Linking T cell receptor sequence to transcriptional profiles with clonotype neighbor graph analysis (CoNGA)

joint work w/ Kate Guion (USC),
Paul Thomas, Stefan Schattgen, and Jeremy Crawford (St. Jude),
Mike Stubbington and Alvaro M Barrio (10x Genomics)
(manuscript in revision)

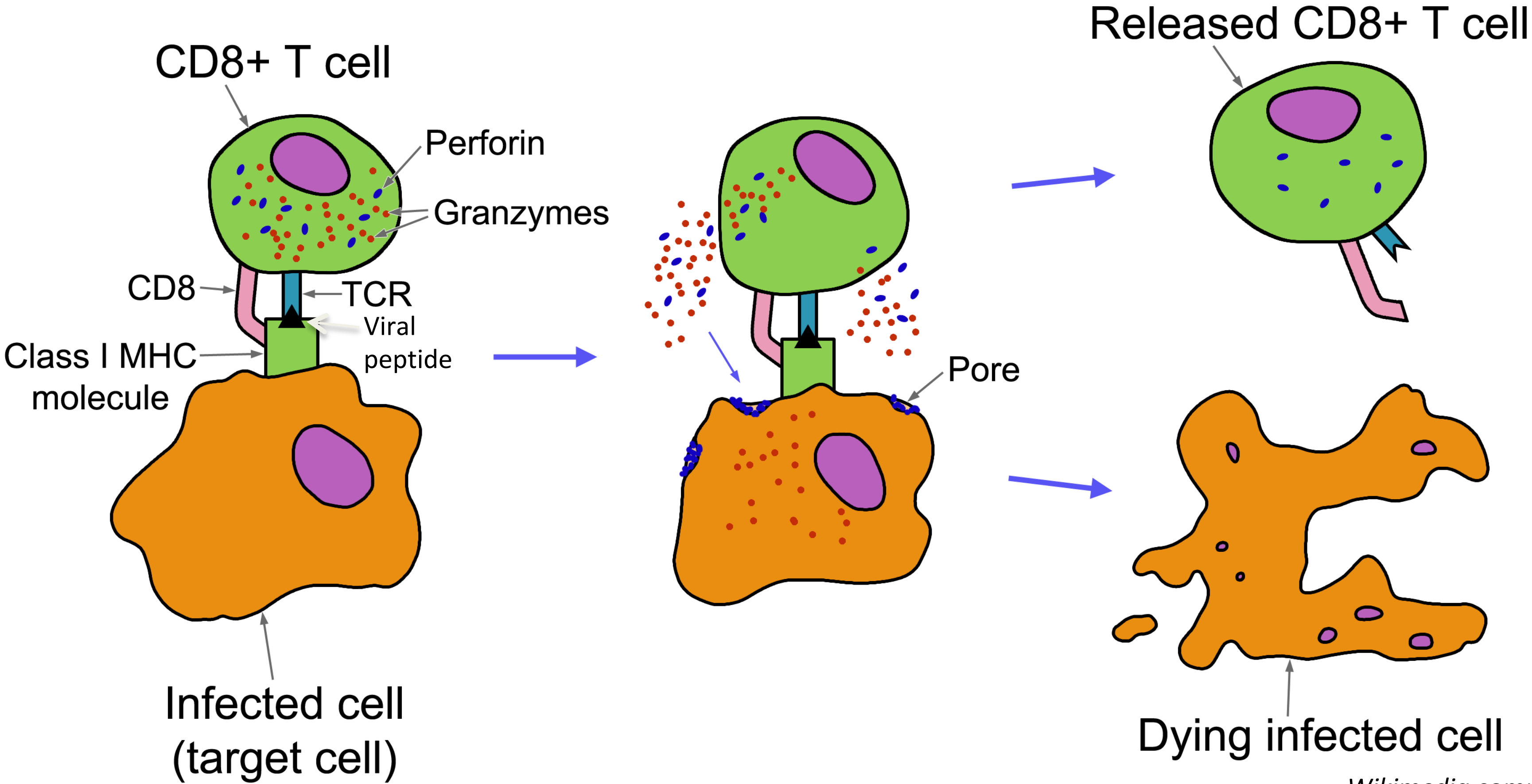
Phil Bradley
Fred Hutch Cancer Center

<u>Abbreviations</u>
TCR = T cell receptor
GEX = gene expression
pMHC = peptide-MHC

Outline

- Background
  - T cells and T cell receptors (TCRs)
  - single-cell gene expression (GEX) analysis
- CoNGA graph-vs-graph analysis
- CoNGA graph-vs-feature analysis

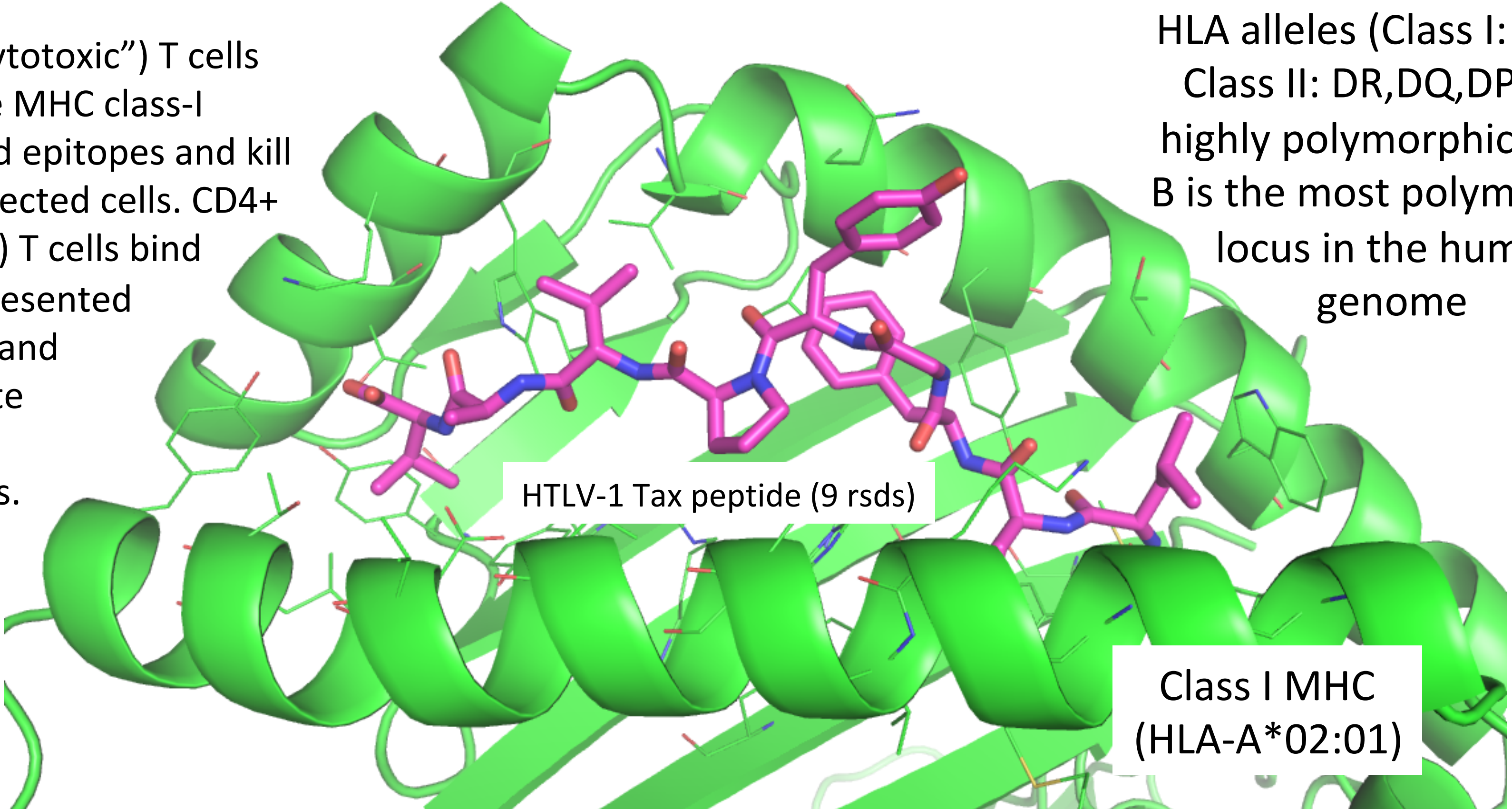
T cells are key regulators and effectors of the adaptive immune response



Wikimedia commons

αβ T cells recognize peptide epitopes presented by MHC (aka HLA) proteins

CD8+ (“cytotoxic”) T cells recognize MHC class-I presented epitopes and kill virally-infected cells. CD4+ (“helper”) T cells bind class-II presented epitopes and coordinate immune responses.



HTLV-1 Tax peptide (9 rsds)

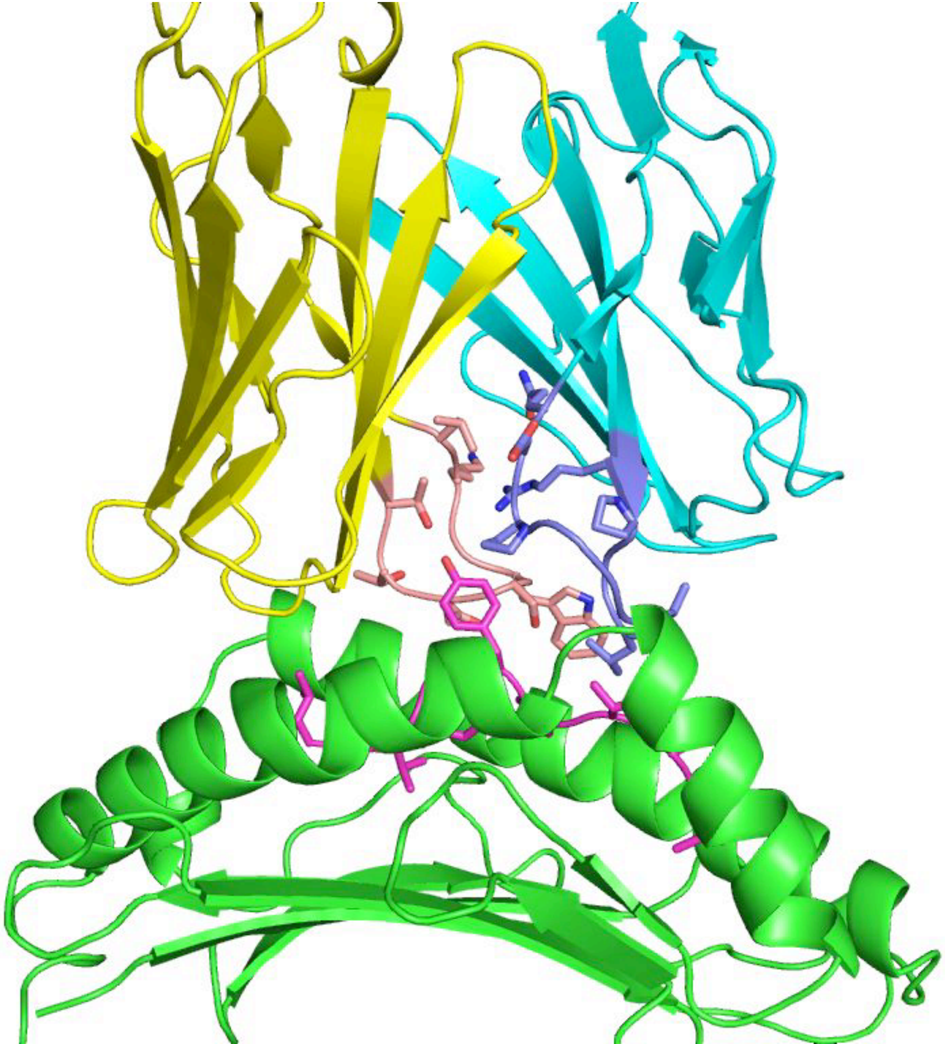
Class I MHC (HLA-A*02:01)

HLA alleles (Class I: A,B,C; Class II: DR,DQ,DP) are highly polymorphic. HLA-B is the most polymorphic locus in the human genome

The peptide-MHC specificity of a T cell is determined by the sequence of its heterodimeric T cell receptor (**TCR**).

