Learning the language of T cell receptor repertoires

Quantitative Immunology

February 2016

Benny Chain

Innate2Adaptive

Immunology with numbers

UCL

- **John Shawe-Taylor**
- **Yuxin Sun**

- Niclas Thomas
- **James Heather**
- **Theres Matjeka**
- **Mattia Cinelli**
- **Katharine Best**
- **Mazlina Ismail**
- **Connor Husovsky**
- **Kroopa Joshi**

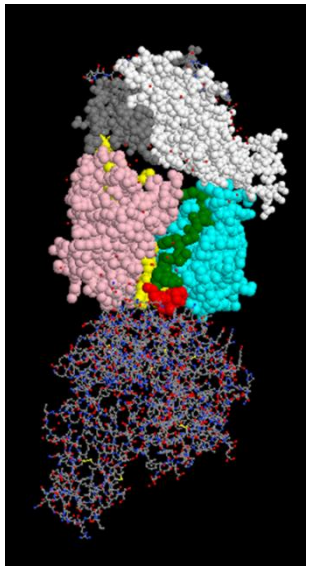
Weizmann Institute

- **Nir Friedman**

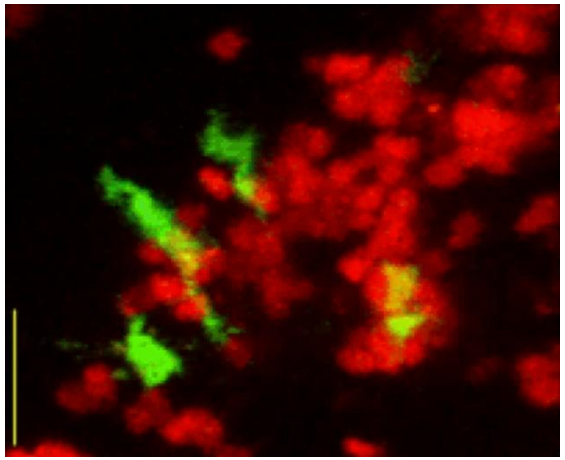
- Hila Gal
- Wilfred Ndifon
- **Shlomit Reich-Zeliger**
- **Eric Shifrut**
- **Michal Marks**

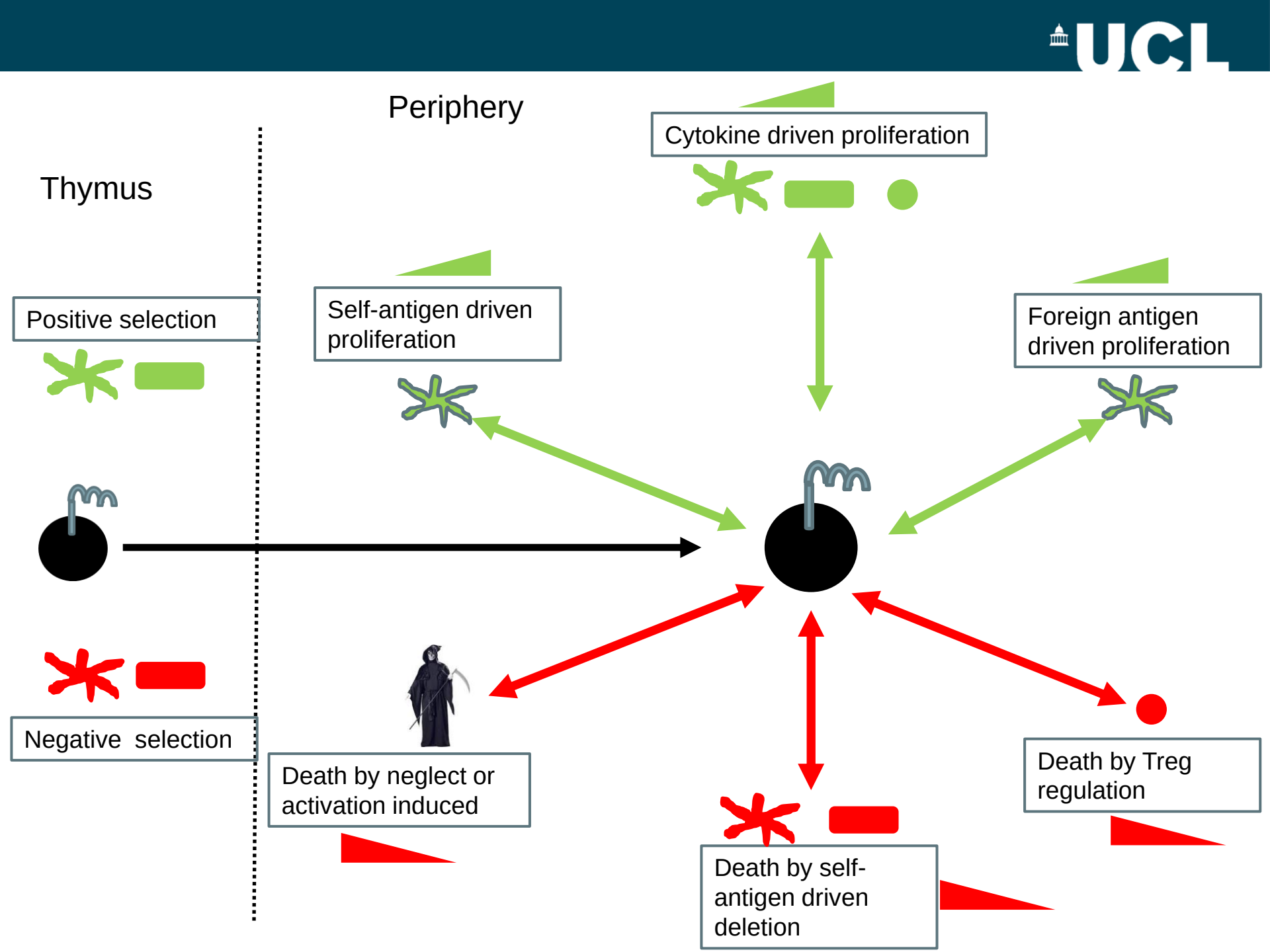
Studying the repertoire is an approach to understanding how T cells recognise antigen …

At an individual receptor level ?



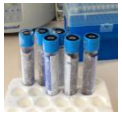
At a population level ?




Periphery
Thymus
Cytokine driven proliferation
Positive selection
Self-antigen driven proliferation
Foreign antigen driven proliferation
Negative selection
Death by neglect or activation induced
Death by Treg regulation
Death by self-antigen driven deletion

… and may provide biomarkers, tools for diagnosis and stratification , and correlates of protection or pathology for infectious diseases.

