Codon arrangement modulates MHC-I peptides presentation

Tariq Daouda[1,2,7*], Maude Dumont-Lagacé[1,3,7], Albert Feghaly[1], Yahya Benslimane[1,3], Rébecca Panes[1,4], Mathieu Courcelles[1,5], Mohamed Benhammadi[1,3], Lea Harrington[1,3], Pierre Thibault[1,5], François Major[1,6], Yoshua Bengio[6], Étienne Gagnon[1,4], Sébastien Lemieux[1,2,6], Claude Perreault[1,3]

[1]Institute for Research in Immunology and Cancer; Université de Montréal; Montréal, Québec H3C 3J7, Canada

[2]Department of Biochemistry; Université de Montréal; Montréal, Québec H3C 3J7, Canada

[3]Department of Medicine; Université de Montréal; Montréal, Québec H3C 3J7, Canada

[4]Department of Microbiology, Infectiology and Immunology; Université de Montréal; Montréal, Québec H3C 3J7, Canada

[5]Department of Chemistry; Université de Montréal; Montréal, Québec H3C 3J7, Canada

[6]Department of Informatics and Operational Research; Université de Montréal; Montréal, Québec H3C 3J7, Canada

[7]These authors contributed equally.

Running title: Co-translational regulation of MHC-I peptide presentation

Tariq Daouda is now affiliated to: (1) Broad Institute of MIT and Harvard, Cambridge, United States; (2) Center for Cancer Research, Massachusetts General Hospital, Charlestown, United States; (3) Department of Medicine, Harvard Medical School, Boston, United States; and (4) Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital, Charlestown, United States.

*Corresponding author: Claude Perreault (claude.perreault@umontreal.ca).

Abstract

MHC-I associated peptides (MAPs) play a central role in the elimination of virus-infected and neoplastic cells by CD8 T cells. However, accurately predicting the MAP repertoire remains difficult, because only a fraction of the transcriptome generates MAPs. In this study, we investigated whether codon arrangement (usage and placement) regulates MAP biogenesis. We developed an artificial neural network called Codon Arrangement MAP Predictor (CAMAP), predicting MAP presentation solely from mRNA sequences flanking the MAP coding regions, while excluding the MAP-coding codons *per se*. CAMAP predictions were significantly more accurate when using codon sequences than amino acid sequences. Furthermore, predictions were independent of mRNA expression and MAP binding affinity to MHC-I molecules, and applied to several cell types and species. Combining MAP binding affinity, transcript expression level and CAMAP scores was particularly useful to ameliorate predictions of MAP derived from lowly expressed transcripts. Using an *in vitro* assay, we showed that varying the synonymous codons in the regions flanking MAP sequences (without changing the amino acid sequence) resulted in significant modulation of MAP presentation at the cell surface. Taken together, our results demonstrate the role of codon arrangement in the regulation of MAP presentation and support integration of both translational and post-translational events in predictive algorithms to ameliorate modeling of the immunopeptidome.

Abbreviations: MHC-I: major histocompatibility complex class-I, MAP: MHC-I associated peptides, CAMAP: Codon arrangement MAP predictor, DRiP: defective ribosomal product, ANN: artificial neural network, MCC: MAP-coding codons, B-LCL: B-lymphoblastoid cell line, KL: Kullback-Leibler, BS: binding score, OVA: ovalbumin protein, WT: wildtype, EP: enhanced presentation, RP: reduced presentation.

Introduction

In jawed vertebrates, all nucleated cells present at their surface major histocompatibility complex class-I (MHC-I) associated peptides (MAPs), collectively referred to as the immunopeptidome (1, 2). MAPs play a central role in shaping the adaptive immune system, as they orchestrate the development, survival and activation of CD8 T cells (3). Moreover, recognition of abnormal MAPs is essential to the elimination of virus-infected and neoplastic cells (4). Therefore, systems-level understanding of MAP biogenesis and molecular composition remains a central issue in immunobiology (5, 6).

The generation of the immunopeptidome can be conceptualized in two main events: (a) the generation of MAP candidates (i.e. peptides of appropriate length for MHC-I presentation) through protein degradation, and (b) a subsequent filtering step through the binding of MAP candidates to the available MHC-I molecules. Rules that regulate the second event have been well characterized using artificial neural networks (ANN) and weighted matrix approaches (7, 8). However, accurately predicting which peptides will ultimately reach MHC-I molecules following a multistep processing in the cytosol and endoplasmic reticulum remains an open question (6). Most efforts at modeling MAP generation have focused on post-translational events and their regulation by the amino acid sequence of MAPs and of directly adjacent residues (typically 10-mers at the N- and C-termini). While the consideration of preferential sites of proteasome cleavage has proven useful to enrich for MAP candidates (9), it remains insufficient for MAP prediction, due to prohibitive false discovery rates (10–12). A large body of evidence suggests that a substantial portion of MAPs are produced co-translationally (13–15), deriving from defective ribosomal products (DRiPs), that is, polypeptides that fail to achieve a stable conformation during translation and are consequently rapidly degraded. This concept was initially supported by two observations: (i) viral MAPs can be detected within minutes after viral infection, much earlier than their associated proteins half-life (16), and (ii) MAP presentation correlates more closely with translation rate than with overall protein abundance (17, 18). In addition, while all proteins contain peptides that are predicted to bind MHC-I molecules, mass spectrometry analyses have revealed that the immunopeptidome is not a random excerpt of the transcriptome or the proteome (1, 19). Indeed, proteogenomic analyses of 25,270 MAPs isolated from B lymphocytes of 18 individuals showed that 41% of expressed protein-coding genes generated no MAPs (19). These authors also provided compelling evidence that the presentation of MAPs cannot be explained solely by their affinity to MHC-I alleles and their transcript expression levels, while ruling out low mass spectrometry sensitivity as an explanation for the non-presentation of the strong binders. Because (i) MAPs appear to preferentially derive from DRiPs and (ii) codon usage influences both precision and efficiency of protein synthesis (20, 21), we hypothesized that codon usage in the vicinity of MAP-coding codons (MCCs) might significantly contribute to the regulation of MAP biogenesis. We developed an artificial neural network called Codon Arrangement MAP Predictor (CAMAP), trained to identify MCC-flanking regions. We then used CAMAP to uncover key codon features that characterize mRNA sequences encoding for MAPs (i.e. source) when compared to sequences that do not (i.e. non-source).

Experimental Procedures

Experimental design and statistical rationale

The fact that only a specific part of the genome generates MAPs suggests that the generation of the immunopeptidome can be conceptualized in two main events: (a) the biogenesis (or preselection) of MIPs candidates (i.e. peptides of appropriate length for MHC-I presentation through protein degradation), and (b) a subsequent filtering step through the binding of MIP candidates to the available MHC-I molecules. Rules that regulate the second event, i.e. the binding of MIPs to MHC-I molecules, have been well characterized by artificial neural networks (ANN) and weighted matrix approach (7, 8). However, it is currently impossible to predict the first event; that is, which peptides will become MAP candidates and ultimately reach MHC-I molecules following a multistep processing in the cytosol and endoplasmic reticulum. The main objective underlying our study was to uncover key features that characterized transcripts encoding for MAPs (i.e. source transcripts) or not (i.e. non-source transcripts). Since MAPs appear to preferentially derive from DRiPs and codon usage influences both accuracy and efficiency of protein synthesis, we hypothesized that codon usage might contribute to the regulation of MAP biogenesis.

Dataset generation

We elected to analyze a previously published dataset consisting of MAPs presented on B lymphocytes by a total of 33 MHC-I alleles from 18 subjects (19, 22). Since this dataset was assembled using older versions of MHC-I binding prediction algorithms (i.e. using a combination of NetMHC3.4 for common alleles and NetMHCcons1.1 for rare alleles), we verified that the majority of MAPs in this dataset would also be predicted as binders using more recent algorithms (i.e. a rank ≤ 2.0% using NetMHC4.0 or NetMHCpan4.0). We found an overlap of >92% between these methods (see Supplementary Fig. 1), thereby validating this dataset for further analysis. In addition, we reasoned that a transcript should be considered as a genuine positive or negative regarding MAP biogenesis only if it was expressed in the cells. We therefore excluded from the dataset all transcripts with very low expression (<1st percentile in terms of FPKM).

To facilitate data analysis and interpretation, we only included transcripts coding for MAPs with a length of 9 amino acids, for a total of 19,656 9-mer MAPs (which represents 78% of MAPs described in Pearson et al., 2016). We then used pyGeno (23) to extract the mRNA sequences of transcripts coding for these 9-mer MAPs, which constituted our source-transcripts (Fig. 1A). We next created a negative (non-source) dataset from transcripts that generated no MAPs. Importantly, transcripts that encoded for MAPs of any length (i.e. 8 to 11-mer) were excluded from the negative dataset. We then randomly selected 98,290 non-MAP 9-mers from this negative dataset, and extracted their coding sequences using pyGeno. Of note, both positive and negative datasets were derived from the canonical reading frame of non-redundant transcripts.

We analyzed only the MAP context only and excluded the MCCs *per se* from our positive (hits) and negative (decoys) sequences (Fig. 1A). We limited our analyses of flanking sequences to 162 nucleotides (54 codons) on each side of MCCs, because longer lengths would entail the exclusion of >25% of transcripts (Supplementary Fig. 2).

Creation of the shuffled synonymous codon dataset

To create the shuffled synonymous codon dataset, each sequence was re-encoded by replacing each codon with itself or with a random synonym according to the human transcriptome usage frequencies extracted using *pyGeno*. These frequencies were calculated in silico on transcript coding sequences using the annotations provided by *Ensembl* for the human reference genome GRCh37.75. Thus, all codon-specific features differing between the positive and negative datasets was removed from the shuffled datasets. Because codons were replaced by their synonymous codons, the shuffled sequences directly reflected amino acid usage in the positive and negative datasets.

CAMAP architecture, sequence encoding and training

The first (input) layer received either MCC-flanking regions from the positive dataset or sequences of the same length contained in the negative dataset (Fig. 1A). The second layer (Supplementary Fig. 3A) was a codon embedding layer similar to that introduced for a neural language model (24). Embedding is a technique used in natural language processing to encode discrete words, and has been shown to greatly improve performances (25). With this technique, the user defines a fixed number of dimensions in which words should be encoded. When the training starts, each word receives a random vector-valued position (its embedding coordinates) in that space. The network then iteratively adjusts the words' embedding vectors during the training phase and arranges them in a way that optimizes the classification task. Notably, embeddings have been shown to represent semantic spaces in which words of similar meanings are arranged close to each other (25). In the present work, we treated codons as words: each codon received a set of random 2D coordinates that were subsequently optimized during training. The third (output) layer delivered the probability that the input sequence was an MCC-flanking region (rather than a sequence from the negative dataset).