TCRs are built by a stochastic genome rearrangement process that results in astronomical sequence diversity.

Each T cell thus has a ‘unique’ rearranged receptor (clonally-related T cells will share the same TCR)



Each TCR chain has three loops that can contact peptide-MHC. The highly variable CDR3 loops are shown in stick representation and colored pink and purple.

Yellow: TCR α chain
Cyan: TCR β chain
Magenta: peptide
Green: MHC

‘A6’ TCR bound to HTLV-1 Tax peptide presented by HLA-A*02:01

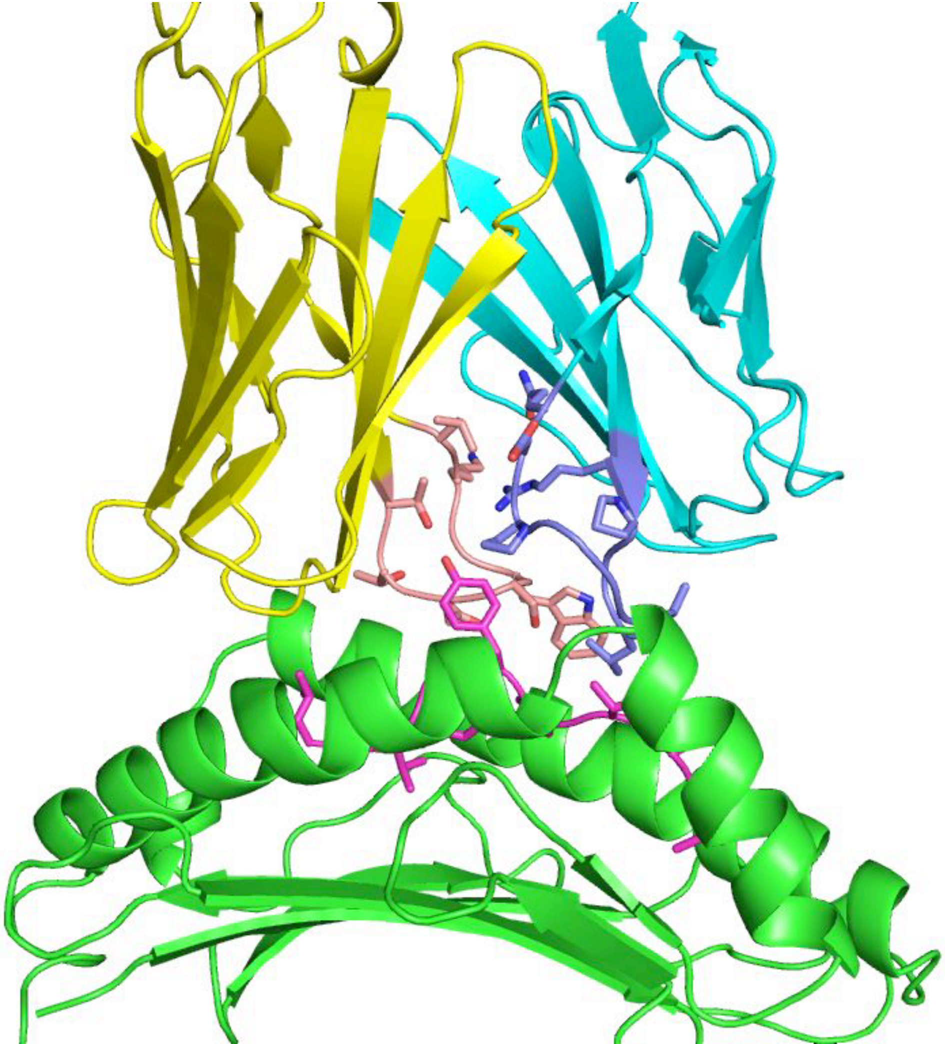
Green: MHC (HLA-A*02:01)
Magenta: epitope (LLFGYPVAV)
Yellow: TCR alpha chain
Cyan: TCR beta chain

CDR3 loops shown as sticks

Vα	TRAV12-2*01
Jα	TRAJ24*02
CDR3α	CAVTTDSWGKLQF
Vβ	TRBV6-5*01
Jβ	TRBJ2-7*01
CDR3β	CASRPGLAGGRPEQYF

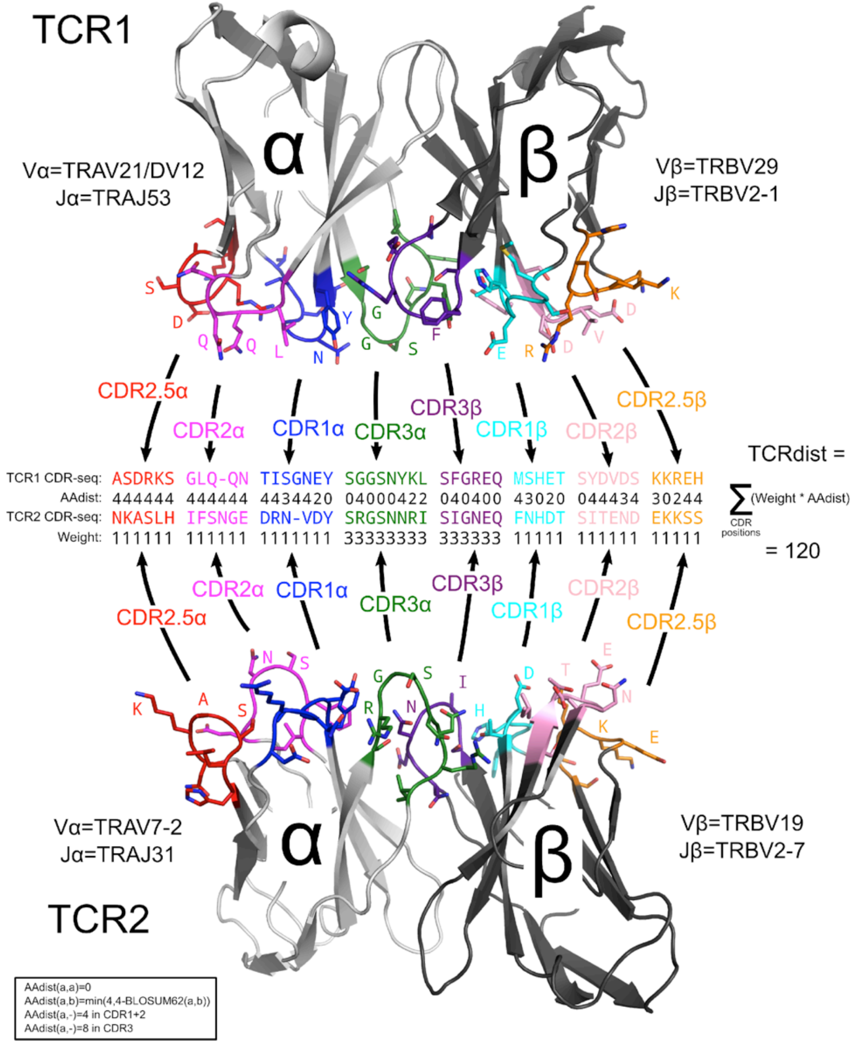
↑

These data (four gene identifiers and two CDR3 sequences) completely describe the TCR protein (no hypermutation)



TCRdist distance measure

To quantify the distance between two TCRs we use a sequence-based distance measure that aligns the CDR loops (the regions of the receptors typically involved in pMHC binding) and tallies an AA-similarity-weighted Hamming distance.



TCR1 CDR-seq:	ASDRKS	GLQ-QN	TISGNEY	SGGSNYKL	SFGREQ	MSHET	SYDVDS	KKREH
AAdist:	444444	444444	4434420	04000422	040400	43020	044434	30244
TCR2 CDR-seq:	NKASLH	IFSNGE	DRN-VDY	SRGSNNRI	SIGNEQ	FNHDT	SITEND	EKKSS
Weight:	111111	111111	1111111	33333333	333333	11111	111111	11111

$$\text{TCRdist} = \sum_{\substack{\text{CDR}\\ \text{positions}}} (\text{Weight} * \text{AAdist}) = 120$$

AAdist(a,a)=0
AAdist(a,b)=min(4,4-BLOSUM62(a,b))
AAdist(a,-)=4 in CDR1+2
AAdist(a,-)=8 in CDR3

