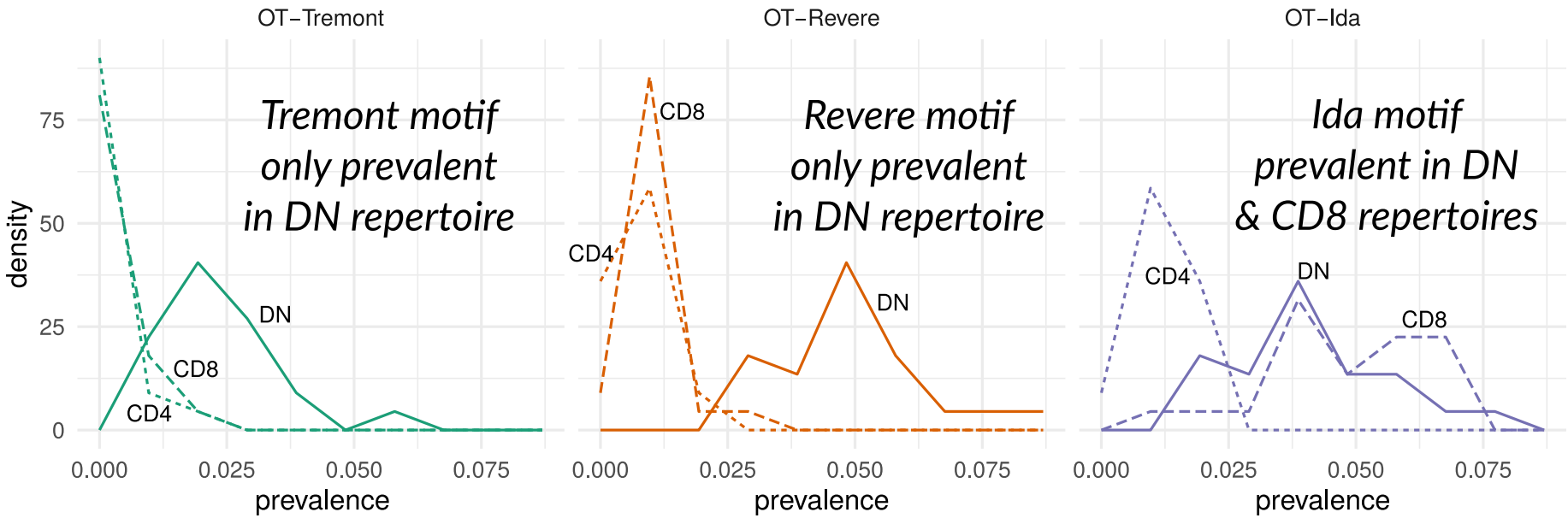
OT-Revere





OT-Ida

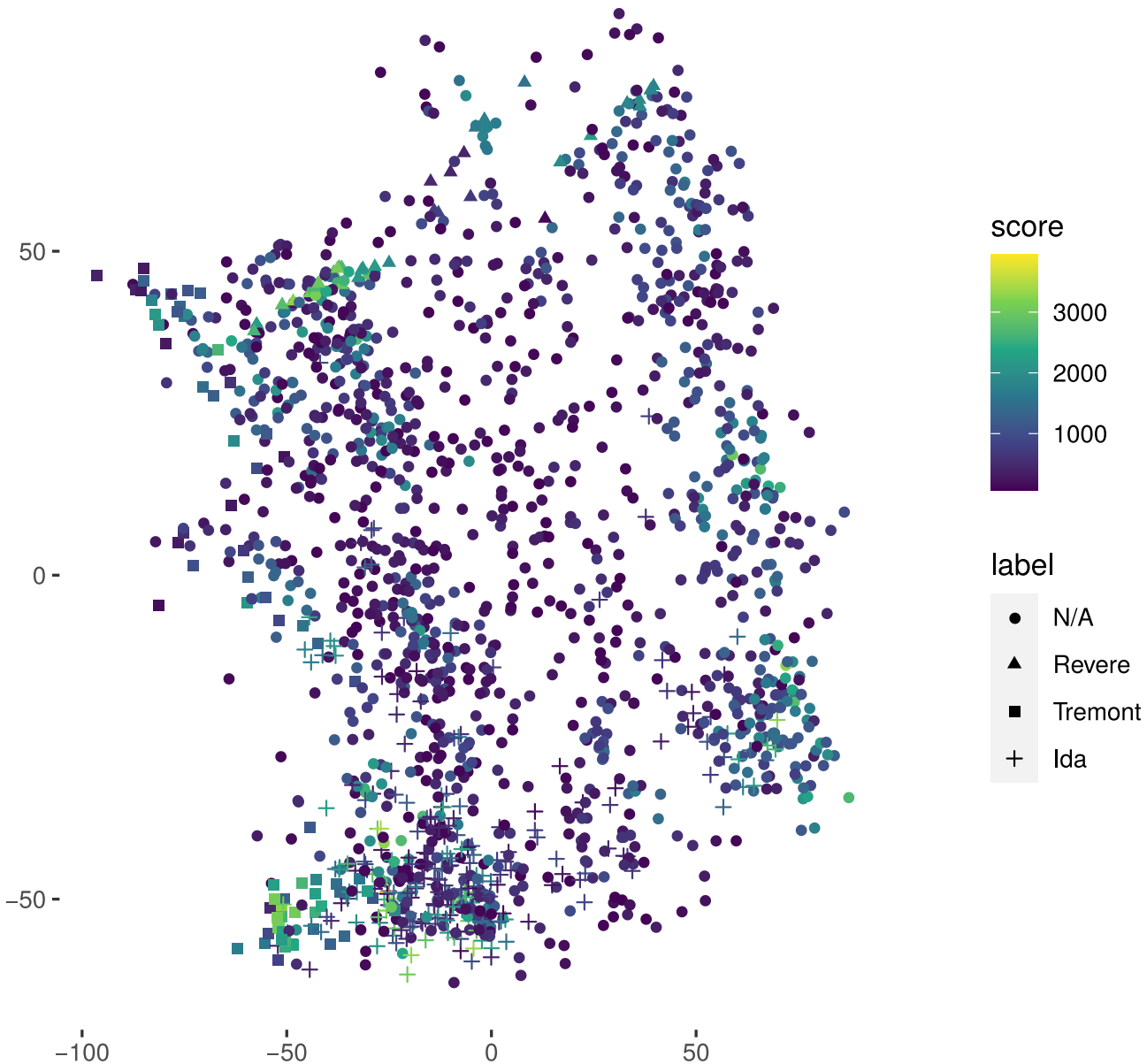




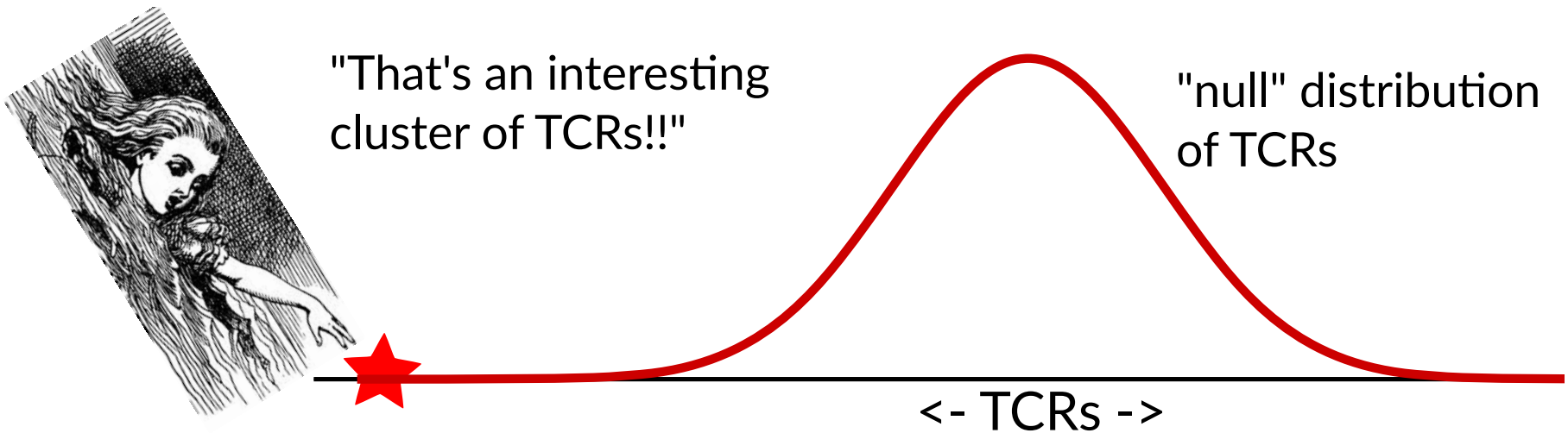
These motifs are now tractable markers in the repertoire:

- What do these TCRs recognize? (clone TCRs for epitope discovery)
- Do they correlate with a particular T cell phenotype or determine their lineage? (perform single-cell gene expression and TCR)
- What is their role in disease and homeostasis? (transgenic T cells)

We wouldn't get these results with dim reduction



ALICE (Pogorelyy, Minervina, et al 2019 PLOS Biology)
is a *parametric* (model-based) approach:



Optimal transport is *nonparametric*, i.e. model-free.