CAMAPs were trained on sequences resulting from the concatenation of pre- and post-MCC regions. Before presenting sequences to our CAMAPs, we associated each codon to a unique number ranging from 1 to 64 (we reserved 0 to indicate a null value) and used this encoding to transform every sequence into a vector of integers representing codons. Neural networks were built using the Python package Mariana (26) [https://www.github.com/tariqdaouda/Mariana]. The *Embedding* layer of Mariana was used to associate each label superior to 0 to a set of 2D trainable parameters; the 0 label represents a *null* (masking) embedding fixed at coordinates (0,0). As an output layer, we used a *Softmax* layer with two outputs (positive / negative). Because negative sequences are more numerous than positive ones, we used an oversampling strategy during training. At each epoch, CAMAPs were randomly presented with the same number of positive and negative sequences. All CAMAPs in this work share the same architecture (Supplementary Fig. 3A), number of parameters and hyper-parameter values: learning rate: 0.001; mini-batch size: 64; embedding dimensions: 2; linear output without offset on the embedding layer; *Softmax* non-linearity without offset on the output layer.

For each condition (e.g. context size), the positive and negative datasets were randomly divided into three non-redundant subsets: (i) the training subsets containing 60% of the positive and negative transcripts, (ii) the validation and (iii) the test subsets each containing 20% of the positive and negative transcripts. Transcripts were assigned through a sequence redundancy removal algorithm, thereby ensuring that no transcript was assigned to multiple subsets. We used an early stopping strategy on validation sets to prevent over-fitting and reported average performances computed on test sets. We trained 12 CAMAPs for each combination of conditions, each one using a different random split of train/validation/test sets. To mask sequences either before or after the MCC, we masked either half with *null* value.

Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence computes how well a given distribution is approximated by another distribution. Its value can be either positive or 0, a null value indicating that the two distributions are identical (see Experimental Procedures for more details). Accordingly, a higher KL divergence for codon distributions vs. amino acid distributions would indicate that codon variations are not entirely accounted for by amino acid variations. KL divergence is not a metric, as it is neither symmetric nor does it satisfy the triangle inequality. It is nevertheless an accurate and most common way of comparing two probability distributions.

We defined the probability of having codon *c* at position *i* as a function of the number of occurrences of *c* at position *i*, divided by the total number of occurrences of that same codon:

$$\boldsymbol{Q_{(c,y,s)}(i) = \frac{N_{c,y,s}(i)}{\sum_j N_{c,y,s}\,(j)}}$$

Here Q is a probability, N is a number of occurrences, c is a codon, y is a class (positive or negative), s indicates if codons have been randomized (true or false), i is a position in sequence. For the remainder of the text we will use the following abbreviations:

$$\boldsymbol{P_c(i) = \ Q_{c,y=positive,s=false}(i)}$$

$$\boldsymbol{D_c(i) = \ Q_{c,y=negative,s=false}(i)}$$

$$\boldsymbol{PS_c(i) = \ Q_{c,y=positive,s=true}(i)}$$

$$\boldsymbol{DS_c(i) = \ Q_{c,y=negative,s=true}(i)}$$

We then used the KL divergence to compute how well P_c distributions approximate D_c distributions and PS_c distributions approximate DS_c distributions.

The KL divergence was defined as:

$$\boldsymbol{D_{KL}(P||Q) = \sum_i P(i)\mathbf{log}\,(\frac{P(i)}{Q(i)})}$$

We performed this calculation for both the original and the shuffled dataset, which we then compared together. If codons and amino acid distributions were equivalent, KL divergence between hits and decoys would be the same for both original and shuffled sequences, and codons would cluster along the diagonal.

Interrogation of CAMAP score to extract codon preferences

Artificial neutral networks (ANNs) still carry the reputation of being undecipherable black boxes. It is true that the interpretation of the inner structures of deep ANNs is still in its infancy. On the other hand, simpler architectures, such as the one used herein, can be more easily probed to yield useful information about the way predictions are being made. Indeed, a trained ANN remains a fixed set of mathematical transformations that can be studied, analyzed and, in theory, interpreted.

We wondered whether some regions were more influential on MAP presentation than others. To address this question, we retrieved the model preferences for each codon at each position, i.e. the prediction score of our best model (trained with original codon sequences for a context size of 162 nucleotides) when a single codon at a single position is provided as input (all other positions being set at [0,0] coordinates in the embedding space). The model's preferences are therefore a measure of each individual codon's propensity to increase or decrease the model's output probability as a function of its position relative to the MCCs. A value of 0.5 denotes a neutral preference, while negative and positive preferences correspond to values below and above 0.5, respectively. Preferences were obtained by feeding the ANN embedding vectors where all codons values were set to null (coordinates (0,0)), except for a single position that received a non-null codon label.

In vitro assay – inducible translation reporter (iTR)-OVA construct design

An inducible translation reporter was generated by flanking the truncated chicken ovalbumin (OVA) cDNA (amino acids 144-386) with EGFP-P2A (in 5') and P2A-Ametrine (in 3') cDNA sequences. MCC-flanking contexts for the EP and RP construct were synthesized as gBlocks (purchased from Integrated DNA Technologies). The fragments were amplified by PCR and joined by Gibson assembly under a doxycycline-inducible Tet-ON promoter in a pCW backbone. Synthetic variants of the OVA coding sequence were generated in silico by varying synonymous codon usage in the MAP context regions (i.e. 162 nucleotides pre- and post-MCC). Importantly, the amino acid sequence was preserved between the different variants; only nucleotide sequences in the MAP context (162 nucleotides on either side) differed. The sequences with the highest (EP) and the lowest (RP) prediction scores were selected for further in vitro validation and swapped into the iTR-OVA plasmid by Gibson assembly (27). OVA-EP and OVA-RP sequences can be found in Supplementary Table 1.

Important features of our inducible translation reporter construct and T cell activation assay were: (i) No changes in amino acid sequence between the three variants: only co-translational events can differ between the three variants, post-translational events being equivalent for the three constructs; (ii) Only one start codon, at the beginning of the eGFP coding sequence: this is important for the translation reporter aspect of our construct (i.e. Ametrine/eGFP ratio), to ensure that translation can only start at the 5'-end of the whole construct, and not at the beginning of the OVA or Ametrine coding sequences; (iii) Separation of the three proteins using P2A peptide: allows the inducible synthesis of three separate proteins in a highly correlated manner; also, the degradation of one protein will be independent from the others. As we hypothesized that codon usage might lead to DRiP formation, we did not want the degradation of OVA-derived polypeptide to induce degradation of attached eGFP or Ametrine, which would affect our translation reporter assay (Ametrine/eGFP ratio); (iv) Because transcript expression level impacts MAP presentation, we normalized T-cell activation results by both the number of transduced cells present in the samples (% of eGFP+ cells) and the Ametrine mean fluorescence intensity of eGFP+ cells (representing whole construct expression level). Because of these four features, any difference between the three constructs could be ascribed solely to synonymous codon variants in the SIINFEKL-flanking OVA codons.

Stable cell line generation

Wildtype and transduced Raw-K^b cells (28) were cultured in DMEM supplemented with 10% Fetal Bovine Serum (FBS), penicillin (100 units/ml), and streptomycin (100mg/ml). B3Z cells (29) were maintained in RPMI medium supplemented with 5% FBS, penicillin (100 units/ml), and streptomycin (100mg/ml).

Lentiviral particles were produced from HEK293T cells by co-transfection of iTR-OVA WT, EP or RP along with pMD2-VSVG, pMDLg/pRRE and pRSV-REV plasmids. Viral supernatants were used for Raw-K^b transduction. Raw-K^b OVA-WT, Raw-K^b OVA-EP were sorted on Ametrine and GFP double positive population after 24h of doxycycline treatment (1 mg/ml).

T-cell activation assay

Raw-K^b OVA-EP, OVA-RP and OVA-WT cells were plated at a density of 250,000 cells/well in 24 well-plates 24h prior to doxycycline treatment (1 mg/ml). After the corresponding treatment duration, cells were harvested and fixed using PFA 1% for 10 minutes at room temperature and washed using DMEM 10% FBS. Raw-K^b were then co-cultured (37°C, 5% CO_2) in triplicates with the CD8 T cell hybridoma cell line B3Z cells at a 3:2 ratio for 16h (7.5 x 10^5 B3Z and 5 x 10^5 Raw-K^b) in 96 well-plates. Cells were lysed for 20 minutes at room temperature using 50 µl/well of lysis solution (25mM Tris-Base, 0.2 mM CDTA, 10% glycerol, 0.5% Triton X-100, 0.3mM DTT; pH 7.8). 170 µl/well CPRG buffer was added (0.15mM chlorophenol red-β-d-galactopyranoside (Roche), 50mM Na_2HPO_4•$7H_20$, 35mM NaH_2PO_4•H_20, 9mM KCl, 0.9mM $MgSO_4$•$7H_2O$). β-galactosidase activity was measured at 575 nm using SpectraMax® 190 Microplate Reader (Molecular Devices). In parallel, cells were analyzed by flow cytometry using a BD FACS CantoII for eGFP and Ametrine fluorescence.

Results

Dataset description

We analyzed a previously published dataset consisting of MAPs presented on B lymphoblastoid cell line (B-LCL) by a total of 33 MHC-I alleles from 18 subjects (19, 22). Because we were searching for features that influence MAP generation and not the binding of MAP to MHC, we elected to analyze the MAP context only and excluded the MCCs *per se* from our positive (hits) and negative (decoys) sequences (Fig. 1A). To facilitate data analysis and interpretation, we restricted our hit dataset to MAPs with a length of 9 amino acids, for a total of 19,656 9-mer MAPs (which represents 78% of MAPs described in Pearson et al., 2016). We next created a decoy dataset from transcripts that generated no MAPs, by randomly selecting 98,290 9-mers from these transcripts. Finally, we used pyGeno (23) to extract MCC-flanking regions corresponding to both hit and decoy MAPs, which constituted our final dataset for CAMAP. Of note, each sequence in the final dataset is unique and derives from the canonical reading frame. In addition, in order to investigate the relative importance of codon vs. amino acid usage in MAP biogenesis, we generated a dataset of shuffled sequences in which original codon sequences were randomly replaced by synonymous codons according to their usage frequency in the dataset (Fig. 1B). The random shuffling causes any codon-specific feature to be shared among synonyms, thereby causing the shuffled codon distribution to reflect the amino acid usage (see Experimental Procedures for more details).