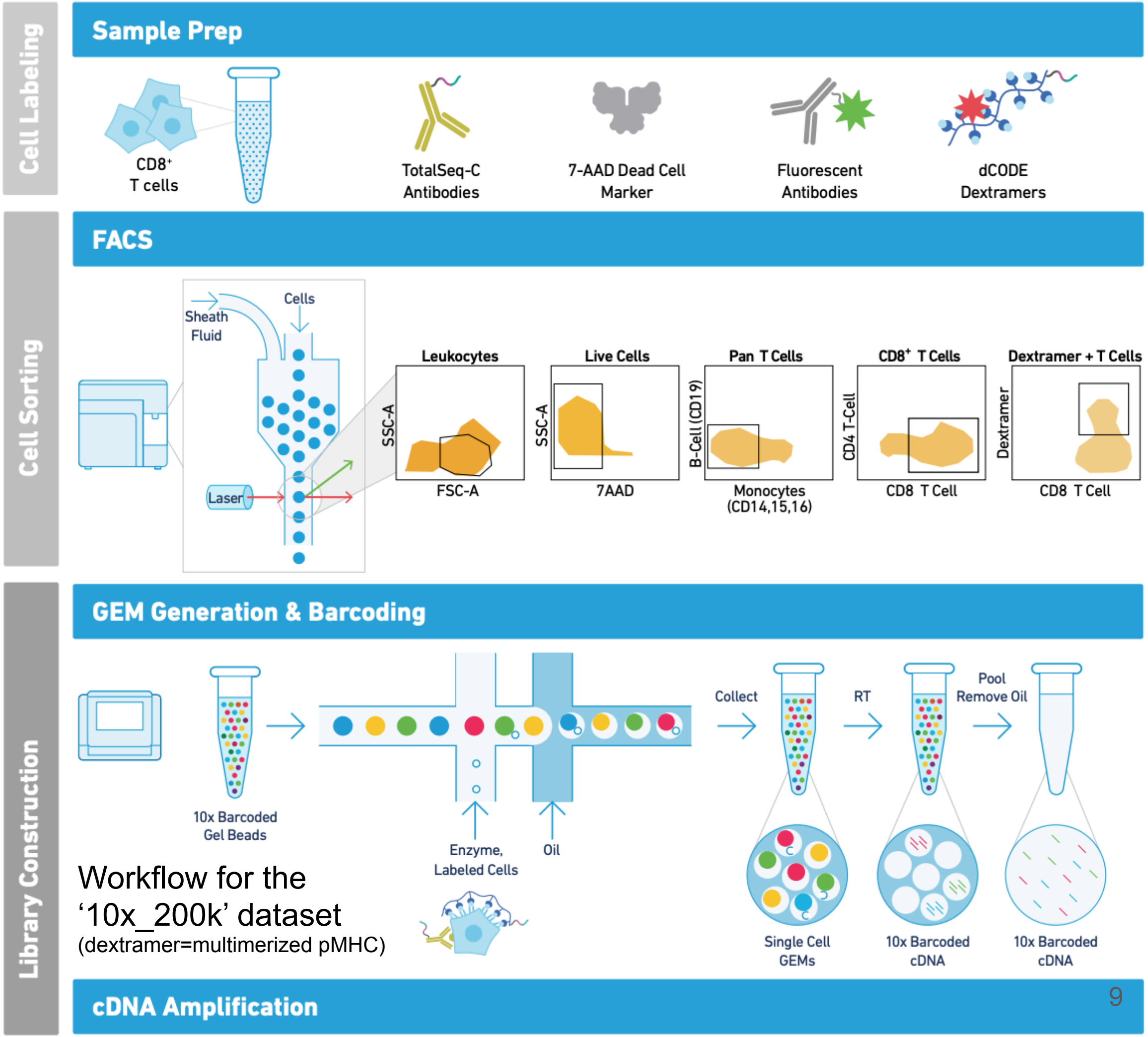
Single-cell T cell data

- Single-cell experiments make it possible to profile gene expression across thousands to millions of individual cells
  - mostly by attaching unique DNA ‘barcodes’ (1/cell) to cDNAs
- By attaching DNA barcodes to other things like antibodies or pMHCs we can profile additional cellular features in the same experiment
  - cell surface protein expression (anti-CD4, anti-CD8, anti-PD1, anti-CCR7)
  - TCR binding specificity (barcoded pMHCs like A*02:01-Flu/M1, B*08:01-EBV/BZLF1, …)
- Single-cell gene expression coverage is very sparse, but by including targeted primers we can focus on specific transcripts
  - like the TCR alpha and beta chains

These ‘multimodal’ single-cell technologies are advancing rapidly, with companies like 10x Genomics and academic labs releasing new protocols and publicly available datasets.

We can access publicly-available single-cell datasets covering millions of T cells, all with gene expression (GEX) and paired TCR sequences (TCR), many with surface protein expression, and several with pMHC binding profiles.

What can we learn from these kinds of datasets about the influence of the TCR sequence on cell phenotypes?



Workflow for the ‘10x_200k’ dataset (dextramer=multimerized pMHC)

Single-cell gene expression (GEX) data

Cells
#=5-10k

$$\begin{matrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&2&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&3&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&8&0&0&6&0&0&0&4&0&0&0&0&3&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{matrix}$$

Genes: #=30k

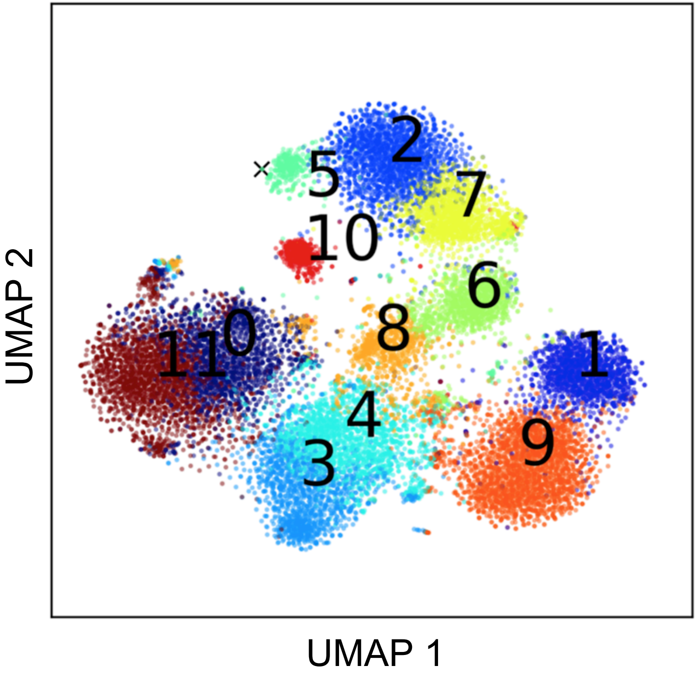
These matrices are typically very sparse (~98% zeros)

Two important techniques for scGEX data analysis: dimensionality reduction and clustering

The raw gene expression data present in the matrix of gene counts is high-dimensional and hard to digest. To help visualizing/analyzing scGEX data it can be useful to project the data into a lower-dimensional space (typically 2D). Popular methods for doing so are tSNE (van der Maaten & Hinton) and UMAP (McInnes & Healy). Of course, projecting from 30,000D to 2D involves some information loss, but these methods are often surprisingly good at revealing structure in the data.

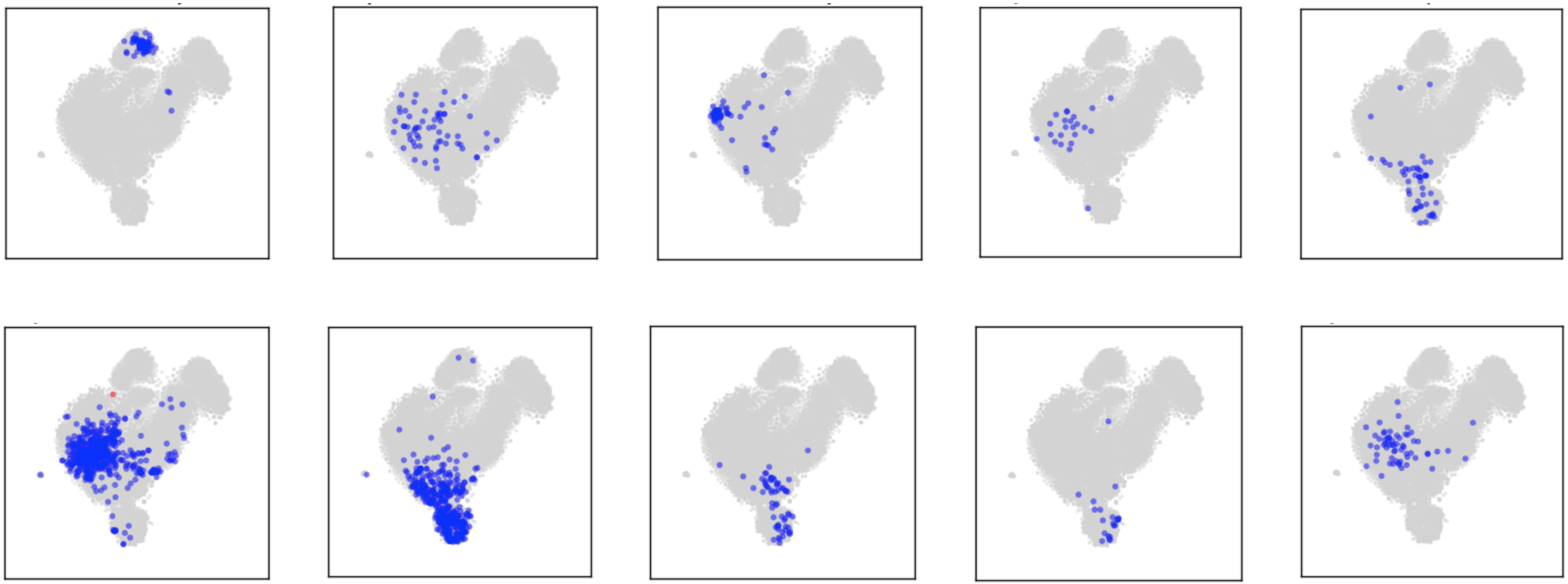
The scGEX data can also be clustered into groups of cells with similar gene expression profiles.

For both analyses, we exclude TRAV/TRBV genes.



Each point is a single cell, colored by cluster assignment, projected so as to preserve similarity relationships.