ALICE ≈ t-test
optimal transport ≈ Mann-Whitney U

Conclusion of optimal transport section

The optimal transport framework enables detailed comparison of repertoires in a space defined by a distance.

We have applied it to find groups of sequences that are *typical of their repertoire* but are *atypical of a second repertoire*.

Our pipeline automatically identified interesting groups of sequences, extending previous results, without careful data cleaning or sequence-gazing.

Lots of other ways to use optimal transport ideas…

… a few recently published projects and a few in the mix …

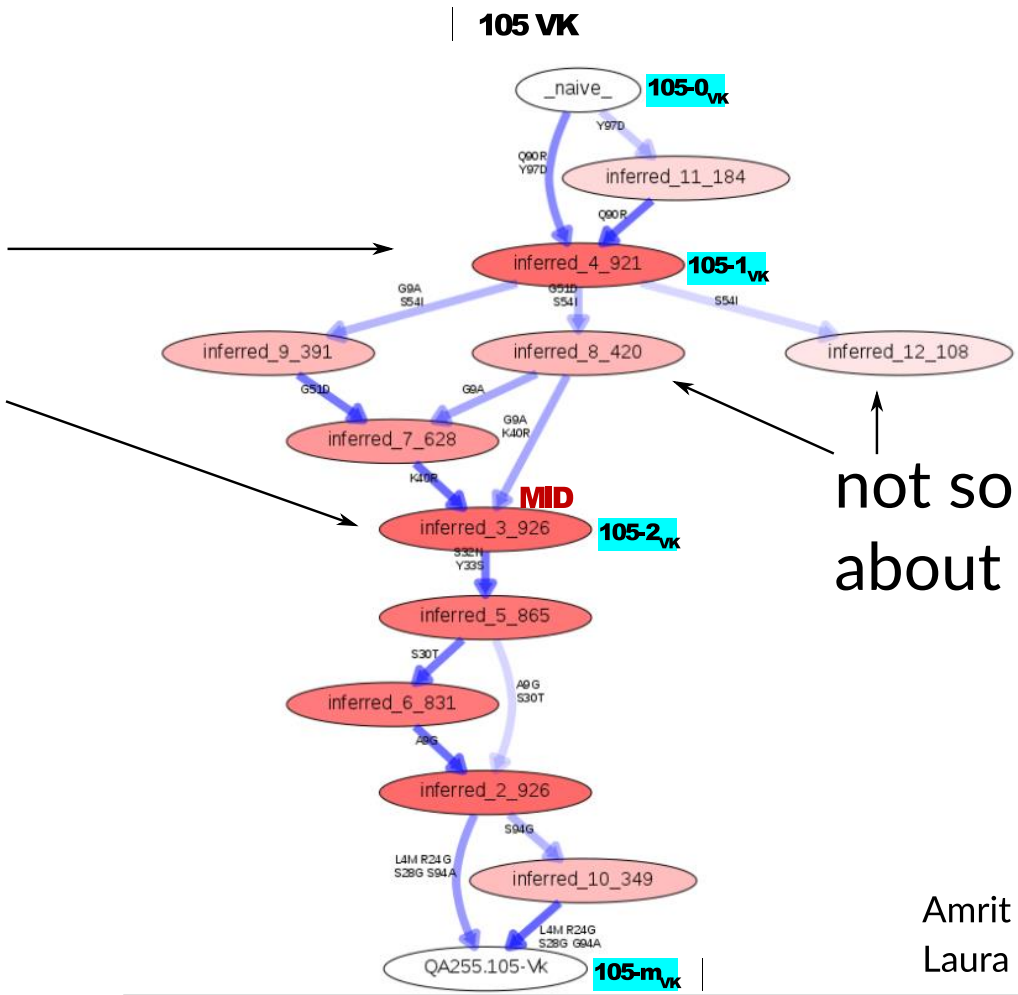
Q: If you reconstruct ancestral BCR sequences, how strong is the evidence for those ancestral sequences?