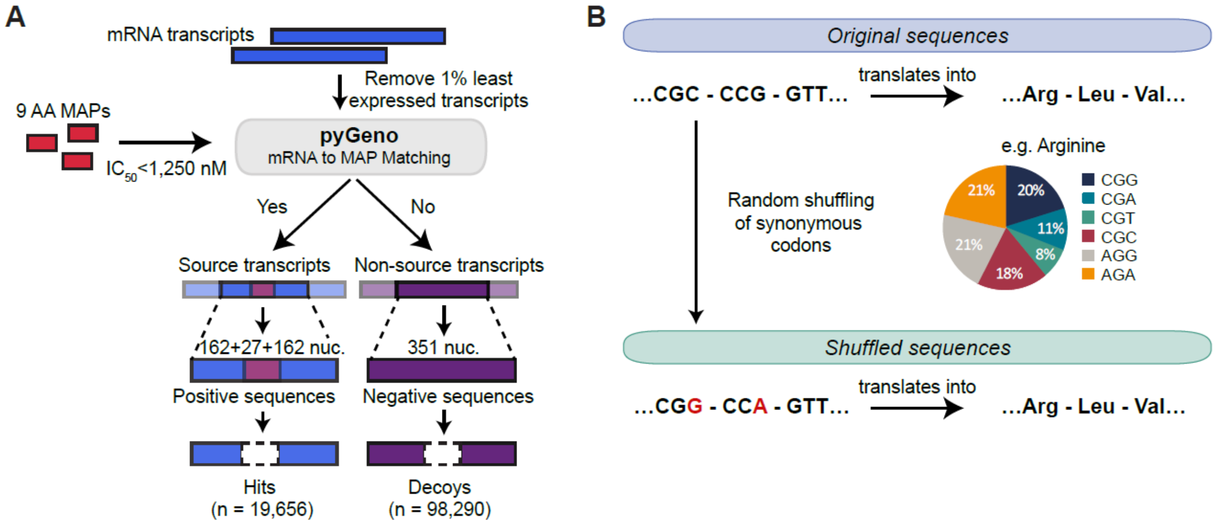


Figure 1. Construction of the dataset. (a) Transcripts expressed in B cells from 18 subjects were considered as source or non-source transcripts depending on their match with at least one MAP. Because we were searching for features that might influence MAP generation and not the binding of MAP to MHC, we focused our attention on mRNA sequences more closely adjacent to the nine MAP-coding codons (MCCs), i.e. up to 162 nucleotides on each side of MCCs. (b) Creation of the shuffled dataset. Codons were randomly replaced by a synonymous codon according to their respective frequencies (i.e. codon usage) in the dataset. The random shuffling causes any codon-specific feature to be shared among synonyms, thereby causing the shuffled codon distribution to reflect the amino acid usage. Importantly, both the original sequence and its shuffled version translates into the same amino acids.

CAMAP links codon usage to MAP presentation

To assess the importance of codon usage in MAP biogenesis, we reasoned that if codons bear important information that is operative at the translational rather than the post-translational level, then: (i) CAMAP trained to identify MCC-flanking regions should consistently perform better when trained on mRNA sequences than on amino acid sequences, and (ii) synonymous codons should have different effects on the prediction. To test these hypotheses, CAMAP received as inputs MCC-flanking regions from hit and decoy sequences from either the original or shuffled datasets. It was then trained to predict the probability that individual input sequences were MCC-flanking regions (i.e. hit) rather than sequences from the negative dataset (Supplementary Figure S3A).

We compared CAMAP performance when predicting MAP presentation from original sequences, representing codon arrangement, versus shuffled sequences representing amino acid arrangement. To evaluate the robustness of our approach, 12 different CAMAPs were trained in parallel, with different train-validation-test splits of the dataset. Our results show that predictions were consistently better when CAMAP received the original codons rather than the shuffled synonymous codons sequences (Fig. 2A and Supplementary Figure S3B). CAMAPs receiving information from both pre-MCC and post-MCC sequences (i.e. whole MCC context) also performed better than when receiving only pre- or post-MCC context (Fig. 2A and Supplementary Figure S3C-D), suggesting that pre- and post-MCC context were not redundant. Indeed, we found a weak correlation between the prediction scores of CAMAPs trained only with pre-MCC or post-MCC sequences (Supplementary Fig. S4). In addition, CAMAPs receiving longer sequences performed better than those receiving shorter sequences (Fig. 2B). Because sequences located far upstream and downstream of the MCC (i.e. in ranges exceeding the direct influence of proteases) are informative regarding MAP presentation, it suggests the existence of factors unrelated to protein degradation modulating MAP presentation.

We next tested our CAMAP trained on 9-mer MAPs derived from B-LCL on 5 datasets containing MAPs identified by proteogenomic analyses of different human and mouse cell types. These included 2 human datasets derived from primary cells (our B-LCL dataset, this time including all peptide lengths (19, 22), and a dataset of peripheral blood mononucleated cells or PBMCs (30)), one human (B721.221 (11)) and 2 murine cell lines (colon carcinoma CT26 and lymphoma EL4 (30, 31)). For all datasets, we created hit and decoy datasets of original and shuffled sequences

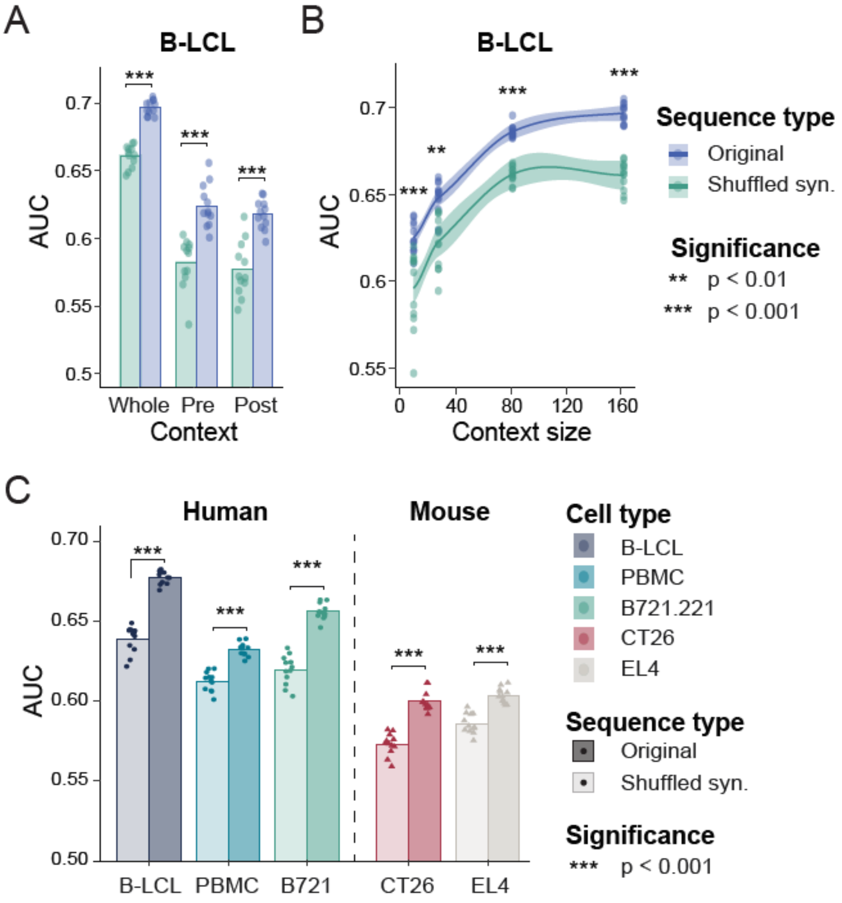


Figure 2. CAMAP predictions on MAP-flanking sequences. (A) Area under the curve (AUC) score for CAMAPs trained with whole MCC-context, versus CAMAPs trained with only pre- or post-MCC context. All CAMAPs presented here were trained with a context size of 162 nucleotides. (B) AUC for CAMAPs trained with context sizes of 9, 27, 81 and 162 nucleotides. (C) CAMAP prediction score for different datasets derived from humans (i.e. B-LCL, PBMCs and B721.221) or mouse (i.e. CT26 and EL4 cell lines). Of note, all CAMAPs were trained on B-LCL-derived sequences encoding for 9-mer MAPs only with a context size of 162 nucleotides. Results are reported for 8 to 11-mer MAPs derived from the 5 datasets. In all panels, 12 CAMAPs trained with original or shuffled synonymous sequences were compared (significance assessed using Student T test).

using the same approach described above, but included MAPs of 8-11 amino acids. Notably, CAMAPs trained on human sequences encoding 9-mers MAPs from one human cell type (i.e. B-LCL) could also predict presentation of 8-11 mers MAPs in other human cell types (Fig. 2C), as well as from mouse cell lines, albeit with lower performances (Fig. 2C). Here again, CAMAPs trained on original sequences consistently outperformed CAMAPs trained on shuffled sequences (Fig. 2C). These results show that the rules derived by CAMAPs to predict MAP presentation are valid across different cell types, and can be applied to different species. These results also support a role for codons in the modulation of MAP presentation, possibly through modulation of the probability of DRiP generation. Of note, CAMAP prediction scores did not correlate with MAP/MHC-binding affinity or transcript expression levels, suggesting that the rules derived by CAMAP are independent of these other known factors regulating MAP presentation (Fig. 3).

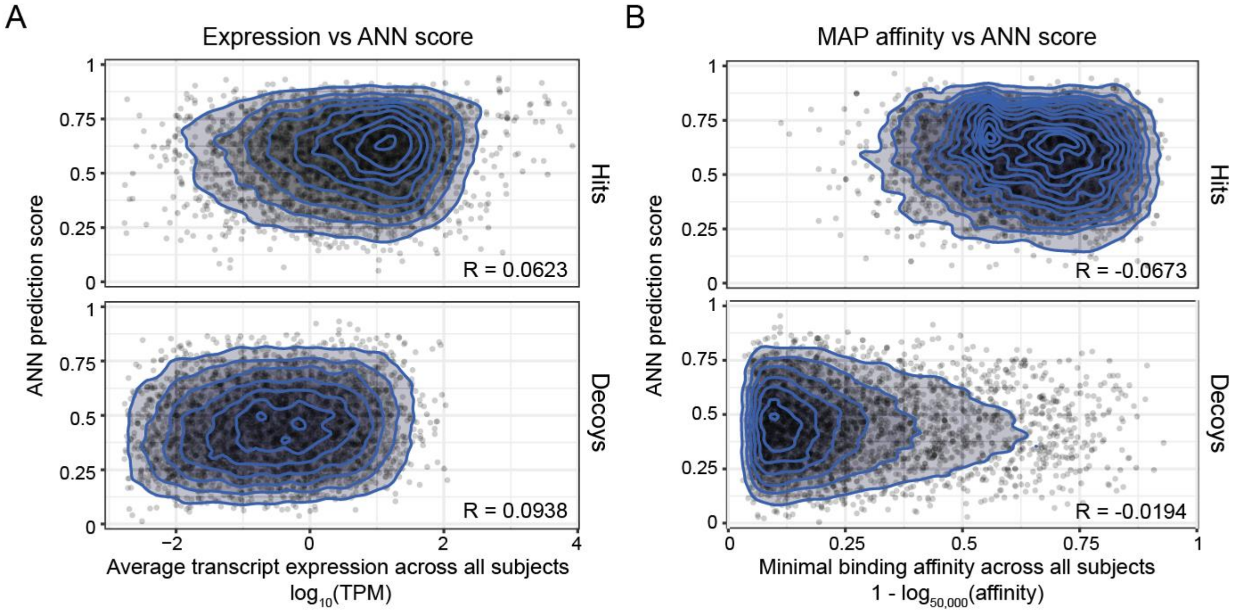


Figure 3. Correlation between CAMAP prediction score and (A) transcript expression levels and (B) MAP binding affinity. CAMAP used here was trained on original codon sequences using a context size of 162 nucleotides (both pre- and post-MCC context). Densities were calculated on all points and drawn using ggplot2. Only a random subset of the points is represented in the figures to limit their size.