T cells occur in clonal families, the members of which descend from the same progenitor and share the same TCR sequence. Clonally related cells tend to have similar gene expression

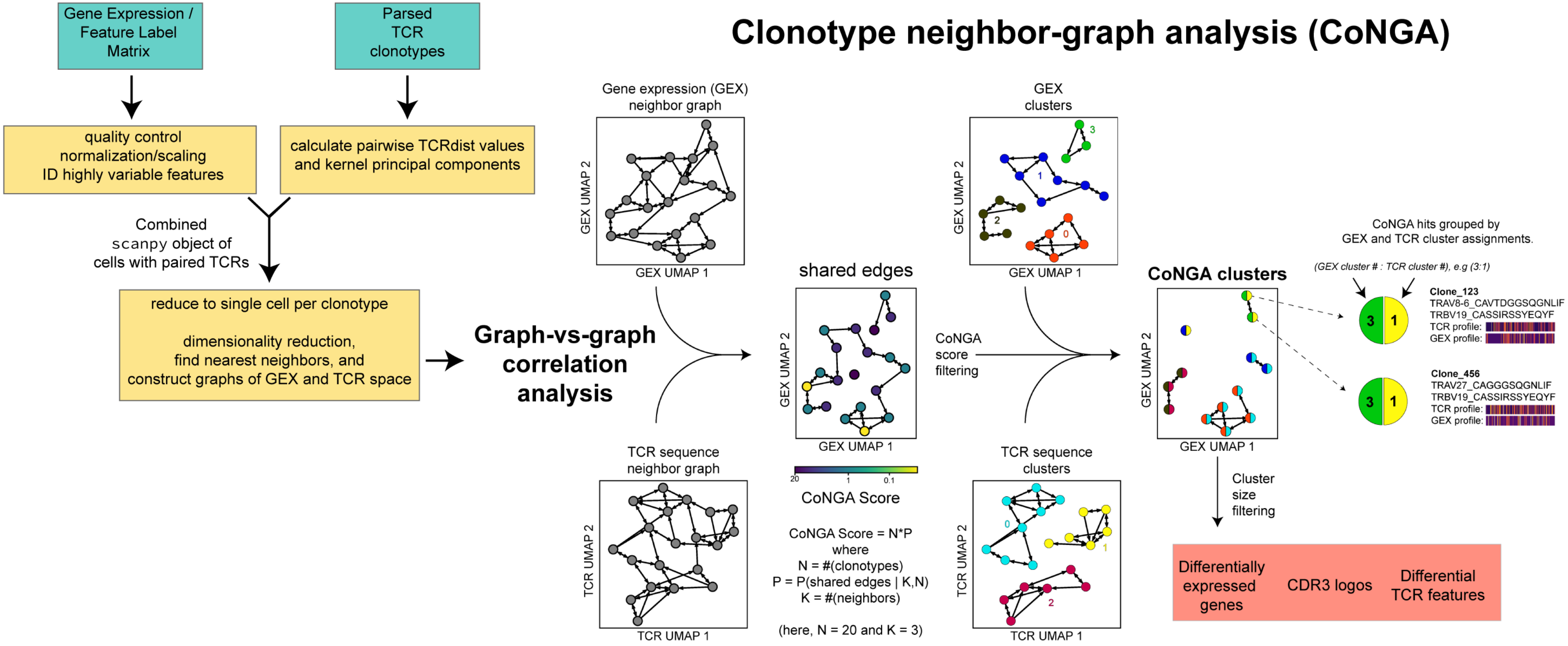


Clonally related cells (which all share the same TCR) are colored blue; other cells are gray. All cells projected to 2D based on GEX similarity. Each panel is a different clone.

Idea: look systematically for TCR/GEX correlations

- The simple idea is to ask whether cells that are nearby in TCR space are also nearby in GEX space, and vice versa. We formalize the notion of 'nearby' using k-nearest neighbor (kNN) graphs, defined based on distances between gene expression profiles or TCR sequences of T cells.
- Since clonally related T cells share identical TCRs and have similar GEX profiles, overlap of kNN graphs of ***cells*** will be dominated by intra-clonotype similarity
- To identify TCR/GEX correlation beyond clonal families, we need to factor out intra-clonotype similarity. We do this by picking a single representative cell for each clonotype (also tried averaging the GEX profiles of all the clones).
- Then compute TCR and GEX distances between ***clonotypes***, define kNN graphs based on these distances, and look for overlap between the graphs.

Clonotype neighbor-graph analysis (CoNGA)



Gene Expression / Feature Label Matrix

quality control
normalization/scaling
ID highly variable features

Parsed TCR clonotypes

calculate pairwise TCRdist values and kernel principal components

Combined `scanpy` object of cells with paired TCRs

reduce to single cell per clonotype

dimensionality reduction,
find nearest neighbors, and
construct graphs of GEX and TCR space

Graph-vs-graph correlation analysis

CoNGA Score

CoNGA Score = N*P
where
N = #(clonotypes)
P = P(shared edges | K,N)
K = #(neighbors)

(here, N = 20 and K = 3)

CoNGA score filtering

CoNGA hits grouped by GEX and TCR cluster assignments.

(GEX cluster # : TCR cluster #), e.g (3:1)

Clone_123
TRAV8-6_CAVTDGGSQGNLIF
TRBV19_CASSIRSSYEQYF
TCR profile:
GEX profile:

Clone_456
TRAV27_CAGGGSQGNLIF
TRBV19_CASSIRSSYEQYF
TCR profile:
GEX profile:

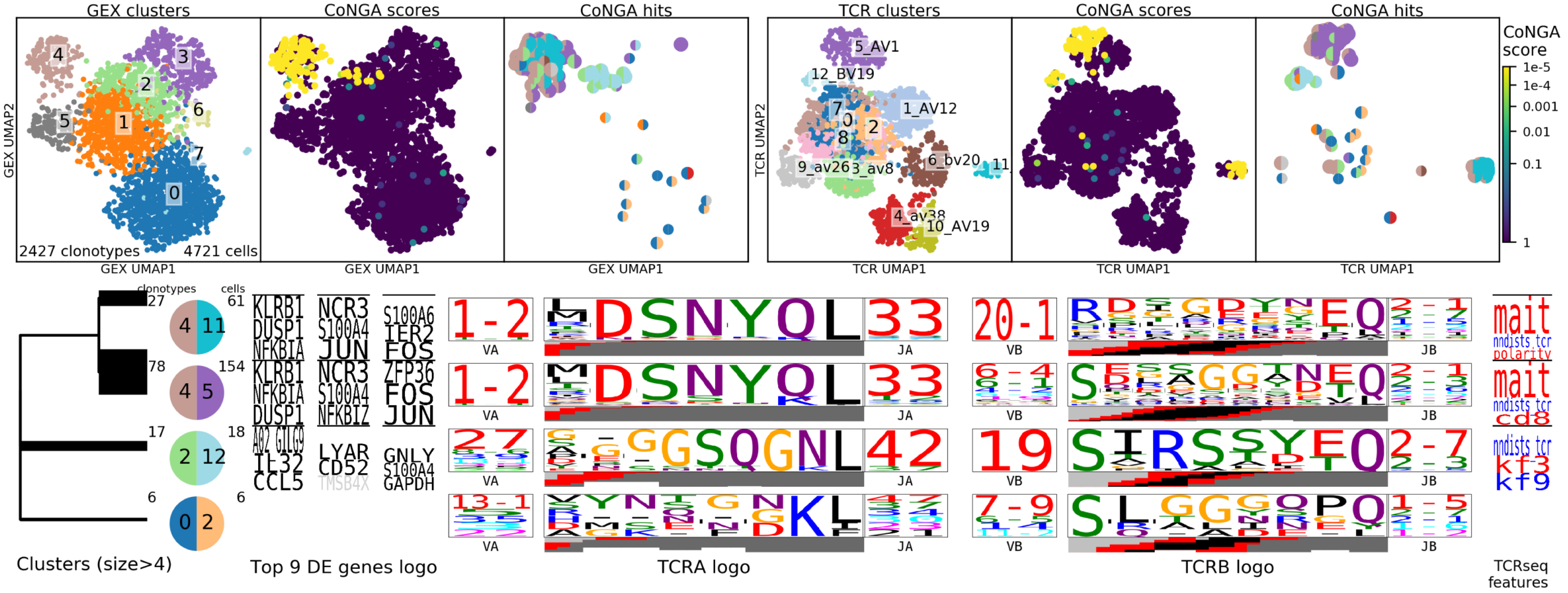
Cluster size filtering

Differentially expressed genes

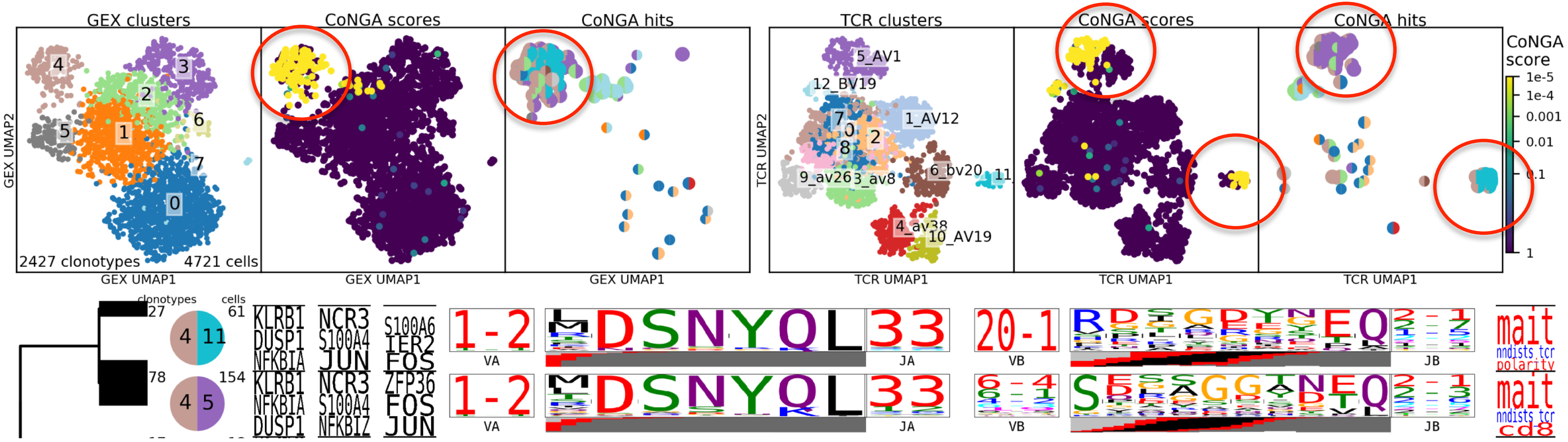
CDR3 logos

Differential TCR features

CoNGA results for a CD8+ T cell dataset from blood



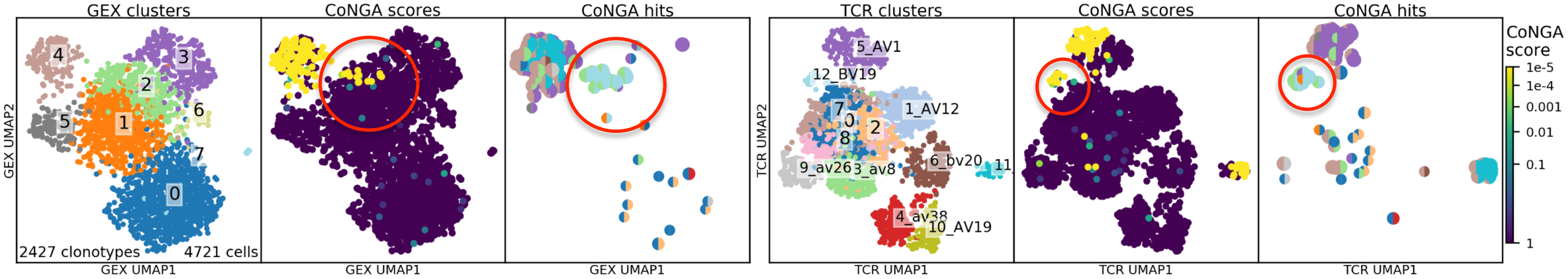
CoNGA results for a CD8+ T cell dataset from blood



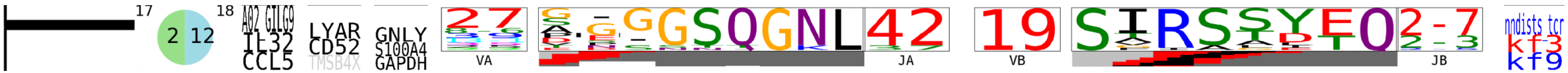
These are MAIT (mucosal-associated invariant T) cells, which bind MR1-presented metabolites.