Bad news: there is typically a lot of tree uncertainty in phylogenetic inference for BCRs

Good news: using Bayesian phylogenetic techniques, we can integrate out the tree uncertainty and evaluate uncertainty on the level of ancestral sequences directly.

Find certain ancestral sequences despite tree uncertainty:

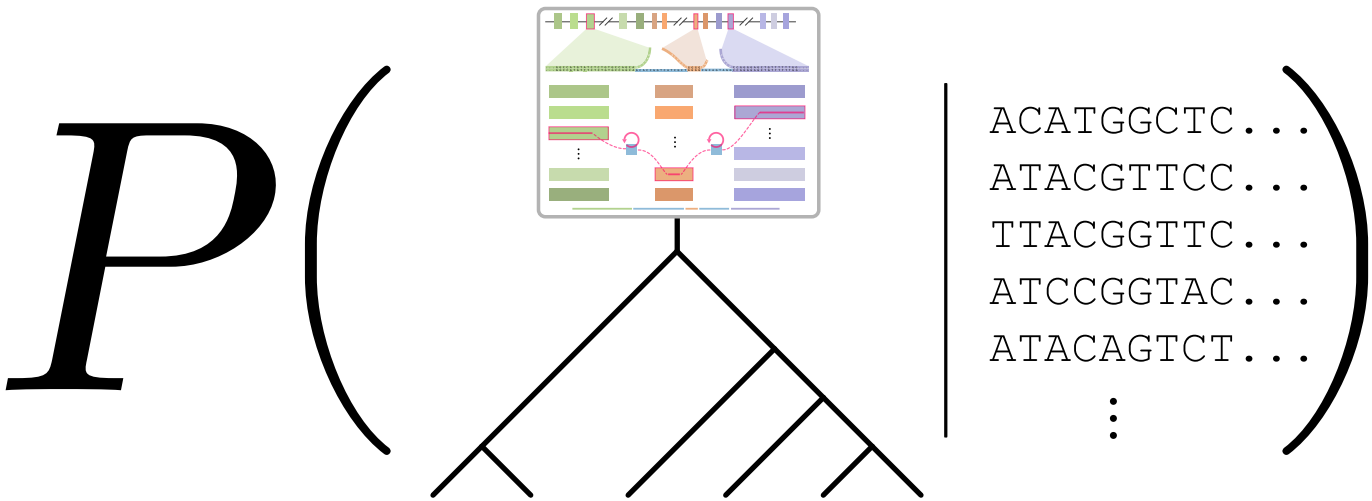
we are fairly certain about these ancestral sequences



not so sure about these

Amrit Dhar &
Laura Doepker

We can also incorporate uncertainty in VDJ recombination



Dhar, Ralph, Minin, & M (2020). *PLOS Computational Biology*

Note: for some important lineages, we still find a lot of ancestral sequence uncertainty!

PS: BEAST is awesome, but I don't think it's appropriate for BCR lineages.

Q: When you identify a BCR clonal lineage with sequences of interest, how do you pick sequences that may have high affinity?

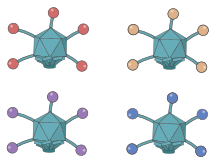
Idea: let's train a machine-learning model to use evolutionary and sequence features to predict good binders.

Result: No model needed. Just pick sequences close to the consensus of the clonal lineage, not the most mutated sequences!

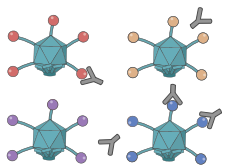
Ralph & M, (2020). *PLOS Computational Biology*

PhIP-seq to understand SARS-CoV-2 serological response

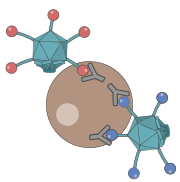
make phage library displaying epitopes of interest