The lower performances of CAMAP trained with shuffled sequences (representing amino acid distribution) suggests that amino acids in MAP-flanking sequences are less informative than codons regarding MAP presentation. We formally quantified this difference in information using the Kullback-Leibler (KL) divergence (see Experimental Procedures for more details). Most codons (47/61, 77%) showed greater KL divergence in the original dataset than the shuffled dataset, indicating that codon distributions contained more information with regards to MAP presentation than amino acid distributions (Supplementary Figure S5). These results suggest that codons in MAP-flanking regions play a role that is non-redundant with amino acids in MAP biogenesis.

Interestingly, while codons closest to the MCC were the most influential on CAMAP scores, some synonymous codons showed opposite effects, demonstrating that codon usage does not recapitulate amino acid usage (Fig. 4A-B and Supplementary Figures S6). The use of embeddings to encode codons has the advantage of arranging them into a semantic space, wherein codons with similar influences are positioned close to each other. We calculated the resulting semantic space as well as the preferences for every codon for the position directly preceding the MCCs (Fig. 4C). Most synonymous codons did not form clusters, with a notable exception being proline codons. This finding indicated that the effect of a given codon on the prediction may be closer to that of a non-synonymous codon than to that of a synonym.

CAMAP increases MAP prediction accuracy

We next compared MAP prediction capacities of CAMAPs scores to that of MAP binding score (BS) and mRNA transcript expression levels. Because MAP binding to the MHC molecule is essential for its presentation at the cell surface, we elected to only compare hits and decoys encoding potential binders, i.e. with a minimal BS <1,250 nM (approximately corresponding to <2% rank) for ≥1 allele in the B-LCL dataset. Using a linear regression model, we compared the

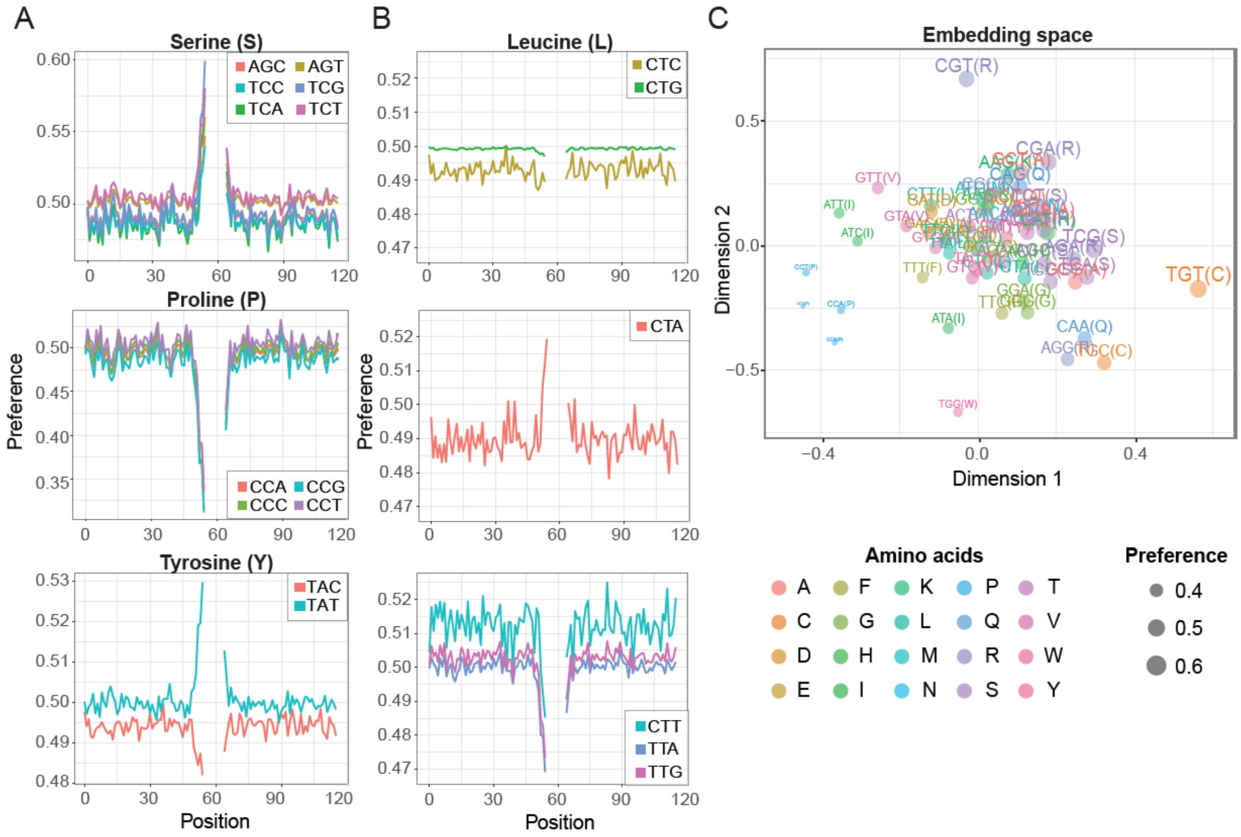


Figure 4. CAMAP interpretation of codon impact on MAP biogenesis. Preferences for a network trained on a context of 162 nucleotides (54 codons) for (A) serine, proline and tyrosine codons, and (B) leucine codons. (C) Learned codon embeddings and preferences at the position directly preceding the MCCs. Proline codons were the only synonyms that formed a conspicuous cluster. As indicated by the size of the dots, codons on the right-hand side increased the probability of the sequence being classified as source, whereas codons on the left-hand side of the graph had the opposite effect. See Experimental Procedures for more details.

predictive capacity of each parameter using Matthews correlation coefficient, which measures the quality of binary classifications. In line with previous studies (11, 19), the mRNA expression level had the highest predictive capacity, followed by BS and CAMAP score (Fig. 5A). Combining CAMAP score with either BS alone, or expression level alone, or both resulted in significant increases in predictive performances (Fig. 5B). Interestingly, CAMAP score was most helpful to predict MAP presentation for transcripts with low expression levels (Fig. 5B), contributing almost as much as the BS to the prediction (Fig. 5C). These results show that combining the CAMAP score with the MAP's BS and corresponding transcript expression level improves prediction of MAP, especially for transcripts with low expression.

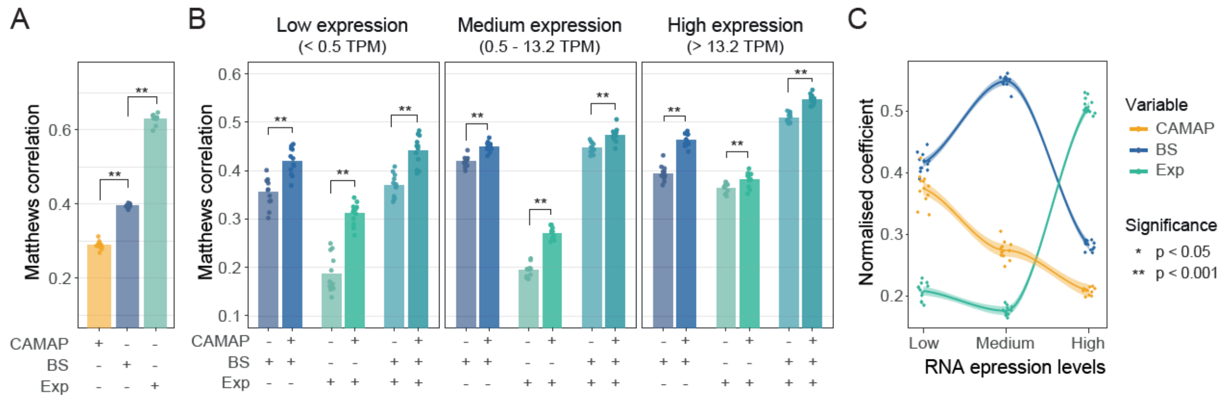


Figure 5. ANN prediction score contributes to the prediction of MIPs especially in transcripts with low expression. (A) Matthews correlation coefficient for MAP prediction using a single variable. (B) Matthews correlation coefficient for MAP prediction using multivariable regression models. The B-LCL dataset (all MAP lengths) was filtered for MAP with a minimal binding affinity ≤1,250 nM and further separated into tiers based on transcripts expression level (low, intermediate and high expression). (C) Contribution of each variable to MAP prediction in a three-variable model for each tier.

Codon usage modulates MAP presentation

To formally demonstrate the influence of codon arrangement on MAP presentation, we generated three variants of the chicken ovalbumin (OVA) protein, containing the model MAP SIINFEKL (32). All variants encoded the same amino acid sequence but used different synonymous codons. Notably, the sole difference between the three constructs were the 162 nucleotides flanking each side of the SIINFEKL-coding codons (i.e. the RNA sequences coding for $\mathrm{OVA}_{202\text{-}256}$ and $\mathrm{OVA}_{265\text{-}319}$, Supplementary Table S1). One construct encoded the wild type OVA (OVA-WT). For the other two constructs, we used CAMAP (trained on original human B-LCL sequences; Fig. 2) to generate two OVA variants *in silico*, both encoding for the same OVA protein but using different synonymous codons: one predicted to enhance SIINFEKL presentation (OVA-EP), the other predicted to reduce it (OVA-RP). Accordingly, the respective CAMAP scores for OVA-RP, OVA-WT and OVA-EP were: 0.03, 0.65, and 0.96 (Fig. 6A).