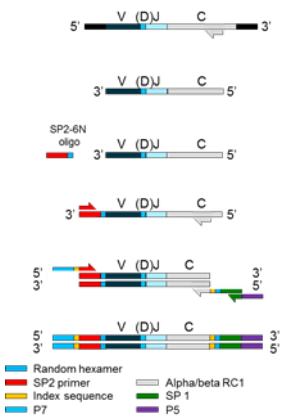
The TCR sequencing pipeline



Cells

RNA

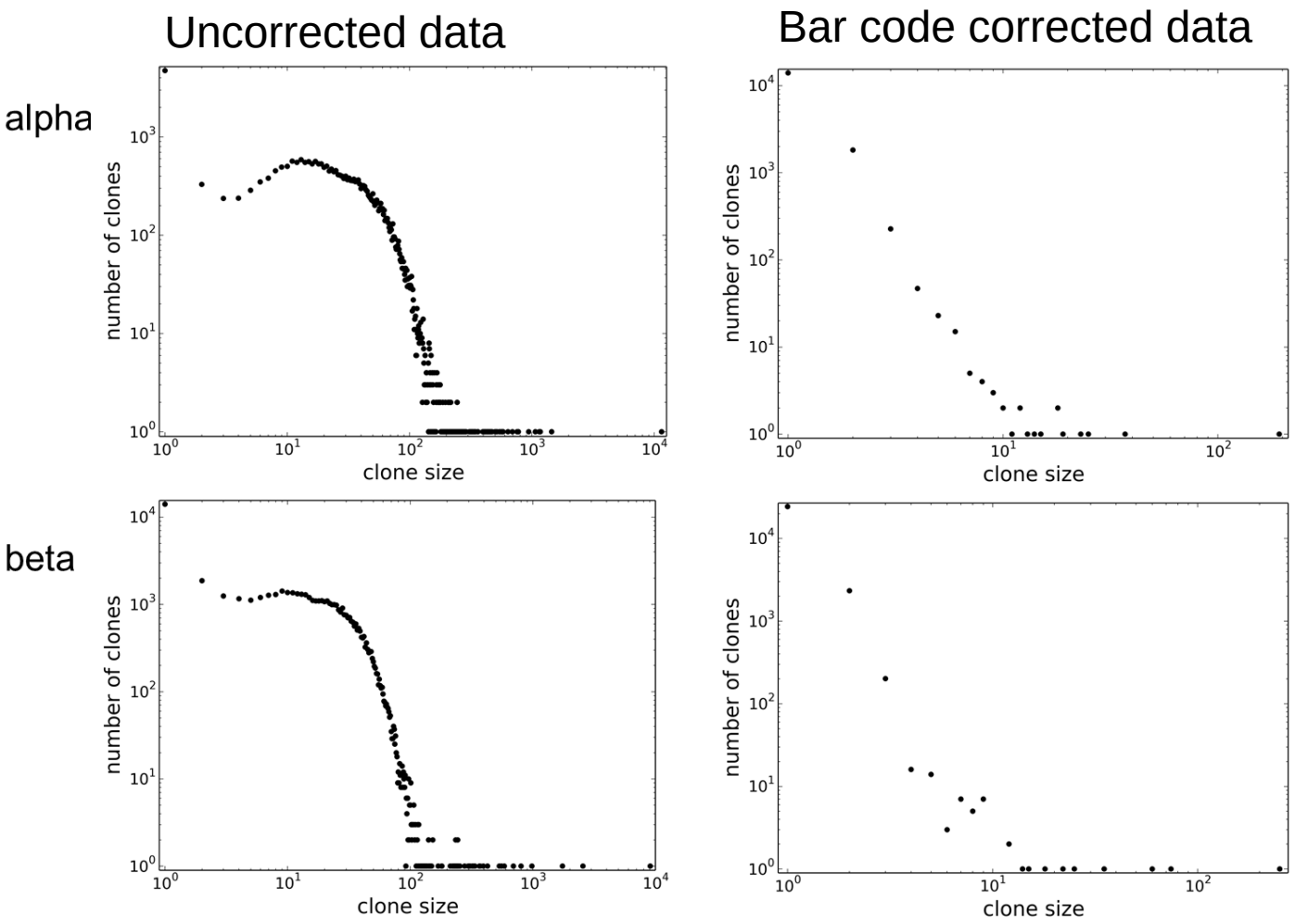
Amplified TCR cDNA



High throughput sequencing

Computational Analysis

Cord blood sample A



1: Best K, Oakes T, Heather JM, Shawe-Taylor J, Chain B. Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. Sci Rep. 2015 Oct 13;5:14629.

What is the T cell repertoire response to antigen ? Can we recognise antigen specific cells from sequences ?

- There are >10^{14} different alpha and beta chains.
- Each individual has undergone a unique set of stochastic recombination events
- Don’t know rules to go from sequence to structure to specificity
- HTS data does not link α and βTCR chains per cell

The experiment

β chain CDR3 TCR repertoire (NOT barcoded) sequenced from CD4 spleen cells.

9 unimmunised mice

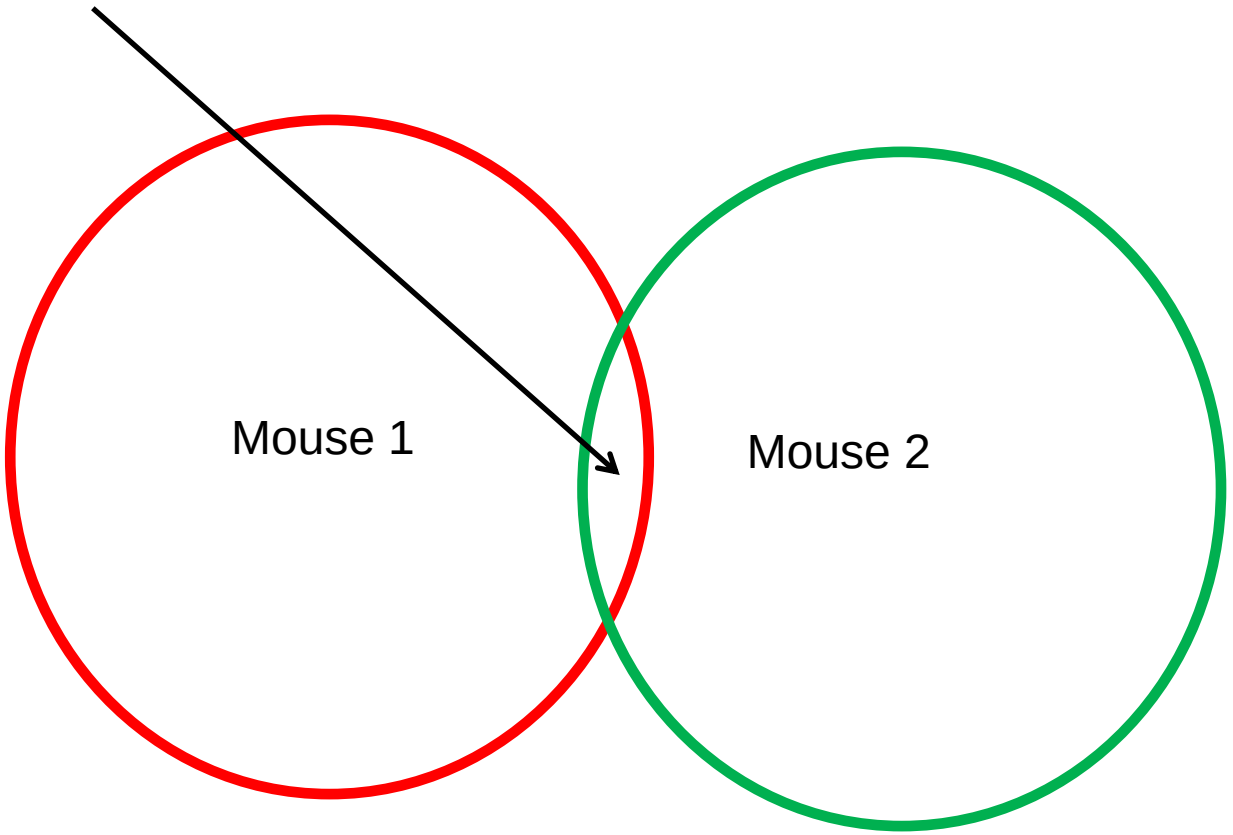
8/9 “early” immunised mice (5,7 and 14 days) immunised with Complete Freund’s Adjuvant (CFA) or ovalbumin with CFA