MAIT cells typically have a nearly invariant TCR alpha chain (see TRAV1-2 and TRAJ33 above) paired with a more diverse (but still restricted) beta chain. Here two TCR clusters can be seen, differentiated by their TRBV gene.

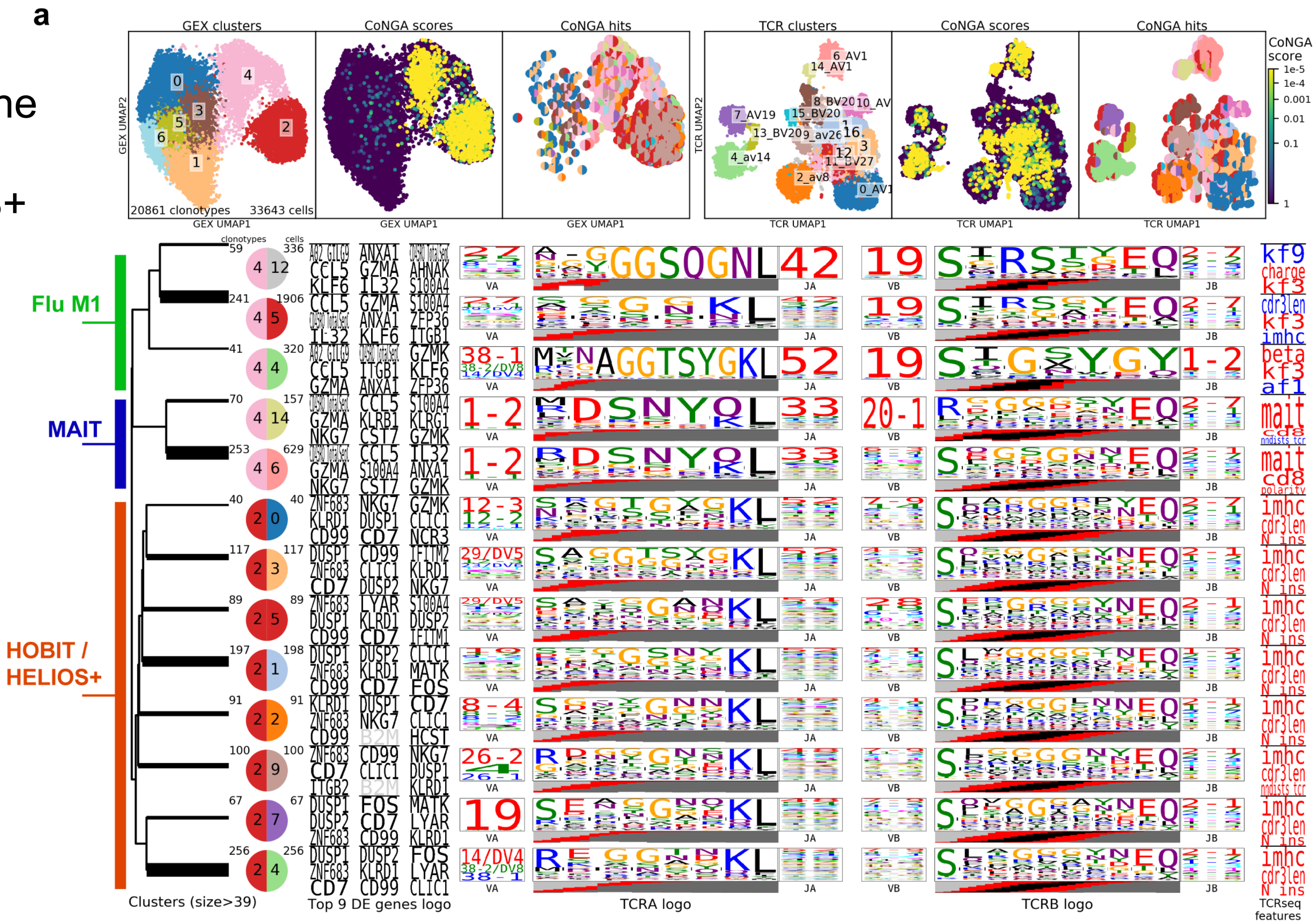
CoNGA results for a CD8+ T cell dataset from blood



These are Flu A*02:M1$_{58}$-positive T cells, based on their sequence features and also based on pMHC-binding data available for this dataset.



Donor 1 from the big ‘10x_200k’ dataset of CD8+ T cells



Interesting population of T cells with long CDR3 regions, expressing *ZNF683* (HOBIT), NK-related genes, HELIOS transcription factor

Comparing these HOBIT+ TCRs to the rest

Here we compare distributions of CDR3 sequence features between the HOBIT+ population and background TCR sequences

We can see that overall length (len_AB) and length of the CDR3beta (len_B) are at the top, then features relating to sequence composition: number of aromatics (aro_AB), number of tryptophans (W_AB and W_B), number of arginines in CDR3B (R_B) and overall (R_AB).

```
                             t-stat      tt pval     MWU  pval
feature len_AB               42.982 tt  0.00e+00 mwu  0.00e+00
feature len_B                40.286 tt  0.00e+00 mwu 1.09e-307
feature aro_AB               26.085 tt 3.70e-149 mwu 2.31e-105
feature aro_B                21.802 tt 5.73e-105 mwu  7.33e-68
feature len_A                21.331 tt 1.39e-100 mwu 3.61e-104
feature W_AB                 20.780 tt  1.44e-95 mwu  1.13e-78
feature W_B                  19.964 tt  2.20e-88 mwu  6.14e-75
feature R_B                  17.877 tt  2.72e-71 mwu  1.26e-54
feature R_AB                 16.225 tt  4.50e-59 mwu  5.24e-44
feature L_AB                 15.297 tt  1.01e-52 mwu  2.01e-41
feature charge_B             15.158 tt  8.35e-52 mwu  1.19e-40
feature L_B                  14.664 tt  1.34e-48 mwu  4.24e-36
feature chargefrac_AB        14.472 tt  2.18e-47 mwu  3.01e-50
feature aro_A                13.971 tt  2.76e-44 mwu  2.64e-34
feature charge_AB            13.905 tt  6.89e-44 mwu  5.41e-36
feature Wfrac_AB             13.719 tt  9.10e-43 mwu  6.09e-64
feature chargefrac_B         13.520 tt  1.36e-41 mwu  5.88e-46
feature Y_AB                 13.214 tt  8.30e-40 mwu  6.06e-32
feature F_AB                 12.498 tt  8.50e-36 mwu  8.53e-29
feature P_AB                 12.083 tt  1.42e-33 mwu  5.85e-28
feature Wfrac_B              12.010 tt  3.45e-33 mwu  1.77e-66
feature C_B                  11.615 tt  3.73e-31 mwu  7.23e-18
feature P_B                  11.302 tt  1.38e-29 mwu  1.54e-23
feature H_AB                 11.056 tt  2.18e-28 mwu  7.63e-26
feature F_B                  10.907 tt  1.13e-27 mwu  8.17e-24
feature V_AB                 10.441 tt  1.69e-25 mwu  4.35e-21
feature C_AB                 10.297 tt  7.64e-25 mwu  4.88e-13
feature Y_B                  10.258 tt  1.14e-24 mwu  5.53e-20
feature arofrac_AB           10.218 tt  1.73e-24 mwu  1.20e-18
feature H_B                   9.982 tt  1.91e-23 mwu  1.14e-22
```

(sorted by t test P value)

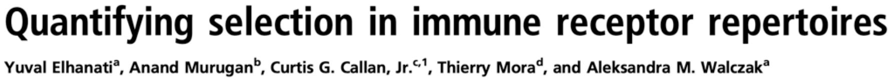
CDR3 sequence composition bias in these T cells suggests that they may not be MHC-restricted?

Molecular constraints on CDR3 for thymic selection of MHC-restricted TCRs from a random pre-selection repertoire

Jinghua Lu[1], François Van Laethem[2], Abhisek Bhattacharya[2], Marco Craveiro[2], Ingrid Saba[2], Jonathan Chu[1], Nicholas C. Love[2], Anastasia Tikhonova[2], Sergei Radaev[1], Xiaoping Sun[3], Annette Ko[3], Tomer Arnon[4,5], Eric Shifrut[4], Nir Friedman[4], Nan-Ping Weng[3], Alfred Singer[2] & Peter D. Sun[1]

“Thus, Cys is specifically excluded from the [CDR3] loops of MHCr repertoires but not MHCi repertoires. Less dramatic than differences in Cys usage, FGβ-loop usages of positively charged amino acids (Arg and His) and hydrophobic amino acids (Trp, Tyr and Pro) are significantly reduced in MHCr TCRs”

MHCr=MHC restricted, MHCi=MHC independent
FG loop = CDR3[4:-4]



Quantifying selection in immune receptor repertoires

Yuval Elhanati[a], Anand Murugan[b], Curtis G. Callan, Jr.[c,1], Thierry Mora[d], and Aleksandra M. Walczak[a]

Similar trends for length and Cys when fitting selection factors from data on in- vs out-of-frame repertoires.