Because codon usage affects translation efficiency, theoretically leading to DRiP formation through premature translation arrest (20, 21), we expected the variable regions of our construct to affect both translation rates and SIINFEKL presentation in our variants. Therefore, each construct also coded for two other proteins, eGFP and Ametrine, placed upstream and downstream of the OVA coding sequence, respectively (Fig. 6A). While the Ametrine fluorescence intensity reflected the translation rate of the whole construct, the ratio of Ametrine/eGFP fluorescence intensity was informative regarding the translation efficiency of the whole construct. Indeed, efficient translation of the full-length construct should produce equivalent quantities of Ametrine and eGFP proteins, while inefficient/interrupted translation of the construct (i.e. leading to DRiP formation) should decrease the Ametrine/eGFP ratio (Fig. 6B). The three protein coding sequences were separated with P2A self-cleaving peptides (33), therefore allowing the synthesis of three separate proteins, controlled by the doxycycline-inducible Tet-On promoter. Importantly, the three proteins were tightly co-expressed because of the presence of only one start codon at the 5' end of the GFP protein, as illustrated by the high correlation between eGFP and Ametrine fluorescence (R>0.9). As we assumed that CAMAP scores reflected the probability of DRiP generation leading to increased MAP presentation, we expected the OVA-RP construct to show both reduced SIINFEKL presentation and enhanced translation efficiency compared to the OVA-EP and OVA-WT constructs. However, as both the OVA-EP and OVA-WT have CAMAP scores above the neutral threshold of 0.5, we expected these constructs to behave more similarly.

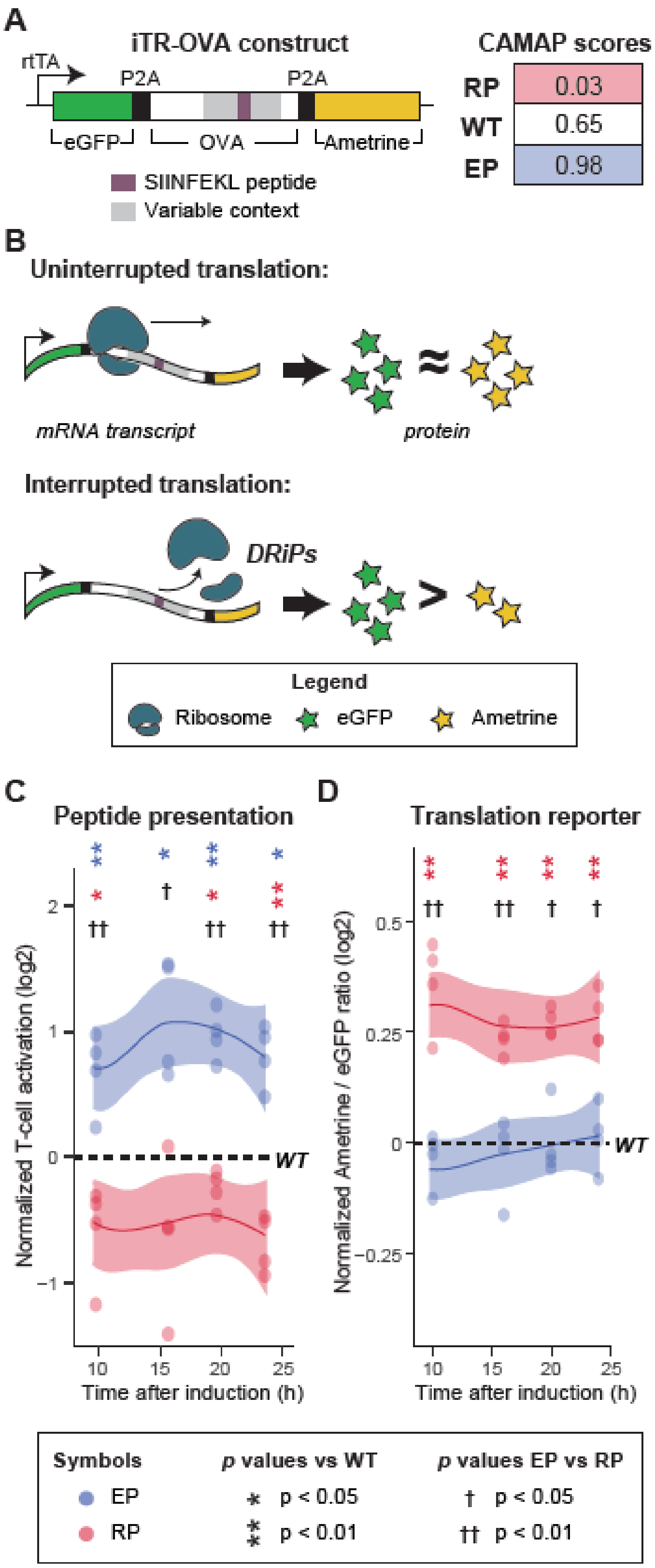


Figure 6. Codon usage influences antigen presentation and translation efficiency. (A) Design of the inducible Translation Reporter (iTR-OVA) constructs and prediction scores for OVA-WT, OVA-EP and OVA-RP sequences. (B) Schematic representation of possible translation events. When mRNA codon usage leads to efficient (uninterrupted) translation, similar amounts of eGFP and Ametrine proteins would be synthesized. When codon usage in the MCC-flanking regions enhances the frequency of translation interruption, a lower Ametrine/eGFP ratio would be observed. (C) Kinetics of SIINFEKL MAP presentation following induction of iTR-OVA constructs expression by doxycycline, measured in a T-cell activation assay. To remove the influence of differential expression levels on antigenic presentation and of varying proportion of transduced cells between samples, T-cell activation levels were normalized to both mean Ametrine fluorescence intensity and proportion of eGFP+ cells (i.e. cells expressing the construct). (D) Translation efficiency as measured by Ametrine/eGFP ratio following iTR-OVA construct induction. For C and D, results are normalized over the WT sample from the same experiment (n=4). Statistical differences at each time point were determined using bilateral paired Student T tests. Comparison against WT are indicated with *, while comparison of EP vs RP is indicated with †.

We then used a SIINFEKL-H2-K^b specific T-cell activation assay (34) to measure SIINFEKL presentation at the cell surface following doxycycline induction. Results for the T-cell activation assay were normalized by both the Ametrine mean fluorescence intensity and the percentage of transduced (eGFP+) cells in each specific sample, so that any difference in T-cell activation observed between our constructs could only be ascribed to synonymous codon variants in the SIINFEKL-flanking OVA codons. Two main findings emerged from our analyses. First, in accordance with CAMAP predictions, variation in codon usage led to a 2.3-fold difference in SIINFEKL presentation between the OVA-EP and OVA-RP variants, with OVA-WT in between (Fig. 6C). Second, translation efficiency (Ametrine/eGFP ratio) was higher with OVA-RP than with OVA-EP or OVA-WT (Fig. 6D). Hence, synonymous codon variations led to slightly divergent outcomes in OVA-EP and OVA-RP: they modulated the levels of SIINFEKL presentation in both OVA-EP and OVA-RP, but enhanced translation efficiency only in OVA-RP. These data show that codon arrangement modulates MAP presentation strength, and strongly support a role for translation efficiency and DRiP formation in the modulation of MAP presentation.

Discussion

Our analyses of large datasets using diverse bioinformatics approaches provides compelling evidence that codon arrangement in mRNA sequences regulates MAP biogenesis. The functional link between codon arrangement and MAP biogenesis was further strengthened by our *in vitro* analyses of SIINFEKL biogenesis. Indeed, we were able to modulate SIINFEKL presentation solely by substituting synonymous codons in mRNA regions flanking SIINFEKL codons, without changing the protein sequence This experiment also highlighted co-translational degradation modulated by codon arrangement as a mechanism regulating differential MAP presentation. Interestingly, rules derived by CAMAP also applied to various human and mouse cell types, suggesting that codon usage is instrumental in MAP biogenesis across different species and cell types.

Further analyses of large datasets will be needed to assess the full extent of codon arrangement's impact on both classic MAPs (i.e. derived from canonical reading frames of coding sequences) and cryptic MAPs (i.e. derived from non-canonical reading frames and non-coding sequences) (35, 36). A more practical implication of our work is the integration of both translational (codon arrangements) and post-translational events (e.g., MHC-binding affinity) in predictive algorithms to enhance the predictive modeling of the immunopeptidome for cancer immunotherapy and peptide-based vaccines, where discovery of suitable target antigens remains a formidable challenge (37, 38).

Acknowledgements

This work was supported by grants from the Canadian Cancer Society (number 705604 and 705714), the Oncopole and the Leukemia & Lymphoma Society of Canada. Perreault' lab is supported in part by The Katelyn Bedard Bone Marrow Association. MDL was supported by a studentship from the Canadian Institute of Health Research. Y.Benslimane was supported by a fellowship from the Cole Foundation and a Canadian Institutes of Health Research operating grant (to LH: #13784). EG lab is supported by the Canadian Institute for Health Research operating grant (MOP-133726). The B3Z CD8+ T cell hybridoma cell line was a kind gift from Nilabh Shastri. RAW-Kb cells were kindly provided by Michel Desjardins.

Data Availability

The datasets analyzed for this study can be found:

- Human B-LCL: RNA-Seq data can be accessed on the NCBI Bioproject database (http://www.ncbi.nlm.nih.gov/bioproject/; accession PRJNA286122).
- Human PBMC: RNA-sequencing data for human PBMC were extracted from healthy donors in Zucca et al (2019) (39) and can be accessed under the GEO accession number GSE106443 and GSE115259, while MAPs were extracted from Murphy et al (2017) (30).
- Human B721.221: The B721.221 dataset was retrieved from Abelin et al (2017) (11); RNA sequencing data can be accessed under the GEO accession number GSE93315.
- Murine CT26: RNA-Seq data can be accessed under the GEO accession number GSE111092. Mass spectrometry data can be found on the ProteomeXchange Consortium via the PRIDE partner repository (human B-LCL: PXD004023 and murine CT26: PXD009065 and 10.6019/PXD009065).
- Murine EL4: MAP dataset was extracted from Murphy et al (2017) (30) and EL4 RNA sequencing dataset was extracted from Sidoli et al (2019) (40) and can be accessed under the GEO accession number GSE125384.

All figures were generated using R’s package “ggplot2”. Source code for pyGeno (https://github.com/tariqdaouda/pyGeno, doi: 10.12688/f1000research.8251.2) and Mariana (https://github.com/tariqdaouda/Mariana, doi: [to be provided after acceptance]) are freely available online.