5/6 “late” mice (60 days) with additional boost at 14 days

Isolate CD4 cells, amplify (NO barcoding, multiplex) and sequence on HiSeq

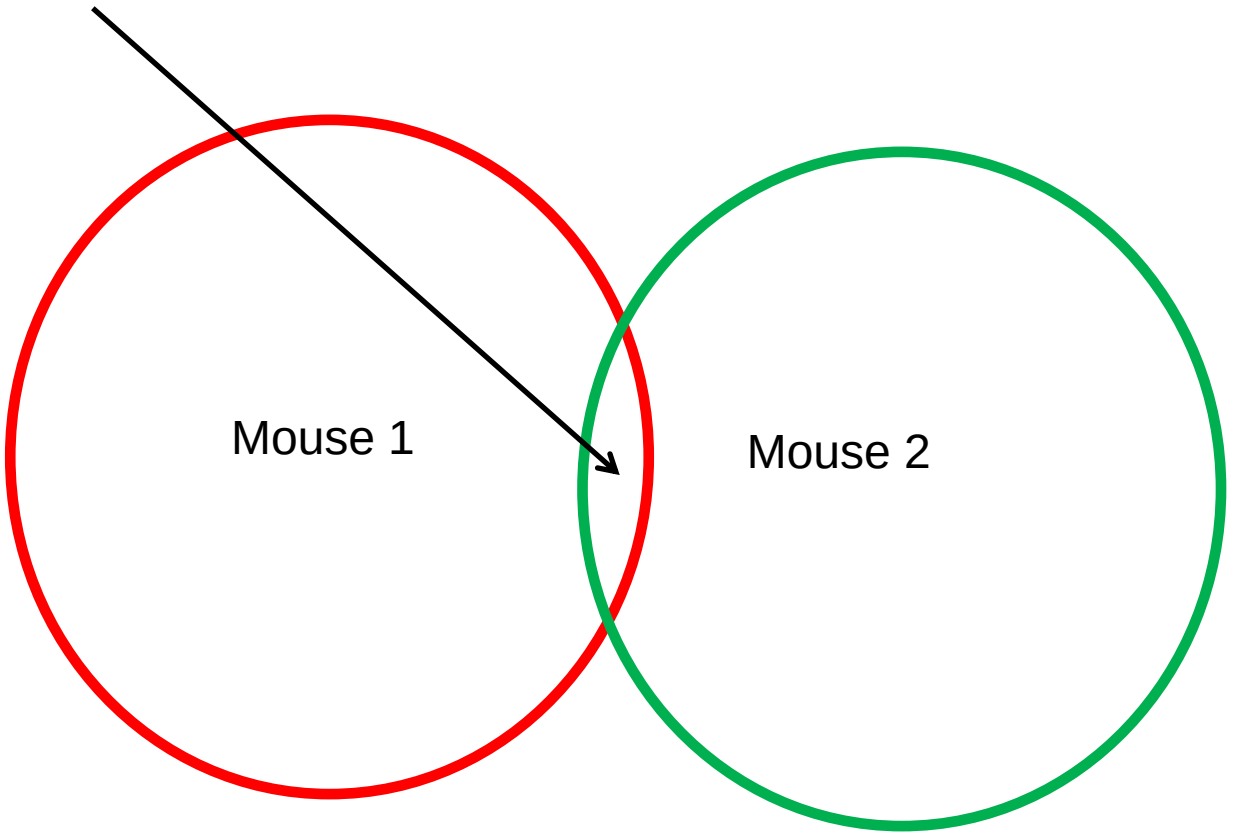
Simplistic approach

The Jaccard index measures overlap

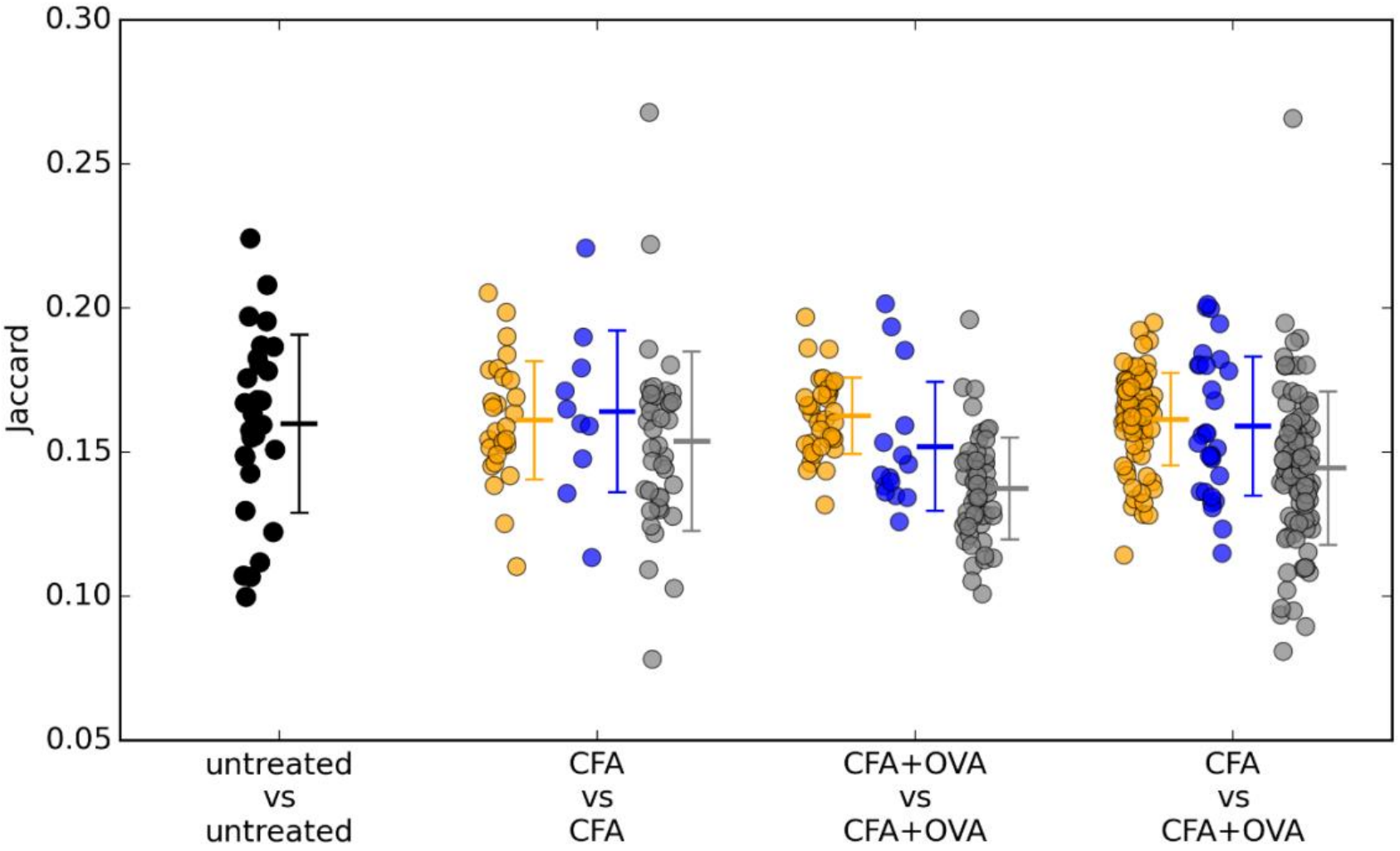


$J = (M1 \cap M2) / (M1 \cup M2)$

The Jaccard index measures overlap

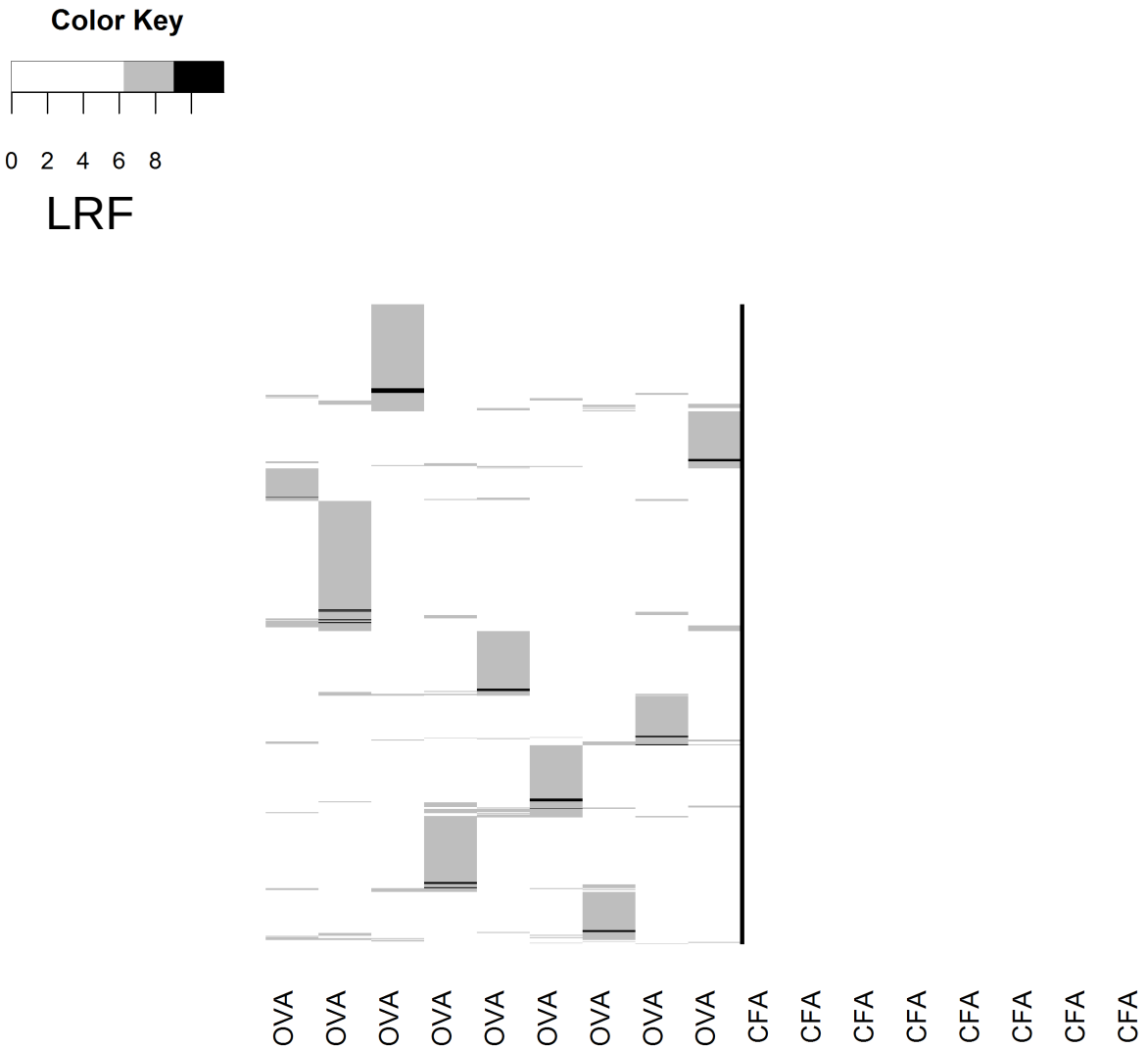


Does immunisation drive increased overlap ?