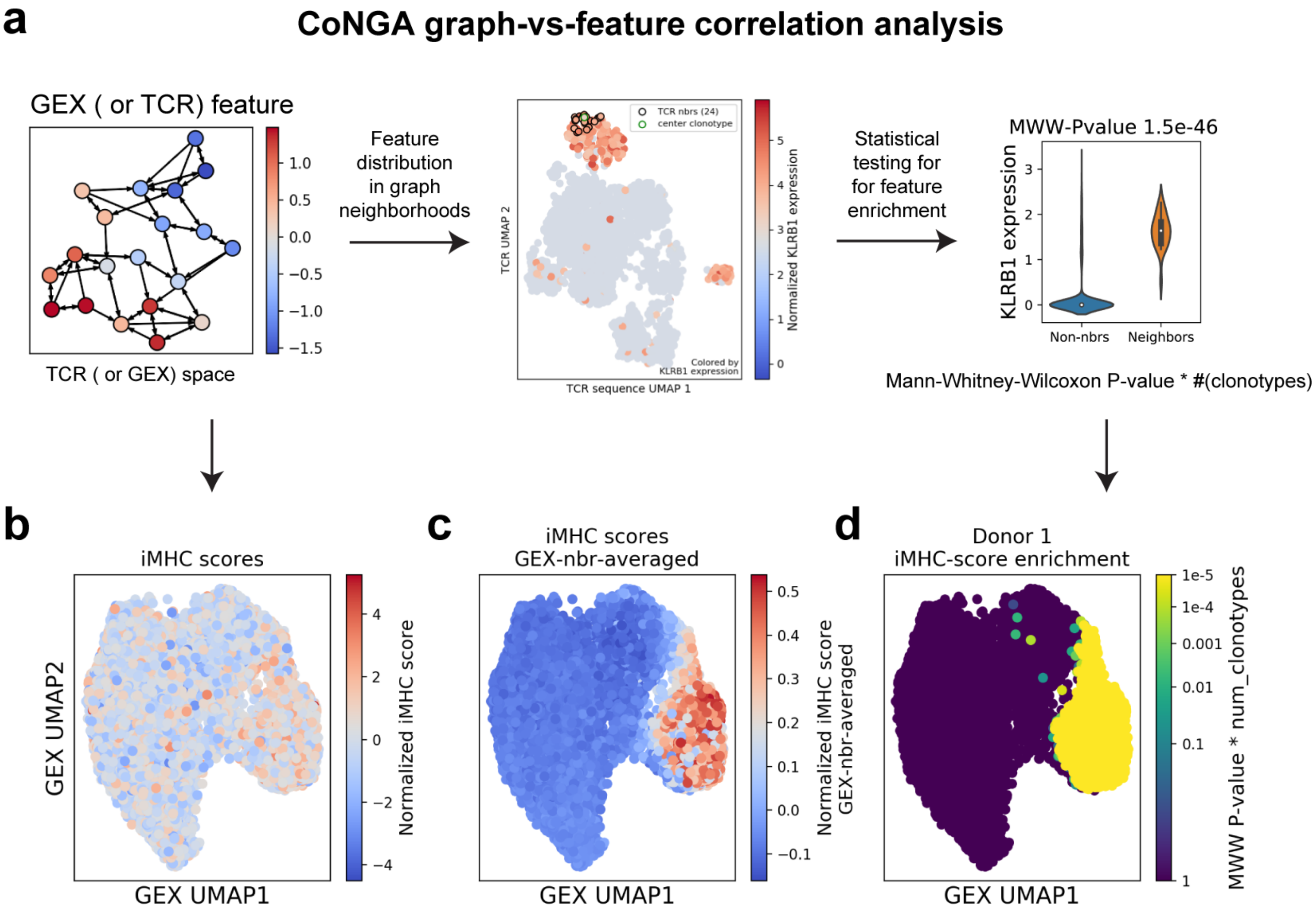
```
                          t-stat    t-pval    mwu-pval      enrichment
aafrac Wfrac_AB           13.719  9.10e-43  6.09e-64 enr     2.127
aafrac Wfrac_B            12.010  3.45e-33  1.77e-66 enr     2.343
aafrac Dfrac_AB           -9.544  1.43e-21  6.14e-24 enr     0.714
aafrac Cfrac_B             9.503  2.12e-21  7.42e-18 enr     8.158
aafrac Gfrac_B            -8.685  3.89e-18  1.70e-22 enr     0.814
aafrac Rfrac_B             8.587  9.15e-18  2.82e-26 enr     1.387
aafrac Cfrac_AB            8.391  4.92e-17  5.15e-13 enr     5.642
aafrac Gfrac_AB           -7.576  3.61e-14  9.88e-15 enr     0.894
aafrac Rfrac_AB            7.450  9.49e-14  6.56e-14 enr     1.283
aafrac Dfrac_A            -7.106  1.21e-12  1.02e-09 enr     0.673
aafrac Nfrac_A            -6.472  9.75e-11  5.18e-09 enr     0.790
aafrac Lfrac_AB            6.280  3.41e-10  3.05e-10 enr     1.225
aafrac Hfrac_AB            6.228  4.76e-10  4.81e-21 enr     1.600
aafrac Dfrac_B            -6.090  1.13e-09  2.93e-06 enr     0.752
aafrac Ffrac_AB            5.954  2.63e-09  5.90e-15 enr     1.348
aafrac Nfrac_AB           -5.740  9.49e-09  9.34e-11 enr     0.865
aafrac Tfrac_B            -5.698  1.22e-08  2.07e-07 enr     0.785
aafrac Kfrac_AB            5.583  2.37e-08  7.33e-17 enr     1.459
aafrac Wfrac_A             5.138  2.79e-07  1.50e-13 enr     1.727
aafrac Efrac_B            -5.061  4.18e-07  2.85e-03 enr     0.726
aafrac Hfrac_B             4.923  8.56e-07  6.24e-20 enr     1.595
aafrac Lfrac_B             4.832  1.36e-06  7.42e-11 enr     1.200
aafrac Ffrac_B             4.656  3.22e-06  2.86e-18 enr     1.441
aafrac Pfrac_AB            4.632  3.63e-06  1.48e-08 enr     1.184
aafrac Vfrac_AB            4.279  1.88e-05  2.59e-09 enr     1.216
aafrac Kfrac_B             3.836  1.25e-04  3.57e-13 enr     1.440
aafrac Pfrac_B             3.625  2.89e-04  1.06e-09 enr     1.175
aafrac Ifrac_AB            3.541  3.98e-04  3.26e-08 enr     1.250
aafrac Tfrac_AB           -3.492  4.79e-04  3.46e-06 enr     0.897
aafrac Ifrac_A             3.464  5.32e-04  2.20e-06 enr     1.382
```

Top sequence composition features by t-test P-value
('Wfrac_AB' = number of W in CDR3a and CDR3b divided by total length)

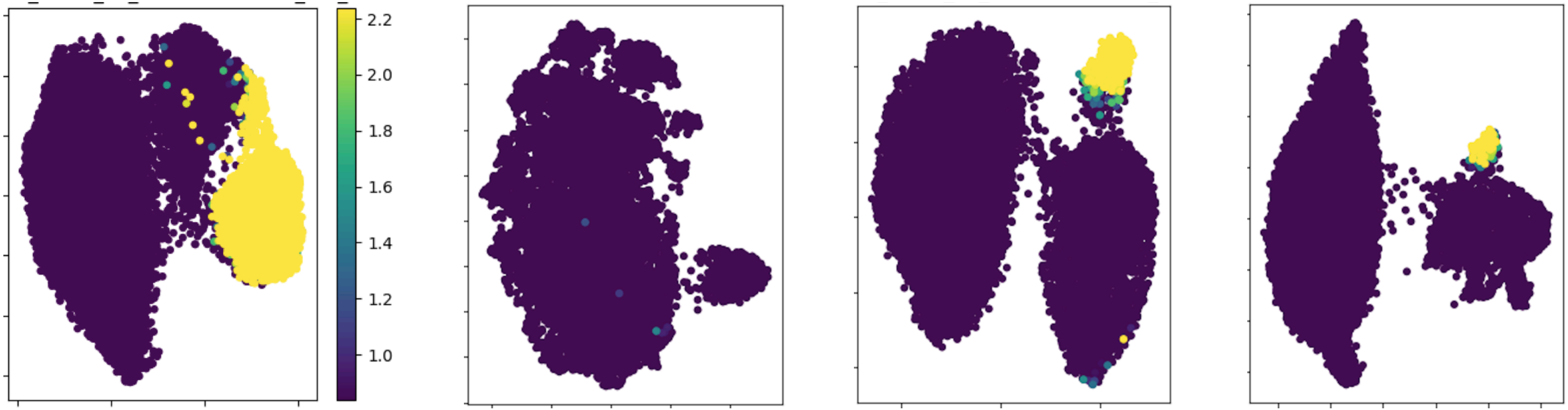
We used logistic regression to fit a simple CDR3 sequence score that captures the TCR biases seen in the HOBIT+ population.

Can we use this TCR-derived score to look systematically for gene expression neighborhoods with biased score distributions?

More generally, can we look for other TCR/GEX features that show biased distributions?


a
CoNGA graph-vs-feature correlation analysis
GEX (or TCR) feature
1.0
0.5
0.0
−0.5
−1.0
−1.5
TCR (or GEX) space
Feature distribution in graph neighborhoods
TCR nbrs (24)
center clonotype
TCR UMAP 2
TCR sequence UMAP 1
Colored by KLRB1 expression
Normalized KLRB1 expression
0
1
2
3
4
5
Statistical testing for for feature enrichment
MWW-Pvalue 1.5e-46
KLRB1 expression
0
1
2
3
Non-nbrs
Neighbors
Mann-Whitney-Wilcoxon P-value * #(clonotypes)
b
iMHC scores
GEX UMAP2
GEX UMAP1
Normalized iMHC score
4
2
0
−2
−4
c
iMHC scores GEX-nbr-averaged
GEX UMAP1
Normalized iMHC score GEX-nbr-averaged
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
d
Donor 1 iMHC-score enrichment
GEX UMAP1
MWW P-value * num_clonotypes
1e-5
1e-4
0.001
0.01
0.1
1

CoNGA graph-vs-feature analysis applied to the iMHC score, for all four donors in the 10x_200k set



Donor 1 (Age 30)

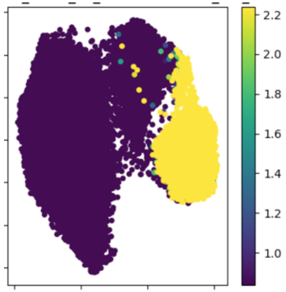
Donor 2 (Age 50)

Donor 3 (Age 38)

Donor 4 (Age 31)

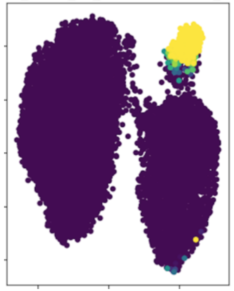
GEX UMAPs colored by adjusted P-value of iMHC-score enrichment in each clonotype's GEX neighborhood

Overlap among differentially expressed genes in these cells



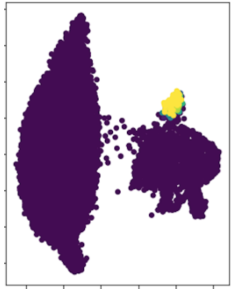
ZNF683
DUSP1
KLRD1
CD7
CD99
DUSP2
NKG7
CLIC1
LYAR
IL2RB