Author Contributions: TD designed and performed all computational experiments (except those performed by AF and MC), wrote pyGeno and Mariana, generated figures, contributed to design of the iTR-OVA construct, co-wrote the first draft of the paper. MDL contributed to data analysis, to design and synthesis of the iTR-OVA construct, performed flow cytometry analysis, with input of EG, figure design, co-wrote the first draft of the paper. AF contributed to data analysis, study design and performed computational experiments (validation on 5 datasets and regressions). Y.Benslimane contributed to design and synthesis of the iTR-OVA construct, with input from LH and EG. RP produced virus for transduction of the iTR-OVA construct, transduced RAW cells, optimized and performed T-cell activation assay using mild fixation, with input from EG, and reviewed the manuscript. MC performed peptide affinity predictions. MB contributed to the optimization of culture conditions for the iTR-OVA assay. PT reviewed the manuscript. Y.Bengio reviewed and contributed to the manuscript. SL and CP contributed to study design, reviewed and contributed to the manuscript. All co-authors reviewed the manuscript.

The authors declare no competing interests.

References

1. Granados, D. P., Laumont, C. M., Thibault, P., and Perreault, C. (2015) The nature of self for T cells—a systems-level perspective. *Curr. Opin. Immunol.* 34, 1–8

2. Caron, E., Espona, L., Kowalewski, D. J., Schuster, H., Ternette, N., Alpízar, A., Schittenhelm, R. B., Ramarathinam, S. H., Lindestam Arlehamn, C. S., Chiek Koh, C., Gillet, L. C., Rabsteyn, A., Navarro, P., Kim, S., Lam, H., Sturm, T., Marcilla, M., Sette, A., Campbell, D. S., Deutsch, E. W., Moritz, R. L., Purcell, A. W., Rammensee, H.-G., Stevanovic, S., and Aebersold, R. (2015) An open-source computational and data resource to analyze digital maps of immunopeptidomes. *eLife* 4, e07661

3. Davis, M. M., Krogsgaard, M., Huse, M., Huppa, J., Lillemeier, B. F., and Li, Q. (2007) T Cells as a Self-Referential, Sensory Organ. *Annu. Rev. Immunol.* 25, 681–695

4. Schumacher, T. N., and Schreiber, R. D. (2015) Neoantigens in cancer immunotherapy. *Science* 348, 69–74

5. Caron, E., Vincent, K., Fortier, M.-H., Laverdure, J.-P., Bramoullé, A., Hardy, M.-P., Voisin, G., Roux, P. P., Lemieux, S., Thibault, P., and Perreault, C. (2011) The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation. *Mol. Syst. Biol.* 7, 533

6. Neefjes, J., Jongsma, M. L. M., Paul, P., and Bakke, O. (2011) Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* 11, 823–836

7. Bassani-Sternberg, M., and Gfeller, D. (2016) Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide–HLA Interactions. *J. Immunol.* 197, 2492–2499

8. Nielsen, M., and Andreatta, M. (2016) NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* 8,

9. Tenzer, S., Peters, B., Bulik, S., Schoor, O., Lemmel, C., Schatz, M. M., Kloetzel, P.-M., Rammensee, H.-G., Schild, H., and Holzhütter, H.-G. (2005) Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell. Mol. Life Sci. CMLS* 62, 1025–1037

10. Nielsen, M., Lundegaard, C., Lund, O., and Keşmir, C. (2005) The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* 57, 33–41

11. Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., Clauser, K. R., Hacohen, N., Rooney, M. S., Carr, S. A., and Wu, C. J. (2017) Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* 46, 315–326

12. Capietto, A.-H., Jhunjhunwala, S., and Delamarre, L. (2017) Characterizing neoantigens for personalized cancer immunotherapy. *Curr. Opin. Immunol.* 46, 58–65

13. Antón, L. C., and Yewdell, J. W. (2014) Translating DRiPs: MHC class I immunosurveillance of pathogens and tumors. *J. Leukoc. Biol.* 95, 551–562

14. Wei, J., Kishton, R. J., Angel, M., Conn, C. S., Dalla-Venezia, N., Marcel, V., Vincent, A., Catez, F., Ferré, S., Ayadi, L., Marchand, V., Dersh, D., Gibbs, J. S., Ivanov, I. P., Fridlyand, N., Couté, Y., Diaz, J.-J., Qian, S.-B., Staudt, L. M., Restifo, N. P., and Yewdell, J. W. (2019) Ribosomal Proteins Regulate MHC Class I Peptide Generation for Immunosurveillance. *Mol. Cell* 73, 1162-1173.e5

15. Yewdell, J. W., Antón, L. C., and Bennink, J. R. (1996) Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules? *J. Immunol.* 157, 1823–1826

16. Croft, N. P., Smith, S. A., Wong, Y. C., Tan, C. T., Dudek, N. L., Flesch, I. E. A., Lin, L. C. W., Tscharke, D. C., and Purcell, A. W. (2013) Kinetics of antigen expression and epitope presentation during virus infection. *PLoS Pathog.* 9, e1003129

17. Milner, E., Barnea, E., Beer, I., and Admon, A. (2006) The turnover kinetics of major histocompatibility complex peptides of human cancer cells. *Mol. Cell. Proteomics MCP* 5, 357–365

18. Hassan, C., Kester, M. G. D., de Ru, A. H., Hombrink, P., Drijfhout, J. W., Nijveen, H., Leunissen, J. A. M., Heemskerk, M. H. M., Falkenburg, J. H. F., and van Veelen, P. A. (2013) The human leukocyte antigen-presented ligandome of B lymphocytes. *Mol. Cell. Proteomics MCP* 12, 1829–1843

19. Pearson, H., Daouda, T., Granados, D. P., Durette, C., Bonneil, E., Courcelles, M., Rodenbrock, A., Laverdure, J.-P., Côté, C., Mader, S., Lemieux, S., Thibault, P., and Perreault, C. (2016) MHC class I–associated peptides derive from selective regions of the human genome. *J. Clin. Invest.* 126, 4690–4701

20. Cannarozzi, G., Schraudolph, N. N., Faty, M., von Rohr, P., Friberg, M. T., Roth, A. C., Gonnet, P., Gonnet, G., and Barral, Y. (2010) A Role for Codon Order in Translation Dynamics. *Cell* 141, 355–367

21. Plotkin, J. B., and Kudla, G. (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* 12, 32–42

22. Granados, D. P., Rodenbrock, A., Laverdure, J.-P., Côté, C., Caron-Lizotte, O., Carli, C., Pearson, H., Janelle, V., Durette, C., Bonneil, E., Roy, D. C., Delisle, J.-S., Lemieux, S., Thibault, P., and Perreault, C. (2016) Proteogenomic-based discovery of minor histocompatibility antigens with suitable features for immunotherapy of hematologic cancers. *Leukemia* 30, 1344–1354

23. Daouda, T., Perreault, C., and Lemieux, S. (2016) pyGeno: A Python package for precision medicine and proteogenomics. *F1000Research* 5, 381

24. Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003) A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* 3, 1137–1155

25. LeCun, Y., Bengio, Y., and Hinton, G. (2015) Deep learning. *Nature* 521, 436–444

26. Daouda, T. (2015) Mariana: The Cutest Deep learning Framework.

27. Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison Iii, C. A., and Smith, H. O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* 6, 343–345

28. Bell, C., English, L., Boulais, J., Chemali, M., Caron-Lizotte, O., Desjardins, M., and Thibault, P. (2013) Quantitative Proteomics Reveals the Induction of Mitophagy in Tumor Necrosis Factor-α-activated (TNFα) Macrophages. *Mol. Cell. Proteomics MCP* 12, 2394–2407

29. Karttunen, J., Sanderson, S., and Shastri, N. (1992) Detection of rare antigen-presenting cells by the lacZ T-cell activation assay suggests an expression cloning strategy for T-cell antigens. *Proc. Natl. Acad. Sci. U. S. A.* 89, 6020–6024

30. Murphy, J. P., Konda, P., Kowalewski, D. J., Schuster, H., Clements, D., Kim, Y., Cohen, A. M., Sharif, T., Nielsen, M., Stevanovic, S., Lee, P. W., and Gujar, S. (2017) MHC-I Ligand Discovery Using Targeted Database Searches of Mass Spectrometry Data: Implications for T-Cell Immunotherapies. *J. Proteome Res.* 16, 1806–1816

31. Laumont, C. M., Vincent, K., Hesnard, L., Audemard, É., Bonneil, É., Laverdure, J.-P., Gendron, P., Courcelles, M., Hardy, M.-P., Côté, C., Durette, C., St-Pierre, C., Benhammadi, M., Lanoix, J., Vobecky, S., Haddad, E., Lemieux, S., Thibault, P., and Perreault, C. (2018) Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* 10, eaau5516

32. Dersh, D., Yewdell, J. W., and Wei, J. (2019) A SIINFEKL-Based System to Measure MHC Class I Antigen Presentation Efficiency and Kinetics. *Methods Mol. Biol. Clifton NJ* 1988, 109–122

33. Kim, J. H., Lee, S.-R., Li, L.-H., Park, H.-J., Park, J.-H., Lee, K. Y., Kim, M.-K., Shin, B. A., and Choi, S.-Y. (2011) High Cleavage Efficiency of a 2A Peptide Derived from Porcine Teschovirus-1 in Human Cell Lines, Zebrafish and Mice. *PLOS ONE* 6, e18556

34. Shastri, N., and Gonzalez, F. (1993) Endogenous generation and presentation of the ovalbumin peptide/Kb complex to T cells. *J. Immunol. Baltim. Md 1950* 150, 2724–2736

35. Laumont, C. M., Daouda, T., Laverdure, J.-P., Bonneil, É., Caron-Lizotte, O., Hardy, M.-P., Granados, D. P., Durette, C., Lemieux, S., Thibault, P., and Perreault, C. (2016) Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.* 7, 10238

36. Zanker, D. J., Oveissi, S., Tscharke, D. C., Duan, M., Wan, S., Zhang, X., Xiao, K., Mifsud, N. A., Gibbs, J., Izzard, L., Dlugolenski, D., Faou, P., Laurie, K. L., Vigneron, N., Barr, I. G., Stambas, J., Van den Eynde, B. J., Bennink, J. R., Yewdell, J. W., and Chen, W. (2019) Influenza A Virus Infection Induces Viral and Cellular Defective Ribosomal Products Encoded by Alternative Reading Frames. *J. Immunol. Baltim. Md 1950* 202, 3370–3380

37. Ehx, G., and Perreault, C. (2019) Discovery and characterization of actionable tumor antigens. *Genome Med.* 11, 29

38. Sette, A., and Fikes, J. (2003) Epitope-based vaccines: an update on epitope identification, vaccine design and delivery. *Curr. Opin. Immunol.* 15, 461–470

39. Zucca, S., Gagliardi, S., Pandini, C., Diamanti, L., Bordoni, M., Sproviero, D., Arigoni, M., Olivero, M., Pansarasa, O., Ceroni, M., Calogero, R., and Cereda, C. (2019) RNA-Seq profiling in peripheral blood mononuclear cells of amyotrophic lateral sclerosis patients and controls. *Sci. Data* 6, 190006