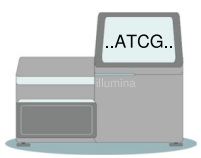
incubate with antibody



pull down phage bound by antibodies

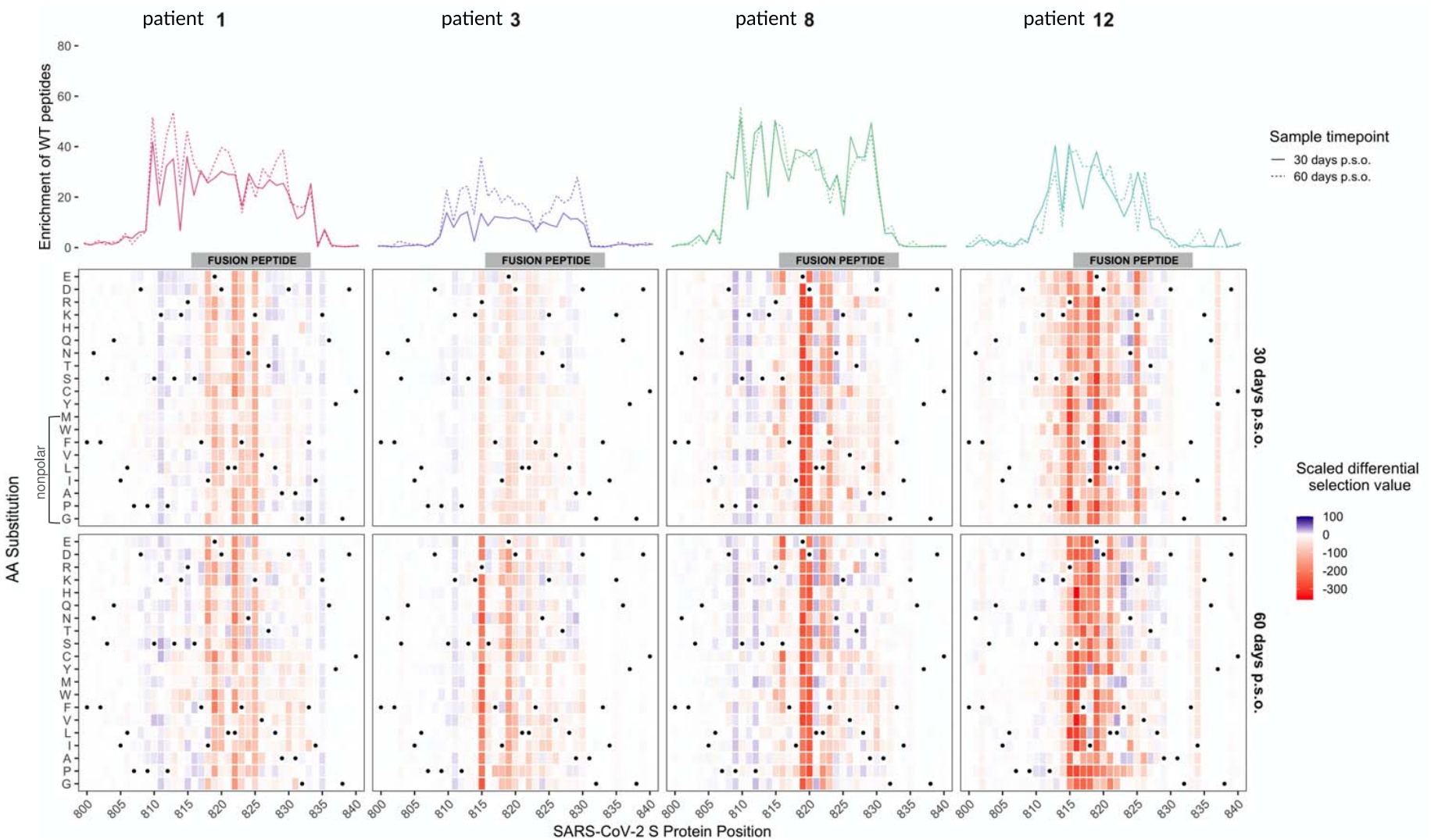


sequence those phage (and some controls)



modified from Mohan...Larman (2018) Nat Protoc

*... how well do antibodies recognize **mutant** peptides?*



Garrett, …, M & Overbaugh, J. (2020). *bioRxiv*. "High resolution…"

Result: *The escape pathway for the virus differs between patients.*

In the pipeline

- Probabilistic mechanistic models of somatic hypermutation
- Modeling epistasis in antigens and antibodies

- GWAS to understand how genetics shapes TCR repertoire (with Phil Bradley and many others)
- HIV antibody deep mutational scan (with Bloom lab)
- Finding broad and potent anti-dengue antibodies (with Goo lab)