Donor 1



ZNF683
IL32
CD7
CTSW
CXCR3
NCR3
CD99
ZFP36L2
DUSP2
CLIC1

Donor 3



CD7
ZNF683
LYAR
DUSP2
KLRD1
CXCR3
NCR3
CD99
ZFP36L2
KLRC3

Donor 4

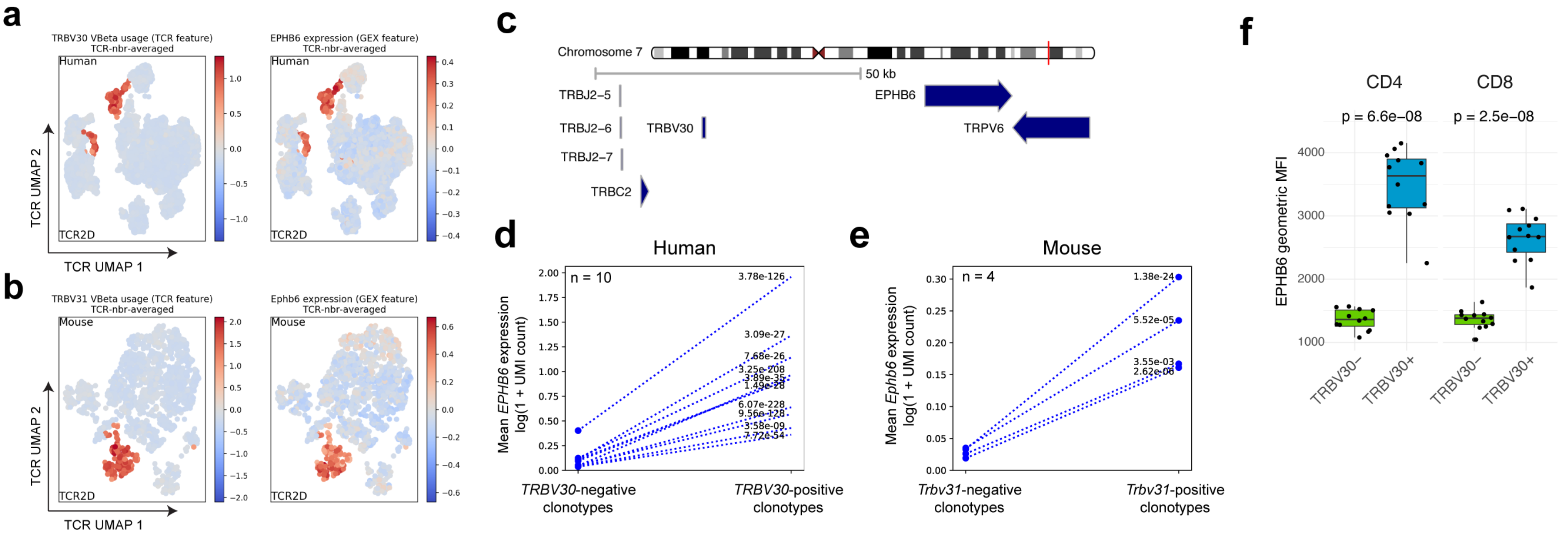
CoNGA graph-vs-feature analysis for GEX features

- We just saw that we can take a TCR-derived feature and look for neighborhoods in the GEX graph with skewed feature distributions
- We can do the same thing in reverse if we have a GEX-derived feature: we can look for neighborhoods in the TCR similarity graph with biased feature scores.
- A good place to start is with the expression levels of individual genes: we took each individual gene and mapped its expression pattern onto the TCR similarity graph. For each gene and each graph neighborhood (ie, clonotype and k nearest neighbors) we compared the distribution of the gene within that TCR graph neighborhood to the distribution outside that graph neighborhood, and looked for statistically significant differences.

Dataset	Gene	*P* value[a]	Enrich[b]	Cluster pair	Vα	Vβ	Invariant fraction[c]	Comment
human_pbmc1	NKG7	2.75e-54	3.86	*(5:8)*	TRAV1-2	TRBV6-4	0.71	MAIT
human_pbmc1	SLC4A10	2.69e-22	3.70	*(5:4)*	TRAV1-2	TRBV20-1	0.91	MAIT
human_pbmc1	GZMA	7.12e-13	4.18	*(0:8)*	TRAV1-2	TRBV6-4	1.00	MAIT
human_pbmc1	RP11-291B21.2	8.33e-04	1.56	*(2:3)*	TRAV14/DV4	TRBV7-9	0.00	CD8 naive?
human_pbmc2	SLC4A10	3.89e-120	6.29	*(4:9)*	TRAV1-2	TRBV6-4	1.00	MAIT
human_pbmc2	NKG7	1.91e-39	5.60	*(4:11)*	TRAV1-2	TRBV20-1	1.00	MAIT
human_pbmc2	CD8B	8.45e-05	1.25	*(2:3)*	TRAV14/DV4	TRBV19	0.00	CD4/CD8 preference
human_pbmc2	CD8A	4.20e-04	1.17	*(2:9)*	TRAV1-2	TRBV6-2	0.28	MAIT
human_pbmc2	S100A4	3.58e-03	0.81	*(4:5)*	TRAV1-1	TRBV20-1	0.45	MAIT
human_pbmc2	CD8B	4.98e-03	1.16	*(2:4)*	TRAV12-1	TRBV10-2	0.00	CD4/CD8 preference
mouse_pbmc	Cxcr6	5.68e-128	7.82	*(7:12)*	TRAV11	TRBV13-2	1.00	MAIT
mouse_pbmc	Ephb6	8.29e-18	3.31	*(2:4)*	TRAV6-6	TRBV31	0.00	EPHB6/TRBV30
mouse_pbmc	Wasf2	2.13e-04	1.19	*(1:0)*	TRAV10D	TRBV13-3	0.00	CD8 naive?
10x_200k_donor2a	SLC4A10	7.79e-64	5.12	*(4:5)*	TRAV1-2	TRBV6-4	0.86	MAIT
10x_200k_donor2a	KLRB1	2.87e-23	5.27	*(4:11)*	TRAV1-2	TRBV20-1	1.00	MAIT
10x_200k_donor2a	CCL5	2.77e-04	2.92	*(2:12)*	TRAV27	TRBV19	0.00	Flu M1
10x_200k_donor2a	HLA-C	4.16e-02	0.29	*(4:6)*	TRAV1-2	TRBV20-1	0.45	MAIT
10x_200k_donor1	SLC4A10	0.00e+00	6.24	*(4:14)*	TRAV1-2	TRBV20-1	0.98	MAIT
10x_200k_donor1	SLC4A10	0.00e+00	7.05	*(4:6)*	TRAV1-2	TRBV6-4	1.00	MAIT
10x_200k_donor1	LGALS3	1.02e-124	4.11	*(4:5)*	TRAV25	TRBV19	0.00	Flu M1
10x_200k_donor1	LGALS3	2.18e-81	3.72	*(4:12)*	TRAV3	TRBV19	0.00	Flu M1
10x_200k_donor1	ZNF683	2.30e-22	0.94	*(2:1)*	TRAV9-2	TRBV11-2	0.00	Hobit+
10x_200k_donor1	ITGB1	6.02e-20	1.92	*(4:0)*	TRAV12-2	TRBV19	0.00	Flu M1
10x_200k_donor1	ZNF683	6.09e-20	0.93	*(2:4)*	TRAV38-2/DV8	TRBV4-3	0.00	Hobit+
10x_200k_donor1	TRBC1	1.55e-19	0.61	*(0:1)*	TRAV36/DV7	TRBV13	0.00	V(D)J recombination
10x_200k_donor1	KLRD1	3.15e-19	0.85	*(2:3)*	TRAV13-2	TRBV11-2	0.00	Hobit+
10x_200k_donor1	GZMK	3.48e-19	0.84	*(2:5)*	TRAV20	TRBV19	0.00	Hobit+
10x_200k_donor2	SLC4A10	1.49e-207	5.25	*(8:5)*	TRAV1-2	TRBV6-4	0.86	MAIT
10x_200k_donor2	SLC4A10	1.33e-182	5.37	*(8:13)*	TRAV1-2	TRBV20-1	1.00	MAIT
10x_200k_donor2	KLRC1	4.47e-39	3.18	*(2:11)*	TRAV12-3	TRBV19	0.00	Flu M1
10x_200k_donor2	ITGB1	1.06e-31	1.15	*(2:6)*	TRAV38-2/DV8	TRBV19	0.00	Flu M1