40. Sidoli, S., Lopes, M., Lund, P. J., Goldman, N., Fasolino, M., Coradin, M., Kulej, K., Bhanu, N. V., Vahedi, G., and Garcia, B. A. (2019) A mass spectrometry-based assay using metabolic labeling to rapidly monitor chromatin accessibility of modified histone proteins. *Sci. Rep.* 9, 13613

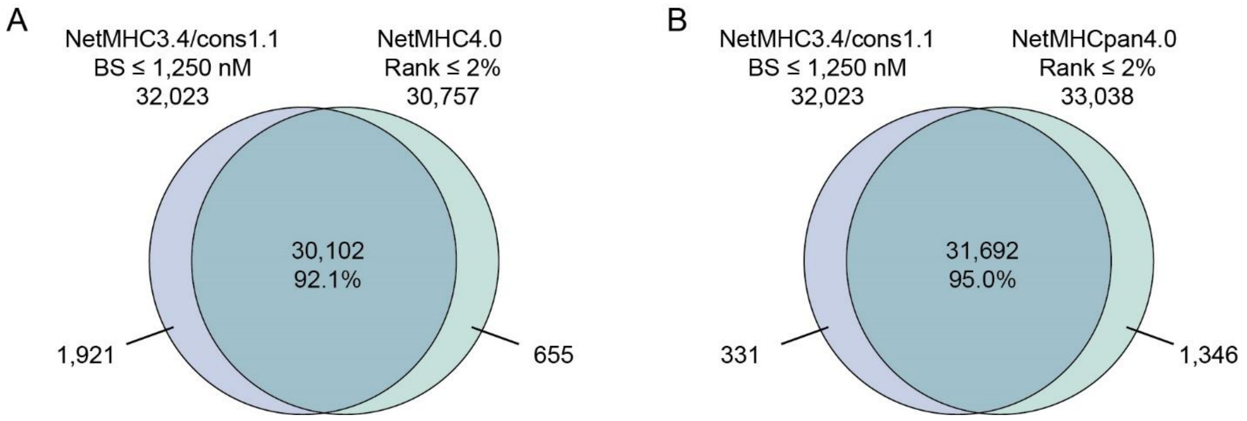
Supplemental Information

Article: Codon arrangement modulates MHC-I peptides presentation

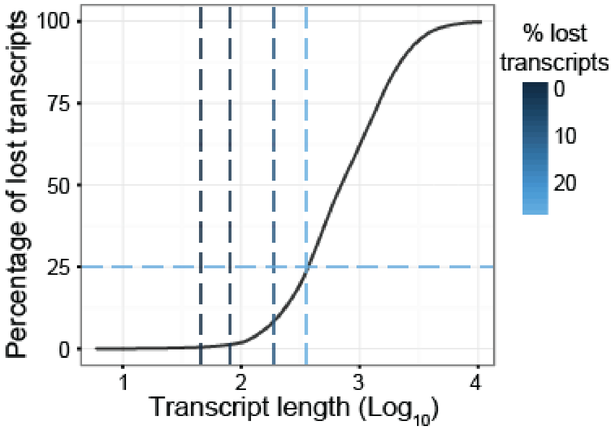
Authors: Tariq Daouda, Maude Dumont-Lagacé, Albert Feghaly, Yahya Benslimane, Rébecca Panes, Mathieu Courcelles, Mohamed Benhammadi, Lea Harington, Pierre Thibault, François Major, Yoshua Bengio, Étienne Gagnon, Sébastien Lemieux, Claude Perreault

Table of Contents

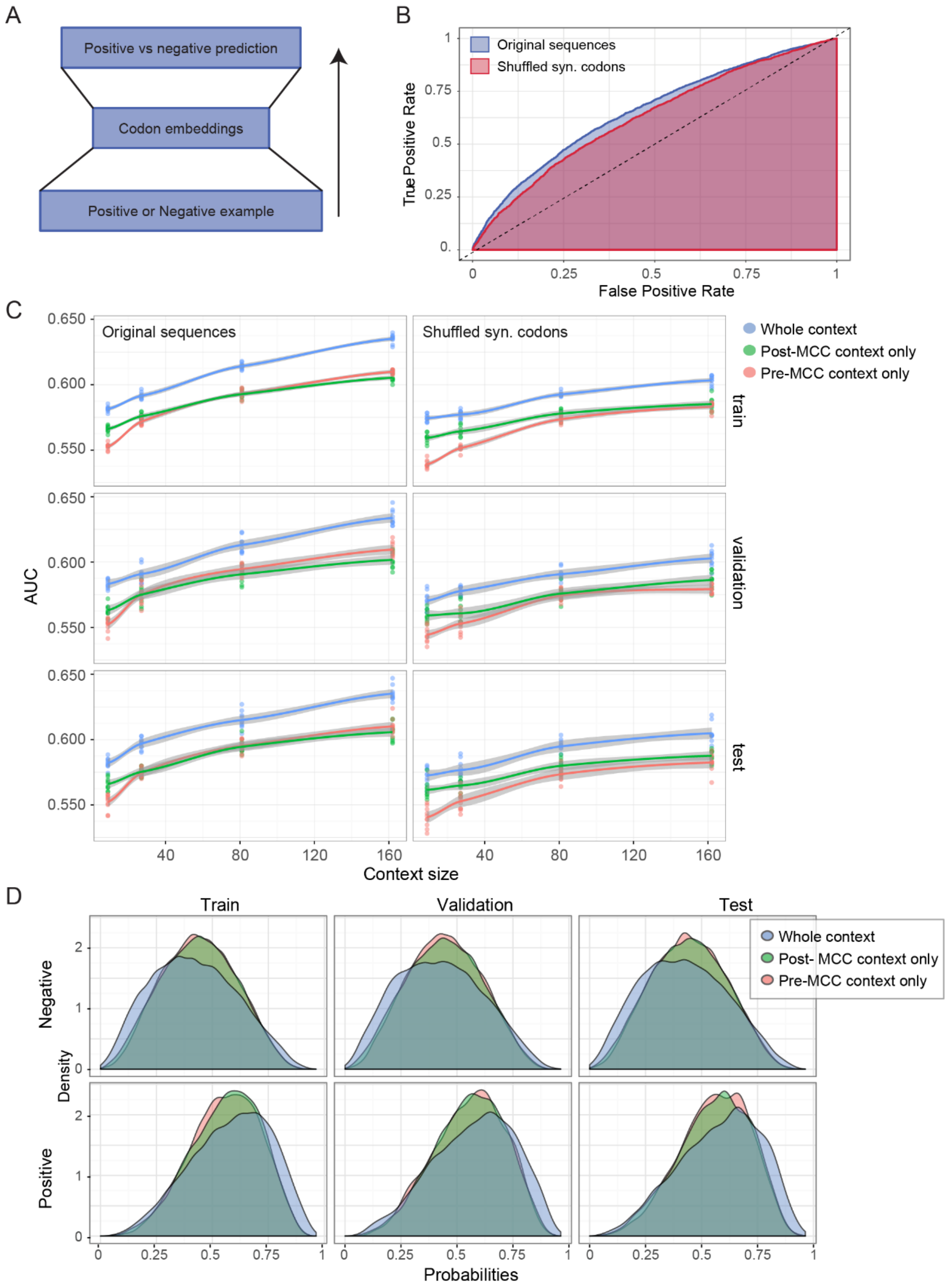
Supplementary Figures



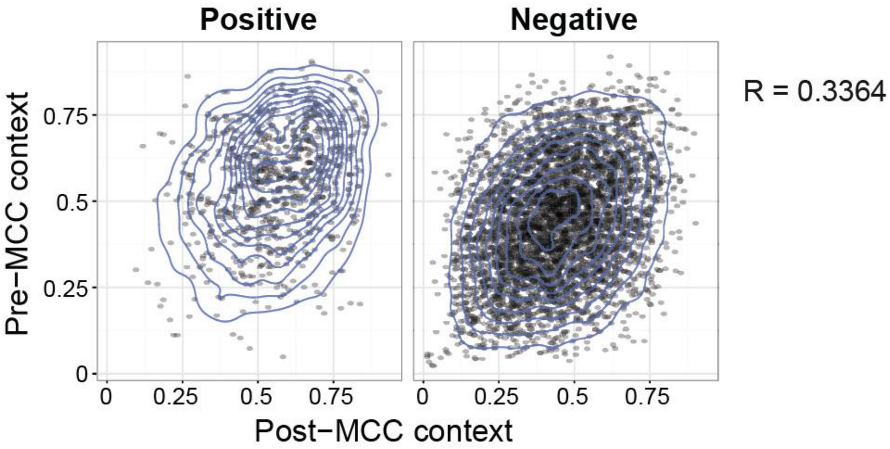
Supplementary Figure S1. Validation of MHC-I associated peptides (MAP) dataset from Pearson H. *et al*. (2016) using the new versions of MAP binding affinity prediction algorithm NetMHC4.0 (A) and NetMHCpan4.0 (B).



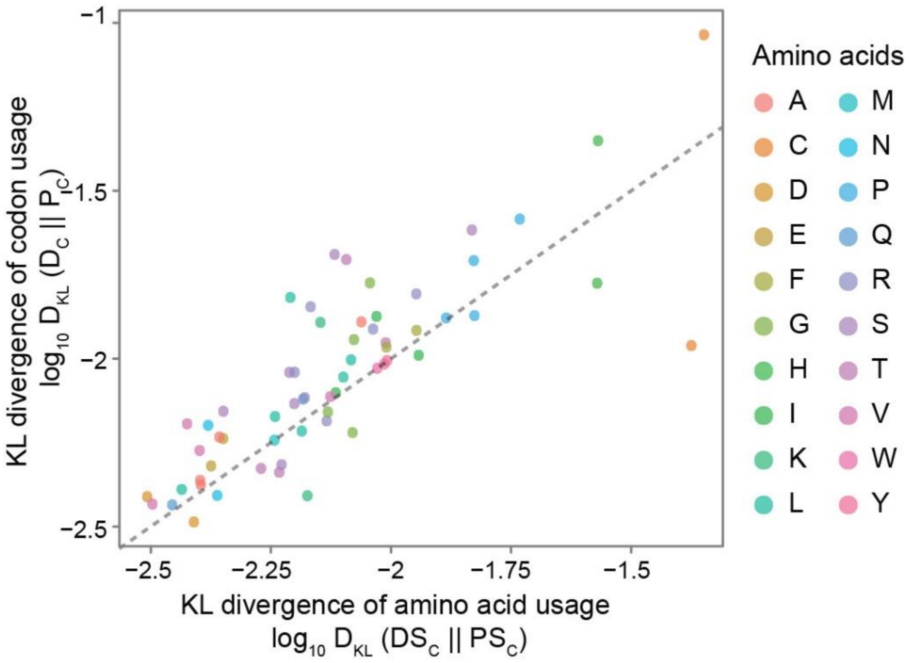
Supplementary Figure S2. Percentage of transcript ineligibility as a function of context size. Transcript length corresponds to *C x 2 + 27*, where *C* is the context size in nucleotides and 27 the length of the MCCs. Related to Figure 1A.