0.30
0.25
0.20
0.15
0.10
0.05
Jaccard
untreated vs untreated
CFA vs CFA
CFA+OVA vs CFA+OVA
CFA vs CFA+OVA

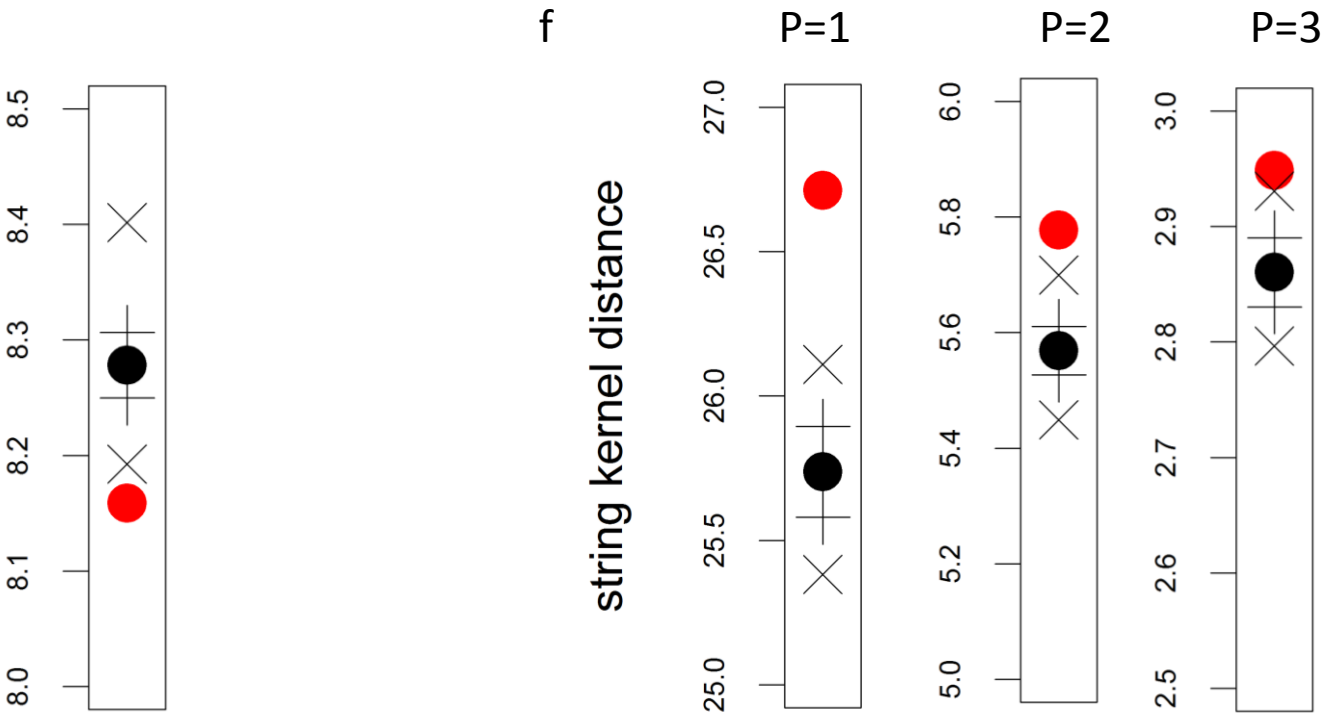
DIVERSITY AND PRIVACY

- Focus on early time point
- Measure abundance of each CDR3 in each immunised mice relative to average frequency in unimmunised mice and express as Log_2(relative frequency) (LRF)
- Identify >2000 sequences with LRF > 6 in OVA immunised mice, but < 2 in CFA only mice
- Plot LRF for each of this set of CDR3s in all mice


Color Key
0 2 4 6 8
LRF
OVA
OVA
OVA
OVA
OVA
OVA
OVA
OVA
OVA
CFA
CFA
CFA
CFA
CFA
CFA
CFA
CFA

SIMILARITY

Are there features of the repertoire which distinguish the OVA responsive repertoire ?



What should we use as features for TCR analysis ?

- Short amino acid motifs as features

CAISLAKDRAYNEQFF CAALRATGGNNKLTFGQG CASSLRATGNLTFGQG ...