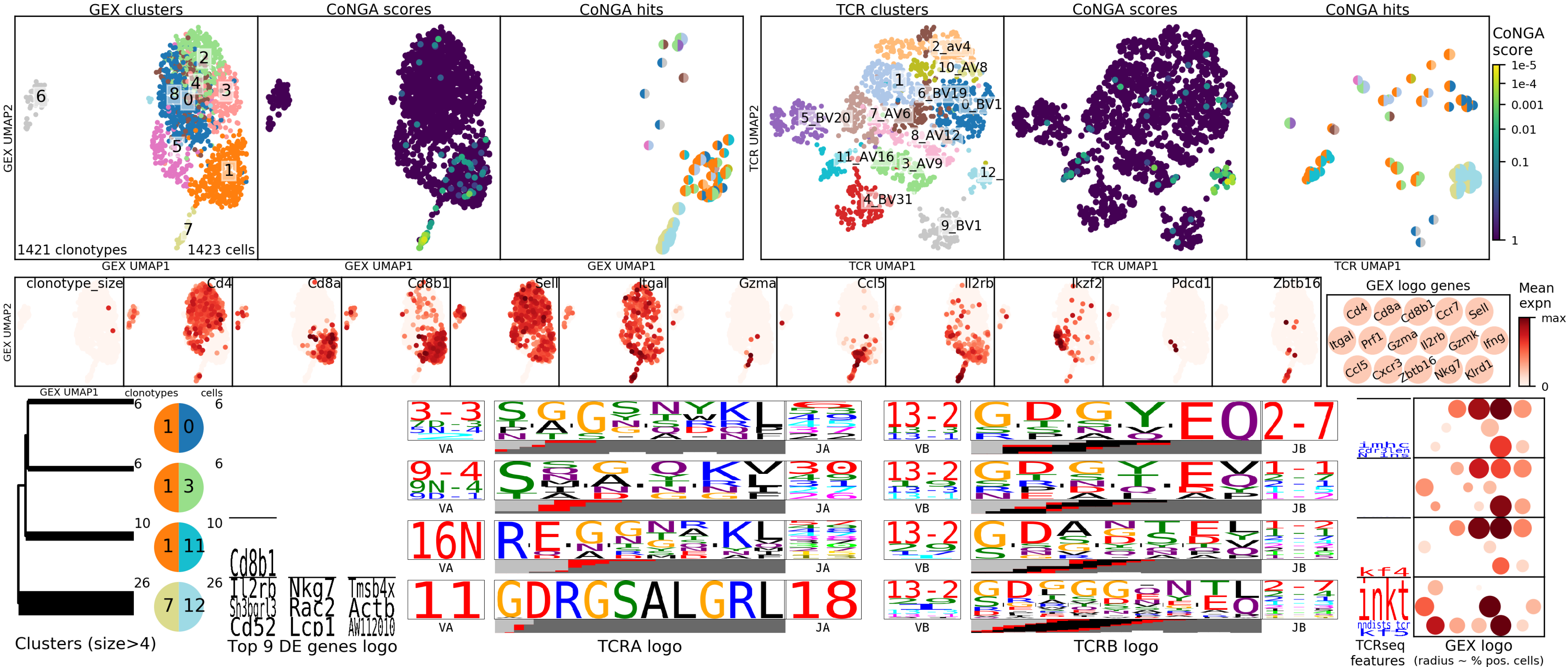
10x_200k_donor2	ITGB1	7.83e-24	1.04	*(2:3)*	TRAV8-3	TRBV19	0.00	Flu M1
10x_200k_donor2	CCL5	3.15e-20	0.97	*(2:1)*	TRAV12-2	TRBV19	0.00	Flu M1?
10x_200k_donor2	ITGB1	2.11e-18	2.12	*(9:11)*	TRAV35	TRBV19	0.00	Flu M1
10x_200k_donor2	GNLY	3.79e-18	3.13	*(2:18)*	TRAV12-3	TRBV19	0.00	Flu M1
10x_200k_donor2	HLA-DRB1	4.02e-13	2.32	*(1:2)*	TRAV13-1	TRBV12-3	0.00	EBV BZLF1
10x_200k_donor3	SLC4A10	0.00e+00	6.71	*(3:5)*	TRAV1-2	TRBV6-4	0.97	MAIT
10x_200k_donor3	KLRB1	1.63e-52	3.99	*(3:14)*	TRAV1-2	TRBV20-1	0.97	MAIT
10x_200k_donor3	GZMA	1.01e-22	2.48	*(2:5)*	TRAV1-2	TRBV6-4	0.73	MAIT
10x_200k_donor3	DAD1	5.82e-07	0.55	*(0:5)*	TRAV1-1	TRBV9	0.05	DAD1/TRAV1
10x_200k_donor3	TRBC1	1.06e-06	0.62	*(1:0)*	TRAV6	TRBV4-1	0.00	V(D)J recombination
10x_200k_donor3	GZMA	2.22e-06	1.88	*(2:4)*	TRAV14/DV4	TRBV18	0.00	other response
10x_200k_donor3	TRBC1	7.70e-06	0.59	*(2:0)*	TRAV39	TRBV6-5	0.00	V(D)J recombination
10x_200k_donor3	TRBC1	9.81e-05	0.58	*(0:0)*	TRAV26-2	TRBV4-1	0.00	V(D)J recombination
10x_200k_donor3	RPL34	6.34e-04	0.38	*(1:5)*	TRAV1-2	TRBV9	0.11	naive?
10x_200k_donor3	TRBC1	7.18e-04	0.55	*(1:3)*	TRAV12-3	TRBV14	0.00	V(D)J recombination
10x_200k_donor4	SLC4A10	0.00e+00	7.17	*(7:8)*	TRAV1-2	TRBV25-1	1.00	MAIT
10x_200k_donor4	EPHB6	3.10e-213	4.16	*(0:13)*	TRAV29/DV5	TRBV30	0.00	EPHB6/TRBV30
10x_200k_donor4	EPHB6	1.30e-66	3.75	*(1:13)*	TRAV12-3	TRBV30	0.00	EPHB6/TRBV30
10x_200k_donor4	GZMK	7.68e-35	2.95	*(7:7)*	TRAV1-2	TRBV20-1	0.67	MAIT
10x_200k_donor4	GZMK	7.06e-14	1.08	*(4:8)*	TRAV1-2	TRBV10-2	0.38	MAIT
10x_200k_donor4	CD3_TotalSeqC	8.55e-05	0.15	*(0:1)*	TRAV14/DV4	TRBV7-9	0.00	CD3↑ in TRAV14/38
10x_200k_donor4	TRBC1	4.40e-04	0.55	*(1:0)*	TRAV6	TRBV30	0.00	V(D)J recombination
10x_200k_donor4	TRBC1	1.38e-03	0.52	*(0:3)*	TRAV17	TRBV28	0.00	V(D)J recombination
10x_200k_donor4	TRBC1	1.21e-02	0.52	*(1:3)*	TRAV6	TRBV19	0.00	V(D)J recombination
thymus_atlas	HIST1H4C	4.11e-34	1.07	*(DP(P):13)*	TRAV41	TRBV19	0.00	DP(P) proliferation
thymus_atlas	DNTT	6.94e-28	1.30	*(DP(Q):13)*	TRAV41	TRBV19	0.00	DP(Q) TCR rearrangement
thymus_atlas	EPHB6	3.23e-26	2.82	*(CD4+T:0)*	TRAV10	TRBV30	0.00	EPHB6/TRBV30
thymus_atlas	EPHB6	1.88e-25	2.68	*(DP(Q):3)*	TRAV6	TRBV30	0.00	EPHB6/TRBV30
thymus_atlas	HIST1H4C	6.47e-25	0.91	*(DP(P):3)*	TRAV20	TRBV12-4	0.00	DP(P) proliferation
thymus_atlas	EPHB6	7.69e-24	2.67	*(CD4+T:3)*	TRAV6	TRBV30	0.00	EPHB6/TRBV30
thymus_atlas	EPHB6	8.18e-23	2.75	*(abT(entry):3)*	TRAV30	TRBV30	0.00	EPHB6/TRBV30
thymus_atlas	HIST1H4C	1.52e-22	0.78	*(DP(P):2)*	TRAV19	TRBV7-9	0.00	DP(P) proliferation
thymus_atlas	TSC22D3	1.59e-22	0.83	*(CD8aa(II):2)*	TRAV19	TRBV7-9	0.00	CD8αα(II)
thymus_atlas	EPHB6	5.51e-22	2.62	*(CD4+T:5)*	TRAV12-3	TRBV30	0.00	EPHB6/TRBV30

EPHB6 expression and TRBV30 usage are correlated

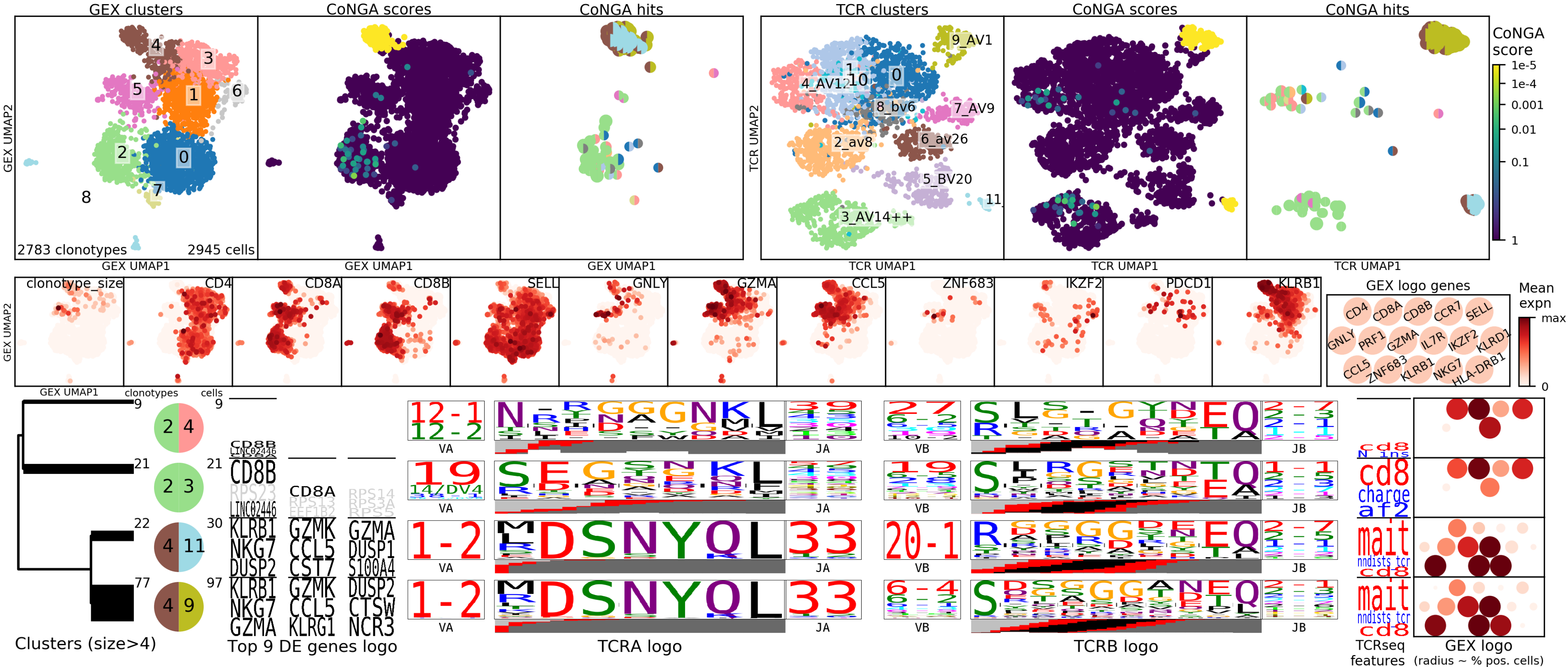


EPHB6 encodes Ephrin-B receptor 6, which has been found to play a role in T cell signalling

CoNGA graph-vs-graph results for mouse PBMC dataset



CoNGA graph-vs-graph results for human PBMC dataset



Conclusions

- Correlation analysis of T cell nearest neighbor graphs reveals known and potentially novel cell subsets and GEX/TCR relationships
    - MAIT and iNKT cells
    - CD8+ T cell sequence preferences
    - epitope-specific T cell subsets
    - a putative MHC-independent T cell subset with diverse but biased TCR sequences
    - correlations between V gene usage and expression of individual genes (*EPHB6, DAD1*)
- Multi-modal single-cell datasets offer many opportunities for algorithm development and biological discovery
- Decoding our information-rich T cell receptor repertoires may facilitate early detection/diagnosis of human disease

Acknowledgments

- Kate Guion (USC undergrad, Fred Hutch summer intern)
- Paul Thomas, Stefan Schattgen, and Jeremy Crawford (St. Jude)
- Mike Stubbington & Alvaro Martinez Barrio (10x Genomics)
- Evan Newell
- Erick Matsen