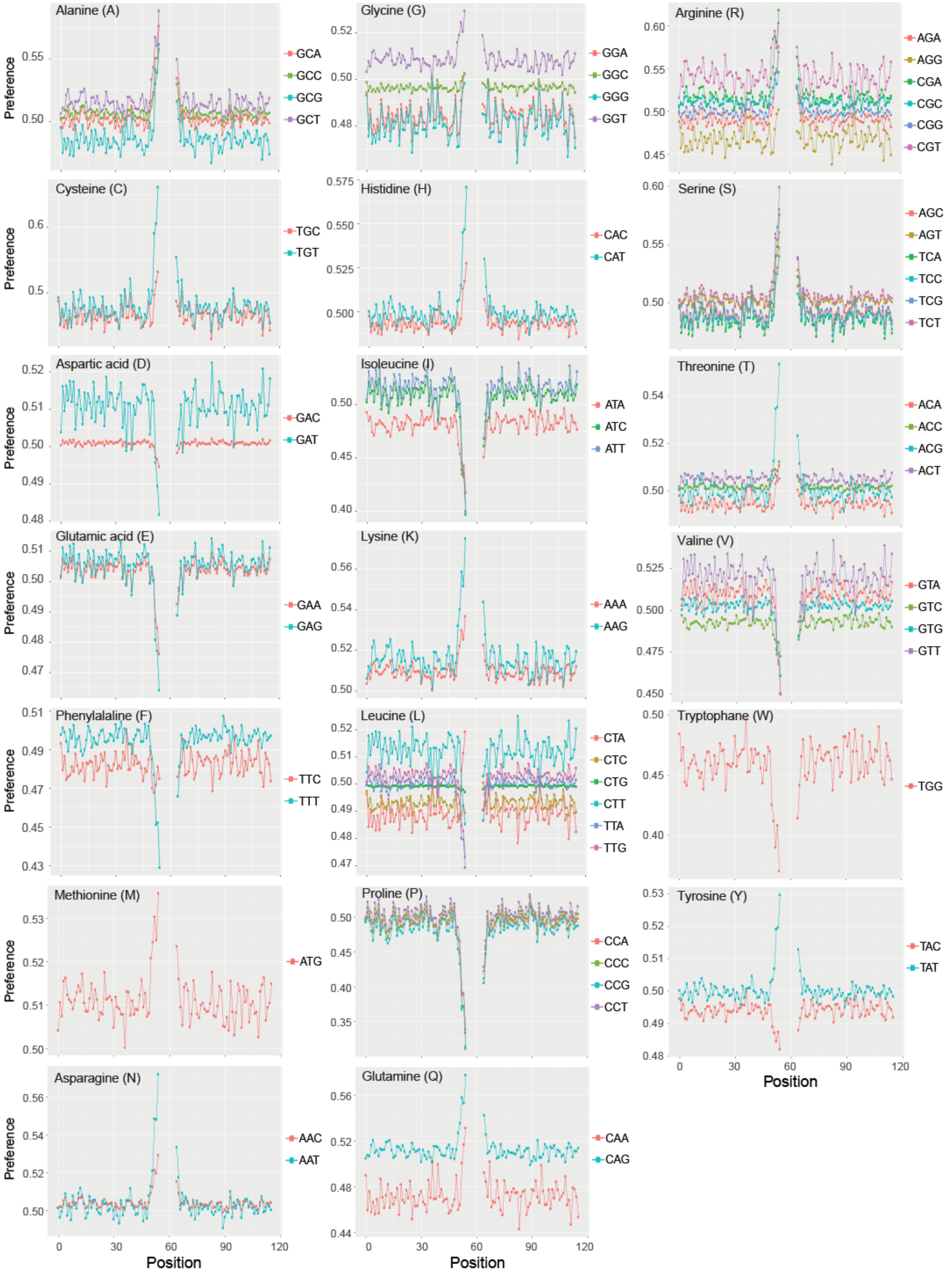
Supplementary Figure S3. CAMAP architecture and detailed predictions. (A) Architecture of the ANN used in this work. (B) ROC curves for a CAMAP trained on a context size of 162 nucleotides on original sequences or sequences with shuffled synonyms. (C) Results for the AUC on all train, validation and test subsets. Grey areas represent the 95% confidence intervals. (D) Distributions of output probabilities of CAMAPs used to calculate correlations in Supplementary Figure S4.



Supplementary Figure S4. Correlation between CAMAP prediction score trained only with pre-MCC or post-MCC sequences. For each sequence in the test set we calculated the average prediction score given by CAMAPs in each condition, and calculated the Pearson correlation using the R software. Densities were calculated on all points and drawn using ggplot2. Only a random subset of the points is represented in the figures to limit their size.



Supplementary Figure S5. Kullback-Leibler divergence between hit and decoy dataset in original codon (y-axis) or shuffled synonymous codon sequences (x-axis). Shuffled sequences represent amino acid usage, as codon-specific information are removed with synonymous codon shuffling.



Supplementary Figure S6. CAMAP preferences per position for all codons. See Experimental Procedures for more details.

Supplementary Tables

Supplementary Table S1. Nucleotide sequences of the EP and RP constructs. SIINFEKL MCC is shown in bold, while the variant regions (pre- and post-MCC contexts of 162-nucleotides) are in blue and italics. Related to Fig. 6.

OVA-EP
ATGGGCTCCATCGGTGCAGCAAGCATGGAATTTTGTTTTGATGTATTCAAGGAGCTCAAAGTCCACCATGCCAATGAGAACATCTTCTACTGCCCCATTGCCATCATGTCAGCTCTAGCCATGGTATACCTGGGTGCAAAAGACAGCACCAGGACACAAATAAATAAGGTTGTTCGCTTTGATAAACTTCCAGGATTCGGAGACAGTATTGAAGCTCAGTGTGGCACATCTGTAAACGTTCACTCTTCACTTAGAGACATCCTCAACCAAATCACCAAACCAAATGATGTTTATTCGTTCAGCCTTGCCAGTAGACTTTATGCTGAAGAGAGATACCCAATCCTGCCAGAATACTTGCAGTGTGTGAAGGAACTGTATAGAGGAGGCTTGGAACCTATCAACTTTCAAACAGCTGCAGATCAAGCCAGAGAGCTCATCAATTCCTGGGTAGAAAGTCAGACAAATGGAATTATCAGAAATGTCCTTCAGCCAAGCTCCGTGGATTCTCAAACTGCAATGGTTCTGGTTAATGCCATTGTCTTCAAAGGACTGTGGGAGAAAGCATTTAAGGATGAAGACACACAAGCAATGCCTTTCAGAGTGACTGAG*CAGGAGTCTAAGCCTGTTCAGATGATGTATCAGATTGGTCTTTTTCGTGTTGCTTCTATGGCTTCTGAGAAGATGAAGATTCTTGAGCTTCCTTTTGCTAGTGGTACTATGTCTATGCTTGTTCTTCTTCCTGATGAGGTTTCTGGTCTTGAGCAGCTTGAA***AGTATAATCAACTTTGAAAAACTG***ACTGAGTGGACTTCTTCTAACGTTATGGAGGAGCGTAAGATTAAGGTTTATCTTCCTCGTATGAAGATGGAGGAGAAGTATAACCTTACTTCTGTTCTTATGGCTATGGGAATTACTGATGTTTTTTCTAGTTCTGCTAACCTTAGTGGTATTTCTTCGGCT*GAGAGCCTGAAGATATCTCAAGCTGTCCATGCAGCACATGCAGAAATCAATGAAGCAGGCAGAGAGGTGGTAGGGTCAGCAGAGGCTGGAGTGGATGCTGCAAGCGTCTCTGAAGAATTTAGGGCTGACCATCCATTCCTCTTCTGTATCAAGCACATCGCAACCAACGCCGTTCTCTTCTTTGGCAGATGTGTTTCCCCTTAA
OVA-RP
ATGGGCTCCATCGGTGCAGCAAGCATGGAATTTTGTTTTGATGTATTCAAGGAGCTCAAAGTCCACCATGCCAATGAGAACATCTTCTACTGCCCCATTGCCATCATGTCAGCTCTAGCCATGGTATACCTGGGTGCAAAAGACAGCACCAGGACACAAATAAATAAGGTTGTTCGCTTTGATAAACTTCCAGGATTCGGAGACAGTATTGAAGCTCAGTGTGGCACATCTGTAAACGTTCACTCTTCACTTAGAGACATCCTCAACCAAATCACCAAACCAAATGATGTTTATTCGTTCAGCCTTGCCAGTAGACTTTATGCTGAAGAGAGATACCCAATCCTGCCAGAATACTTGCAGTGTGTGAAGGAACTGTATAGAGGAGGCTTGGAACCTATCAACTTTCAAACAGCTGCAGATCAAGCCAGAGAGCTCATCAATTCCTGGGTAGAAAGTCAGACAAATGGAATTATCAGAAATGTCCTTCAGCCAAGCTCCGTGGATTCTCAAACTGCAATGGTTCTGGTTAATGCCATTGTCTTCAAAGGACTGTGGGAGAAAGCATTTAAGGATGAAGACACACAAGCAATGCCTTTCAGAGTGACTGAG*CAAGAATCCAAACCGGTCCAAATGATGTACCAAATAGGGCTATTCAGGGTCGCGTCCATGGCGTCCGAAAAAATGAAAATACTAGAACTACCGTTCGCGTCAGGGACGATGTCCATGCTCGTCCTACTACCGGACGAAGTCTCCGGACTCGAACAACTCGAG***AGTATAATCAACTTTGAAAAACTG***ACAGAATGGACATCCTCCAATGTCATGGAAGAAAGGAAAATAAAAGTCTACCTCCCGAGGATGAAAATGGAAGAAAAATACAATCTAACATCCGTCCTAATGGCGATGGGTATAACAGACGTCTTCTCCTCATCCGCGAATCTATCAGGGATATCCAGCGCG*GAGAGCCTGAAGATATCTCAAGCTGTCCATGCAGCACATGCAGAAATCAATGAAGCAGGCAGAGAGGTGGTAGGGTCAGCAGAGGCTGGAGTGGATGCTGCAAGCGTCTCTGAAGAATTTAGGGCTGACCATCCATTCCTCTTCTGTATCAAGCACATCGCAACCAACGCCGTTCTCTTCTTTGGCAGATGTGTTTCCCCTTAA