CAI AIS ISL SLA LAK AKD ……



458
ITALIA



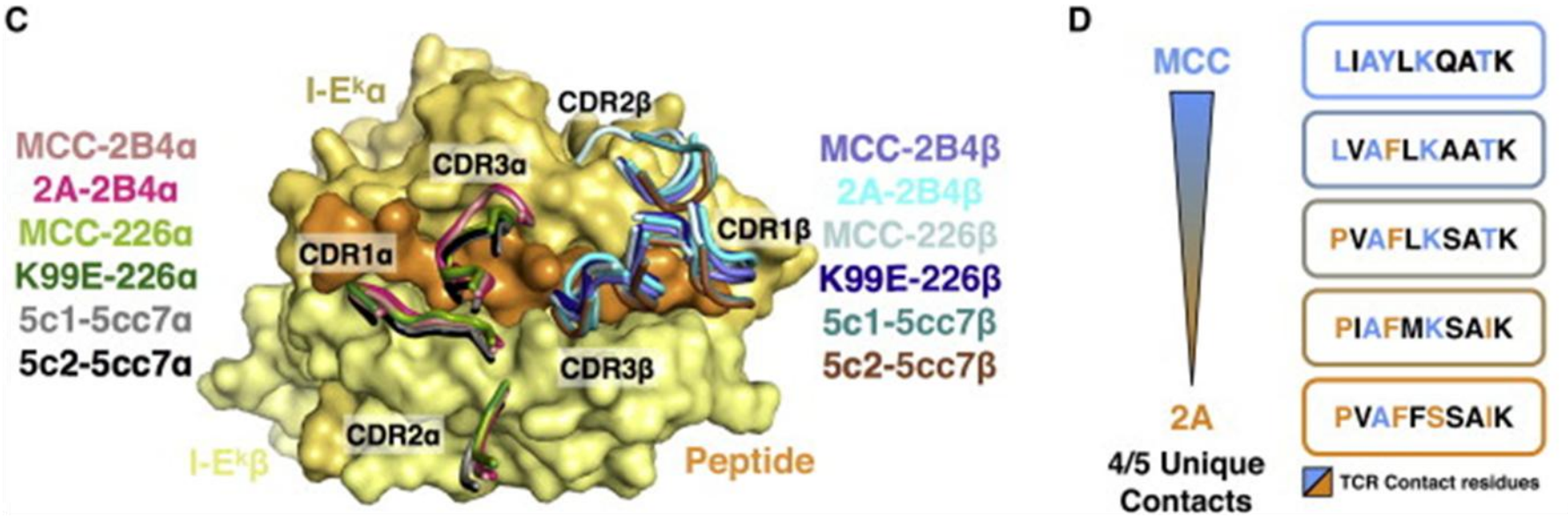




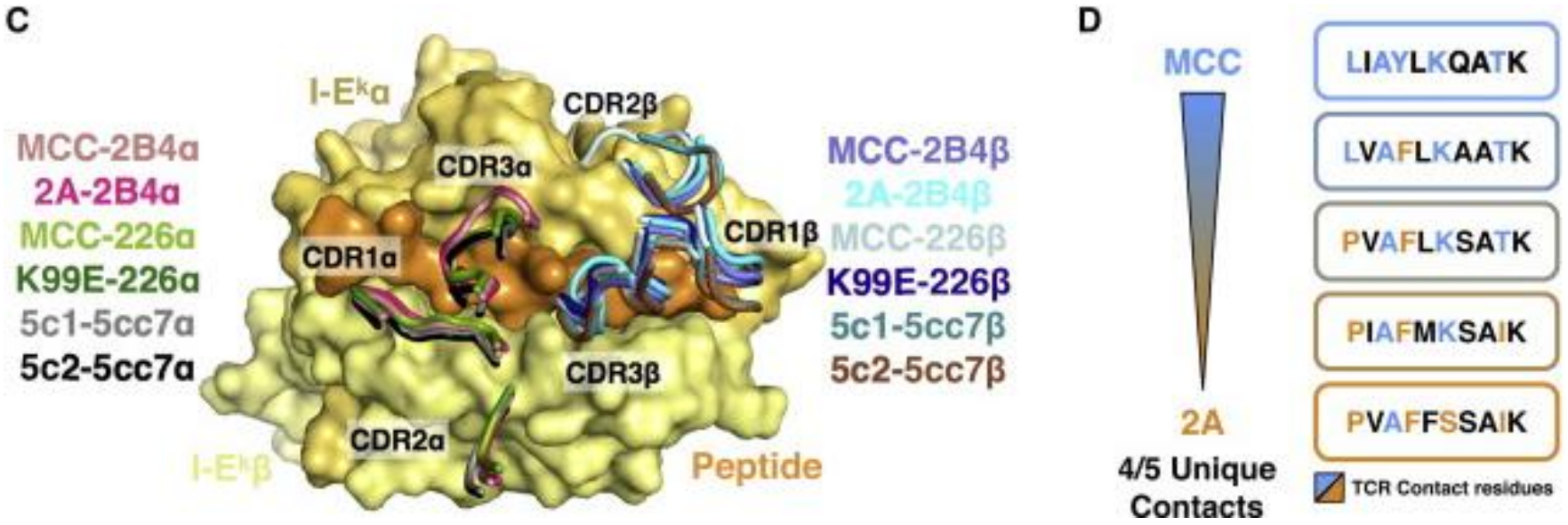



458
ITALIA

Some *post facto* motivation for this approach



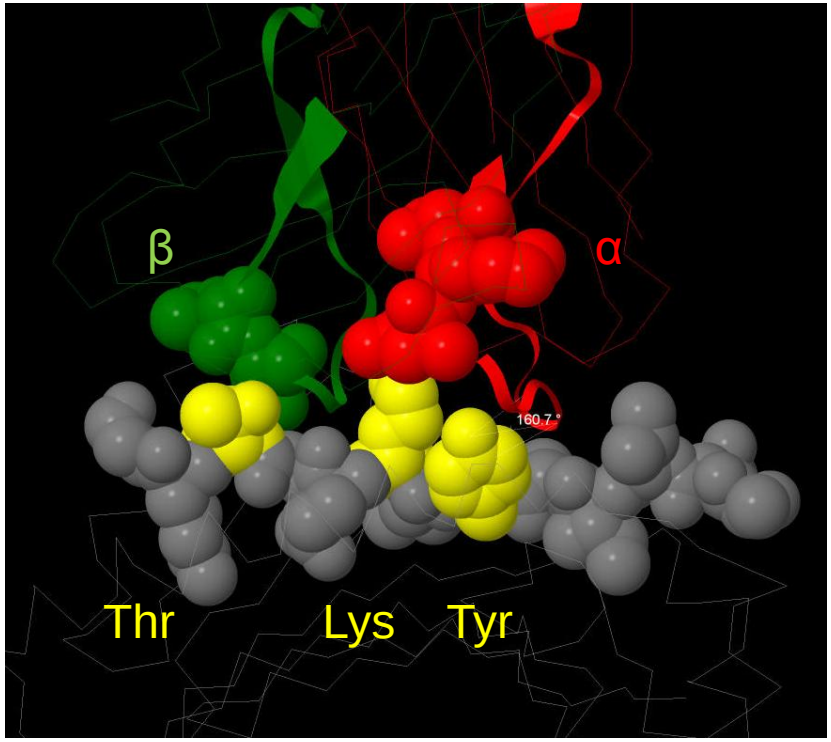
Deconstructing the peptide-MHC specificity of T cell recognition. *Birnbaum ME, Mendoza JL, Sethi DK, Dong S, Glanville J, Dobbins J, Ozkan E, Davis MM, Wucherpfennig KW, Garcia KC. Cell. 2014 May 22;157(5):1073-87*

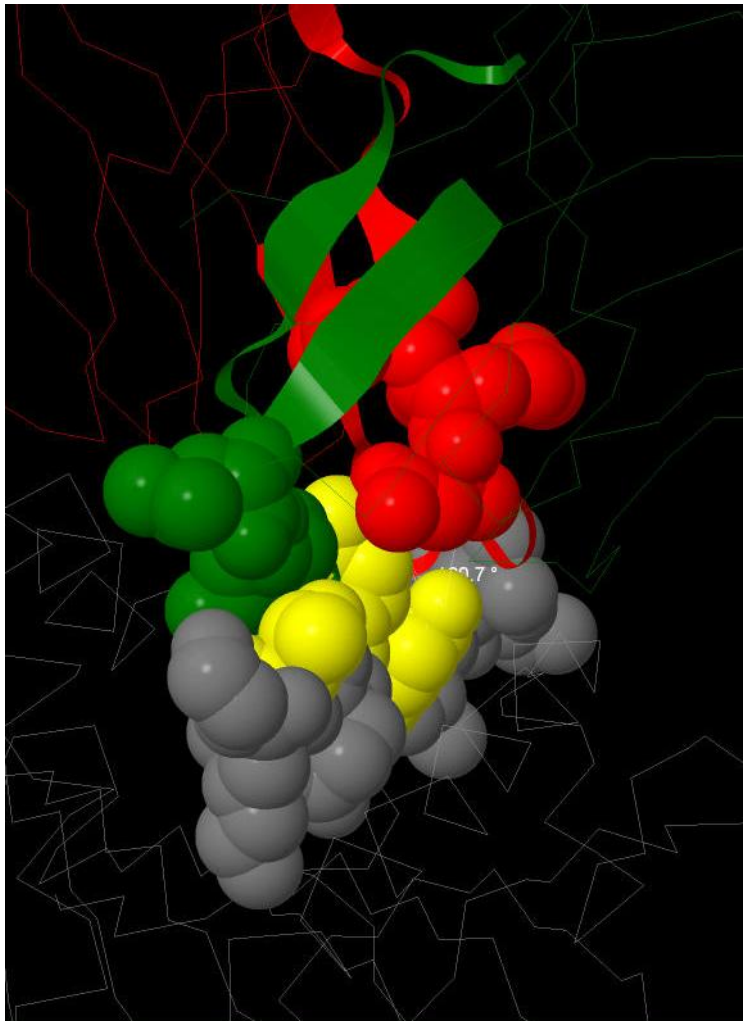


Cyt C peptide IAYLKQATKG

2B4	α	CAALRATGGNNKLTFGQG
226	α	CAAEPSSG--QKLVFGQG
5CC7	α	CAAE-AS-NTNKVVFGTG

2B4	β	CASSLNWSQDTQYFGPG
226	β	CASSLNNANSDYTFGSG
5CC7	B	CASSLNWSQDTQYFGPG





2B4 TCR binding cytochrome C peptide showing conserved residues

Selecting features
Beyond simple counting : a Markov process probability model for sequence generation

Consider a (protein) sequence (pseq) of length l as made up of a series of (contiguous) aa strings of length k.

Define $P_{u \to x}$ as the transition probability of moving from string u (length k-1) to character x.

We might consider the protein as a Markov process where the likelyhood of generating the protein is given by :

$\mathcal{L}(\text{pseq}) = \prod_{j=k}^{l} P(seq_{j-k+1:j-1} \longrightarrow seq_j)$ given a set of parameters $\Theta = P(u \longrightarrow x)$ for all u and x.

Given this generative model the Fisher score is defined as :

$$Fisher\ score\ = \frac{\partial\ log\ (\mathcal{L}(seq))}{\partial \Theta}$$

$$= \frac{\partial log \prod_{j=k}^{l} P(seq_{j-k+1:j-1} \longrightarrow seq_j\ )}{\partial P(u \longrightarrow x)}\ \forall\ ux \in \Theta$$

This is a vector indexed by all possible ux (k-mers) whose elements can be approximated by