… and new foundations for Bayesian phylogenetics (× lots!)…

Thank you

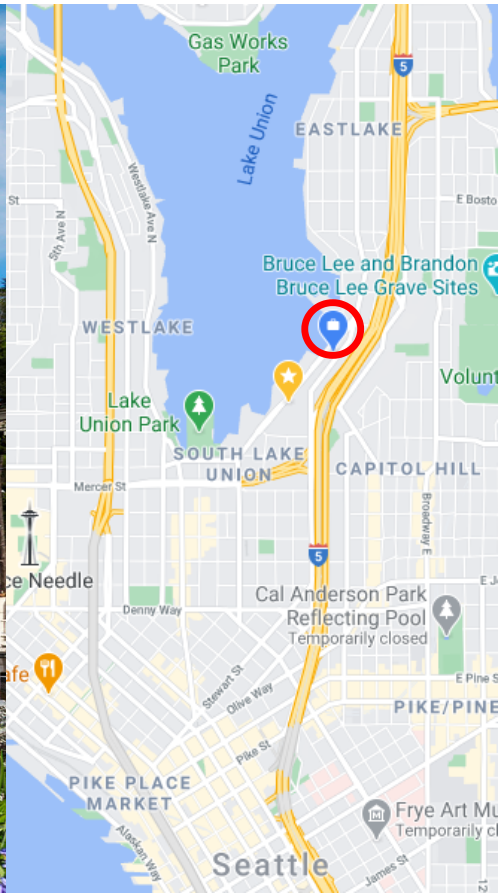
- Branden (Olson) Steele 🎓
- Phil Bradley (Fred Hutch)
- Paul Thomas and Stefan Schattgen (St. Jude's)

- Amrit Dhar, Ducan Ralph, and Vladimir Minin.
- Caitlin Stoddard, Meghan Garrett, Jared Galloway, **Julie Overbaugh** & her lab

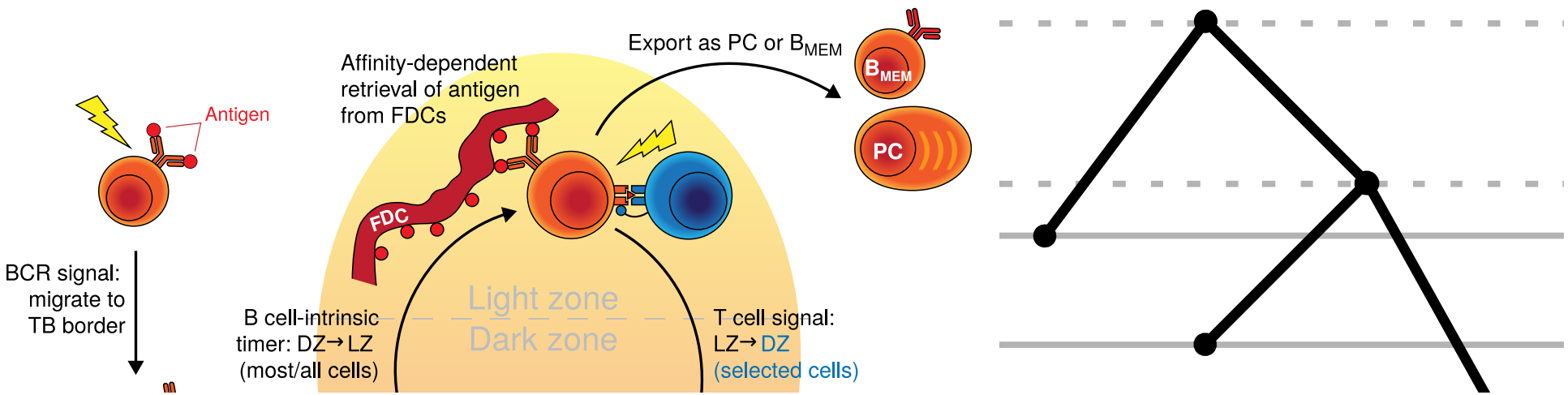
NSF, NIH and HHMI-Simons Faculty Scholars Program

If this work interests you, drop me a line! 👋

The "Steam Plant": data science + immunotherapy



If you are a programmer, postdoc, or aspiring faculty member and are interested in working in an awesome collaborative environment, email me at ematsen@gmail.com!


Antigen
BCR signal:
migrate to
TB border
Affinity-dependent
retrieval of antigen
from FDCs
FDC
Export as PC or B_{MEM}
B_{MEM}
PC
Light zone
Dark zone
B cell-intrinsic
timer: DZ→LZ
(most/all cells)
T cell signal:
LZ→DZ
(selected cells)