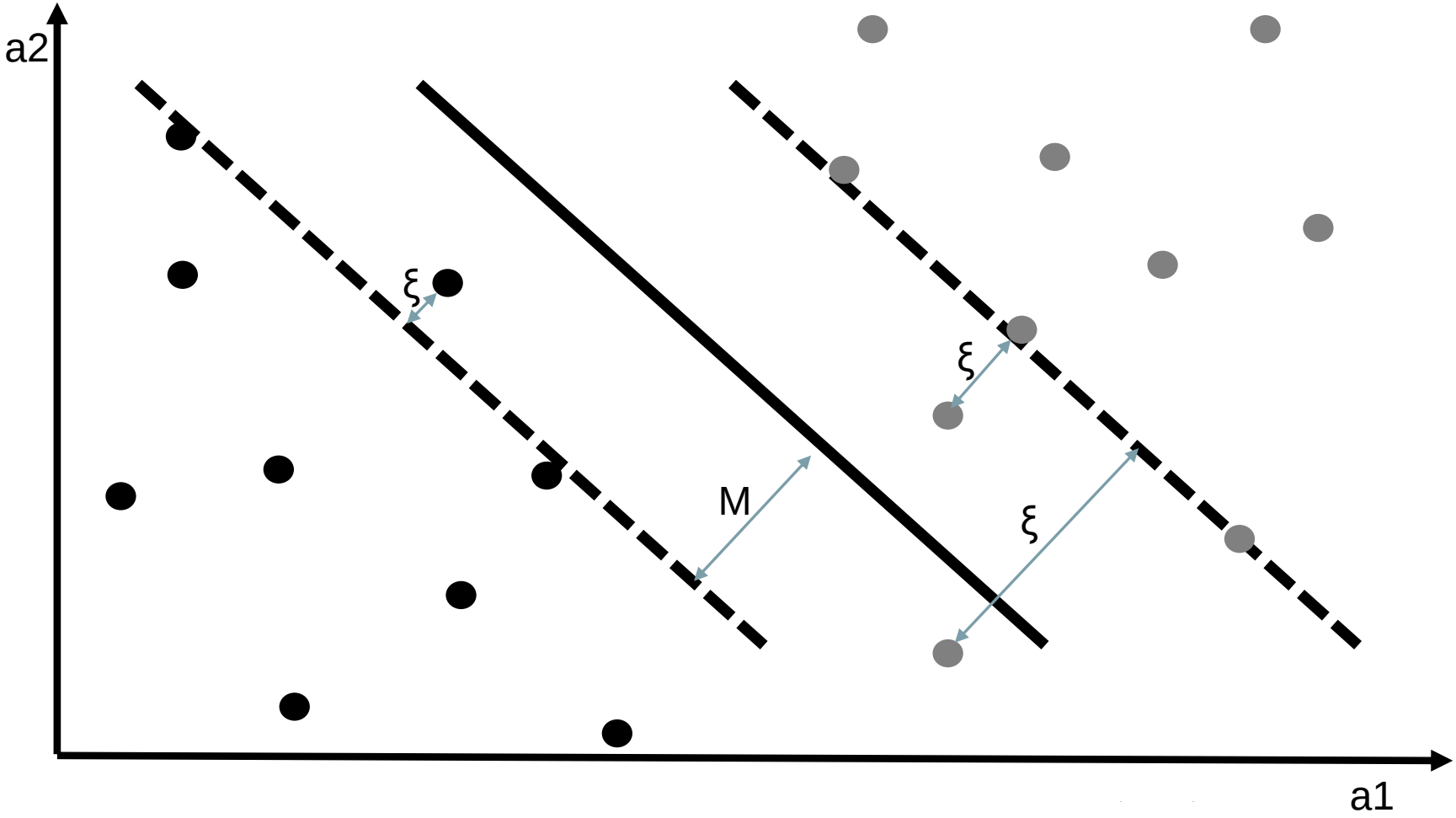
$$\frac{Term\ frequency\ (\iota}{P(u \longrightarrow x)}$$

Saunders, C., Vinokourov, A. & Shawe-taylor, J. S. String kernels, fisher kernels and finite state automata. in *Adv. Neural Inf. Process. Syst.* 633–640 (2002).

Practically, this corresponds to counting k-mers (e.g. triplets) and weighting according to some observed transition probability (or function of this probability)

Conceptually, the Fisher score can be considered as focusing attention on those features with the largest gradient in parameter space : i.e. those where a small change in parameter will have a large effect on the likelihood. This approach has proved fruitful in the context of image recognition.

Combining feature selection and classification



Looking for linear classifiers of the form : $f(\boldsymbol{x}) = \sum_1^m a_j H_j$

Linear programming boosting

Primal

$$\min_{a,\xi} \sum_{j=1}^{m} a_j + C \sum_{1}^{l} \xi_i$$

Such that

$$\sum_{j}^{m} a_j y_i H_{i,j} \geq 1 - \xi \quad \forall \quad i = 1 \ldots\ldots l$$

$$a_j \geq 0 \quad \forall\, j = 1 \ldots\ldots m$$

$$\xi_i \geq 0 \quad \forall\, i = 1 \ldots\ldots l$$

Dual

$$\max_{u} \sum_{1}^{l} u_i$$

$$\sum_{i=1}^{l} u_i y_i H_{ij} \leq 1 \quad \forall \quad \text{j = 1.....m}$$

$$0 \leq u_i \leq C \qquad \forall \quad \text{i = 1.....l}$$

Support vector machine

$$\min_{a,\xi} \frac{1}{2} \sum_{1}^{m} {a_i}^2 + C \sum_{1}^{l} \xi_i$$

$$\max_{u} \sum_{1}^{l} u_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{k=1}^{l} u_i u_k y_i y_k < H_i . H_k >$$

Such that

$$0 \leq u_i \leq C, \forall\, i = 1 \ldots\ldots l \qquad \sum_{i=1}^{l} u_i y_i = 0$$

Definitions

l	: number of training data points
m	:number dimensions (or weak learners or features)
H	:the set of features associated with each data point
ξ	:slack variables for data points within margin
a	: coefficients defining hyperplane (i.e. feature weights)
m	:sample weights
C	:penalty weights
y	:classification

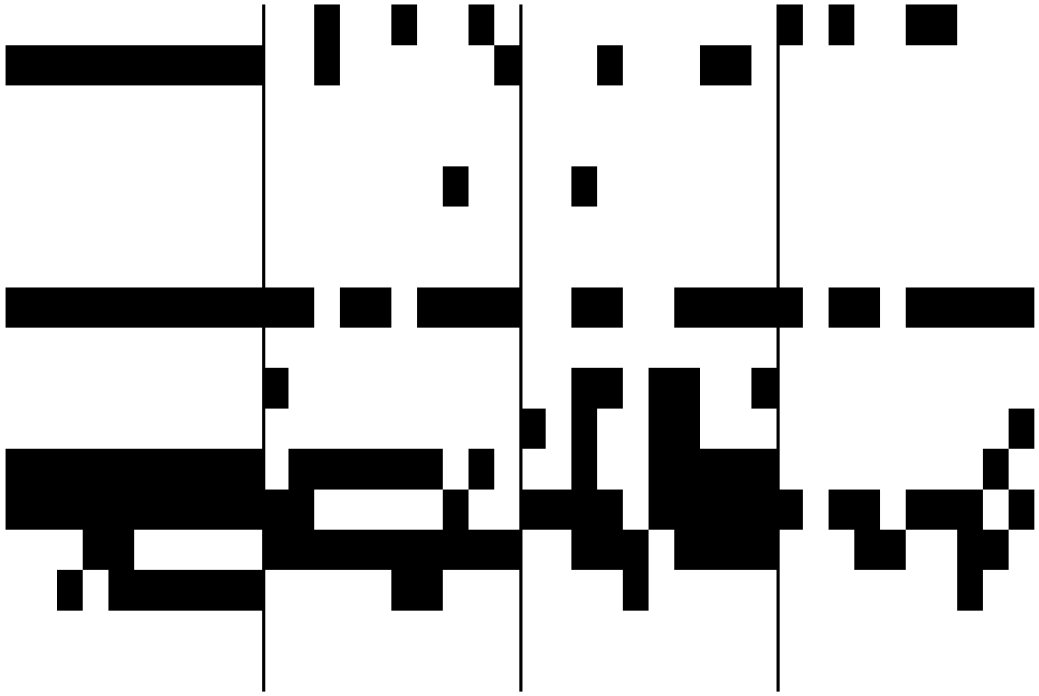
Demiriz, A., Bennett, K. P., & Shawe-Taylor, J. (2002). Linear programming boosting via column generation. Machine Learning, 46(1-3), 225-254.

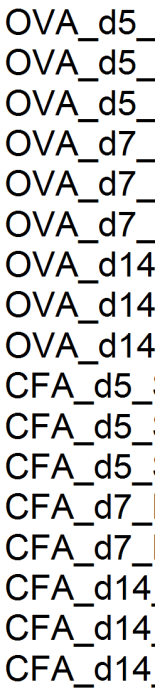
Linear Boost
Triplet counts

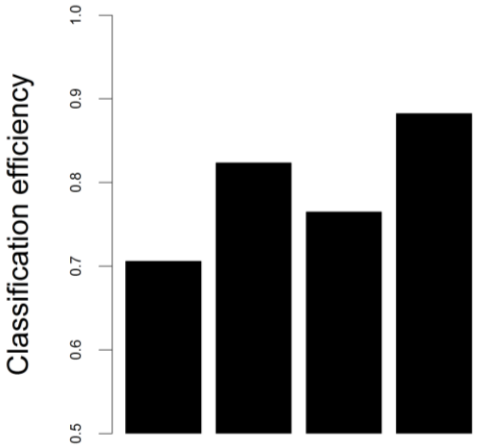
Linear Boost
Triplet counts + SVM

Linear Boost
Fisher scores

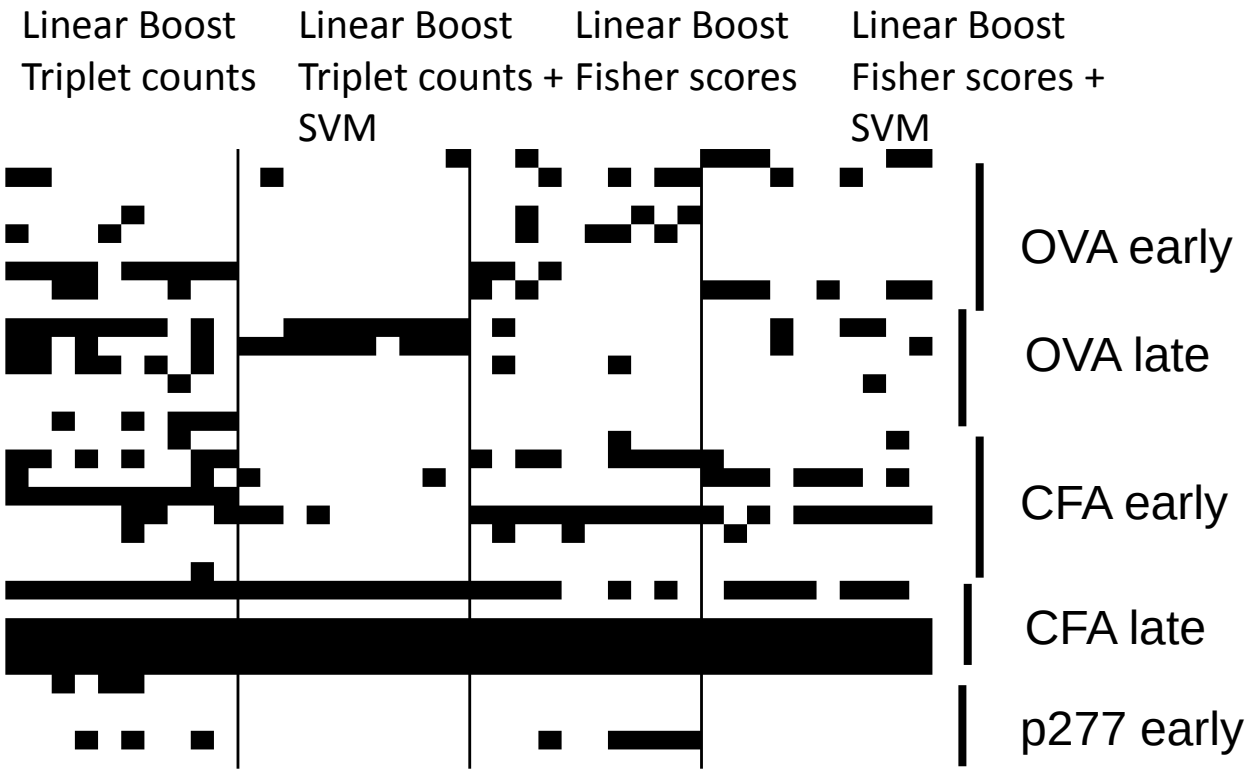
Linear Boost
Fisher scores + SVM

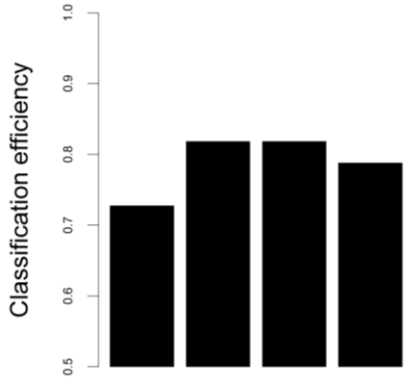






Linear boosting using 50 thousand CDR3s from each mouse, leave-one-out validation

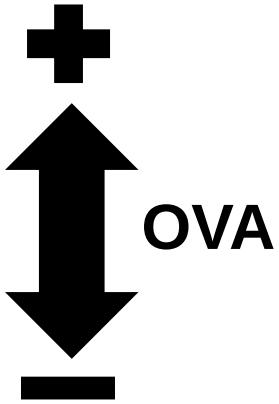




Linear boosting using 50 thousand CDR3s from each mouse, leave-one-out validation

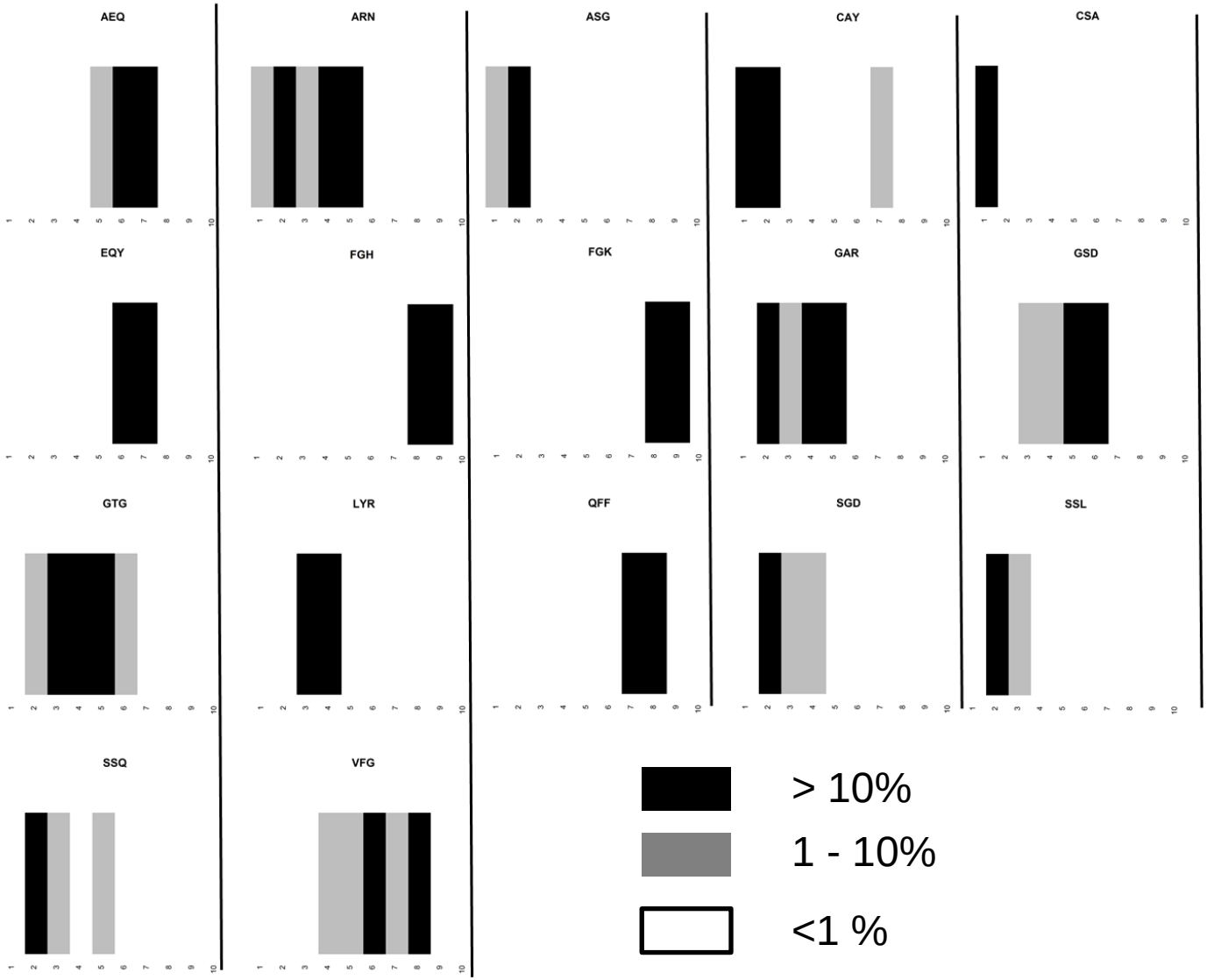
Features selected by LPBoost using counts, Fisher score weighted counts (bold) or both algorithms (red).

AEQ	CSA	EQY	**FGH**	**GAR**	**LYR**	QFF	SGD		**VFG**
ARN	**CAY**		**FGK**	GTG			**SSL**		
ASG				**GSD**			**SSQ**		

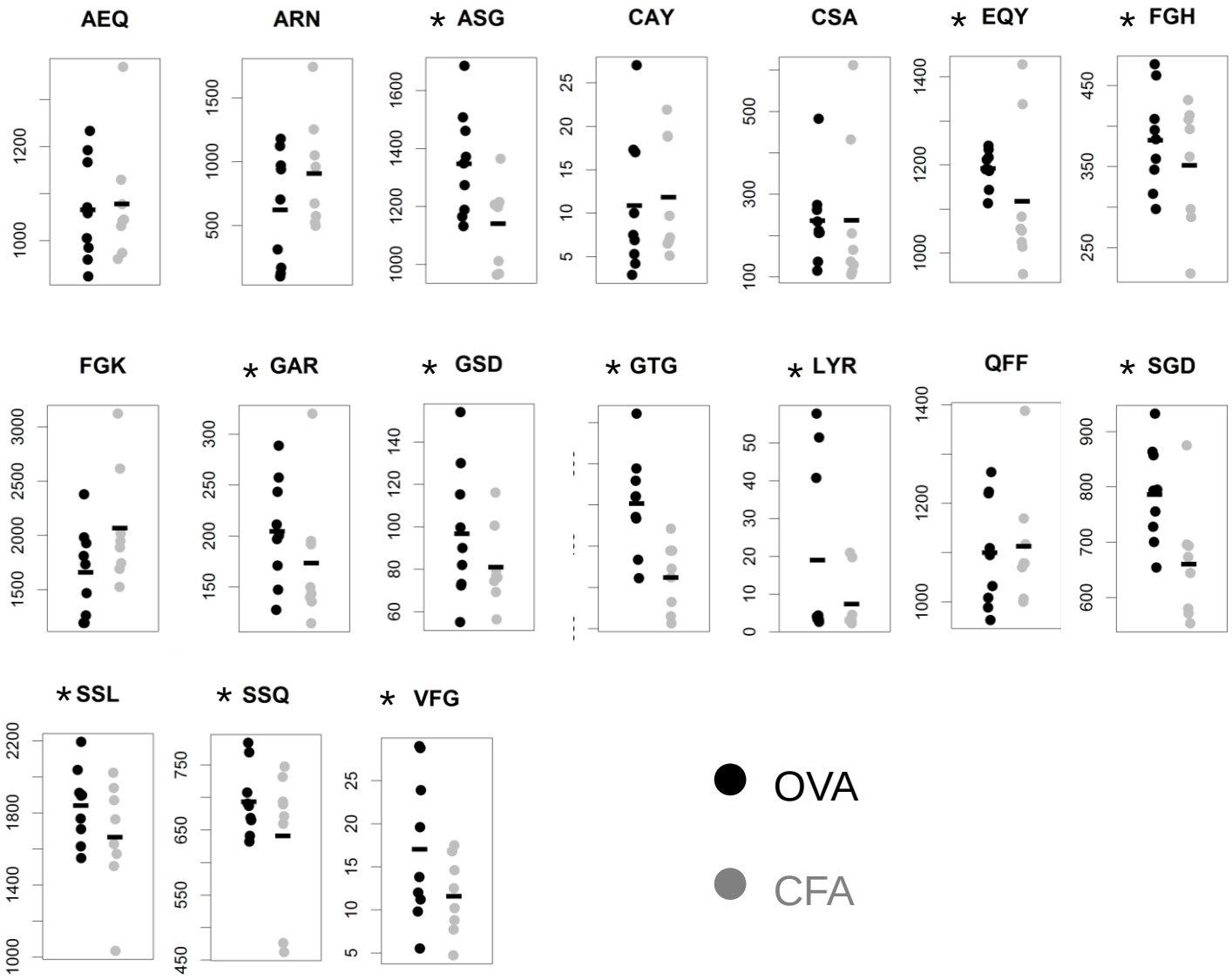


ASS	**CAW**		**FGE**	GQL	LYF	NSD	**SSD**	TGQ	YFG
	CAC			GSG		**NTG**	**SSR**	TQY	YTF
	CAE			**GQS**			**SSY**		
	CAL						SYA		
	CAP						**SYT**		
	CAR						**SSD**		
	CAT						**SSI**		
							SST		

Position of OVA selected triplets along CDR3.



Frequency of OVA selected triplets in early immunised repertoires.

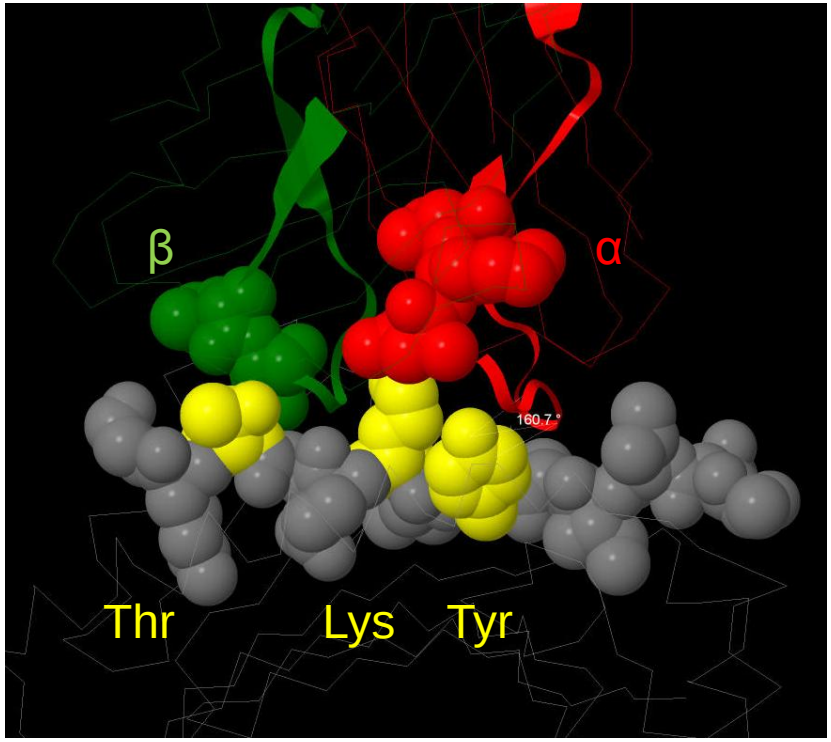


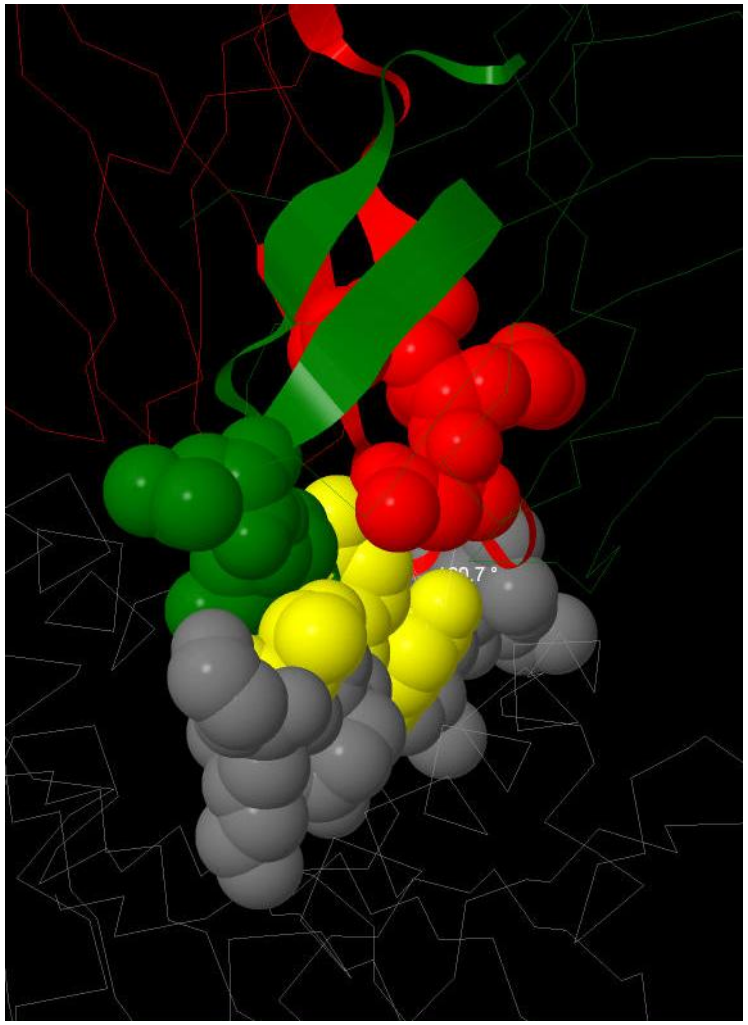
Conclusions

- The OVA response is diverse and predominantly private at the level of CDR3β
- OVA expanded CDR3βs have some sequence similarity
- Amino acid triplets provide features which in combination contribute to defining an OVA response
- Each triplet has a well-defined position along the CDR3
- Many selected triplets are found at the ends of the CDR3, within the sequence coded by V or J region genomic

Speculations (i.e. my future research program)

- The antigen response is distributed across many TCRs of differing frequency
- Antigen specificity is an emerging property of a repertoire and is not defined by individual TCR/pMHC interactions
- Individual TCR/pMHC interactions can be modelled as a set of interactions between individual antigen amino acids and a set of conserved short amino acids motifs within the CDRs





2B4 TCR binding cytochrome C peptide showing conserved residues