CAMAP: Artificial neural networks unveil the role of codon arrangement in modulating MHC-I peptides presentation

Tariq Daouda[1,2,*], Maude Dumont-Lagacé[1,3,*], 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

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: Tariq Daouda

Email: tdaouda@broadinstitute.org

*These authors contributed equally.

30 **Abstract**

31   MHC-I associated peptides (MAPs) play a central role in the elimination of virus-infected and

32   neoplastic cells by CD8 T cells. However, accurately predicting the MAP repertoire remains

33   difficult, because only a fraction of the transcriptome generates MAPs. In this study, we

34   investigated whether codon arrangement (usage and placement) regulates MAP biogenesis. We

35   developed an artificial neural network called Codon Arrangement MAP Predictor (CAMAP),

36   predicting MAP presentation solely from mRNA sequences flanking the MAP-coding codons

37   (MCCs), while excluding the MCC *per se*. CAMAP predictions were significantly more accurate

38   when using original codon sequences than shuffled codon sequences which reflect amino acid

39   usage. Furthermore, predictions were independent of mRNA expression and MAP binding affinity

40   to MHC-I molecules and applied to several cell types and species. Combining MAP ligand scores,

41   transcript expression level and CAMAP scores was particularly useful to increaser MAP prediction

42   accuracy. Using an *in vitro* assay, we showed that varying the synonymous codons in the regions

43   flanking the MCCs (without changing the amino acid sequence) resulted in significant modulation

44   of MAP presentation at the cell surface. Taken together, our results demonstrate the role of codon

45   arrangement in the regulation of MAP presentation and support integration of both translational

46   and post-translational events in predictive algorithms to ameliorate modeling of the

47   immunopeptidome.

48

49

Author summary

51   MHC-I associated peptides (MAPs) are small fragments of intracellular proteins presented at the

52   surface of cells and used by the immune system to detect and eliminate cancerous or virus-infected

53   cells. While it is theoretically possible to predict which portions of the intracellular proteins will

54   be naturally processed by the cells to ultimately reach the surface, current methodologies have

55   prohibitively high false discovery rates. Here we introduce an artificial neural network called

56   Codon Arrangement MAP Predictor (CAMAP) which integrates information from mRNA-to-

57   protein translation to other factors regulating MAP biogenesis (e.g. MAP ligand score and

58   transcript expression levels) to improve MAP prediction accuracy. While most MAP predictive

59   approaches focus on MAP sequences per se, CAMAP's novelty is to analyze the MAP-flanking

60   mRNA sequences, thereby providing completely independent information for MAP prediction.

61   We show on several datasets that the integration of CAMAP scores with other known factors

62   involved in MAP presentation (i.e. MAP ligand score and mRNA expression) significantly

63   improves MAP prediction accuracy, and further validate CAMAP learned features using an *in-*

64   *vitro* assay. These findings may have major implications for the design of vaccines against cancers

65   and viruses, and in times of pandemics could accelerate the identification of relevant MAPs of

66   viral origins.

67

68   **Abbreviations**: MHC-I: major histocompatibility complex class-I, MAP: MHC-I associated

69   peptides, CAMAP: Codon arrangement MAP predictor, DRiP: defective ribosomal product,

70   ANN: artificial neural network, MCC: MAP-coding codons, B-LCL: B-lymphoblastoid cell line,

71   KL: Kullback-Leibler, BS: binding score, OVA: ovalbumin protein, WT: wildtype, EP:

72   enhanced presentation, RP: reduced presentation.

73

Introduction

75 In jawed vertebrates, virtually all nucleated cells present at their surface major histocompatibility

76 complex class-I (MHC-I) associated peptides (MAPs), collectively referred to as the

77 immunopeptidome [1,2]. MAPs play a central role in shaping the adaptive immune system, as they

78 orchestrate the development, survival and activation of CD8 T cells [3]. Moreover, recognition of

79 abnormal MAPs is essential to the elimination of virus-infected and neoplastic cells [4]. Therefore,

80 systems-level understanding of MAP biogenesis and molecular composition remains a central

81 issue in immunobiology [5,6].

82 The generation of the immunopeptidome can be conceptualized in two main events: (a) the

83 generation of MAP candidates (i.e. peptides of appropriate length for MHC-I presentation) through

84 protein degradation, and (b) a subsequent filtering step through the binding of MAP candidates to

85 the available MHC-I molecules. Rules that regulate the second event have been well characterized

86 using artificial neural networks (ANN) and weighted matrix approaches [7,8]. However,

87 accurately predicting which peptides will ultimately reach MHC-I molecules following a multistep

88 processing in the cytosol and endoplasmic reticulum remains an open question [6]. Most efforts at

89 modeling MAP generation have focused on post-translational events and their regulation by the

90 amino acid sequence of MAPs and of directly adjacent residues (typically 10-mers at the N- and

91 C-termini). While the consideration of preferential sites of proteasome cleavage has proven useful

92 to enrich for MAP candidates [9], it remains insufficient for MAP prediction, due to prohibitive

93 false discovery rates [10–12].

94 A large body of evidence suggests that a substantial portion of MAPs are produced co-

95 translationally [13–15], deriving from defective ribosomal products (DRiPs), that is, polypeptides

96 that fail to achieve a stable conformation during translation and are consequently rapidly degraded.

97   This concept was initially supported by two observations: (i) viral MAPs can be detected within

98   minutes after viral infection, much earlier than their associated proteins half-life [16], and (ii) MAP

99   presentation correlates more closely with translation rate than with overall protein abundance

100   [17,18]. In addition, while all proteins contain peptides that are predicted to bind MHC-I

101   molecules, mass spectrometry analyses have revealed that the immunopeptidome is not a random

102   excerpt of the transcriptome or the proteome [1,19]. Indeed, proteogenomic analyses of 25,270

103   MAPs isolated from B lymphocytes of 18 individuals showed that 41% of expressed protein-

104   coding genes generated no MAPs [19]. These authors also provided compelling evidence that the

105   presentation of MAPs cannot be explained solely by their affinity to MHC-I alleles and their

106   transcript expression levels, while ruling out low mass spectrometry sensitivity as an explanation

107   for the non-presentation of the strong binders. Because (i) MAPs appear to preferentially derive

108   from DRiPs and (ii) codon usage influences both precision and efficiency of protein synthesis

109   [20,21], we hypothesized that codon usage in the vicinity of MAP-coding codons (MCCs) might

110   significantly contribute to the regulation of MAP biogenesis. We developed an artificial neural

111   network called Codon Arrangement MAP Predictor (CAMAP), trained to identify MCCs flanking

112   regions. We then used CAMAP to uncover key codon features that characterize mRNA sequences

113   encoding for MAPs (i.e. source) when compared to sequences that do not (i.e. non-source).

114

115   **Results**

116   **Dataset description**

117   We analyzed a previously published dataset consisting of MAPs presented on B lymphoblastoid

118   cell line (B-LCL) by a total of 33 MHC-I alleles from 18 subjects [19,22]. Because we were

119 searching for features that influence MAP generation and not the binding of MAP to MHC-I

120 molecules, we elected to analyze the MCC flanking sequences only and excluded the MCCs *per*

121 *se* from our positive (hits) and negative (decoys) sequences (Fig. 1A). To facilitate data analysis

122 and interpretation, we restricted our hit dataset to MAPs with a length of 9 amino acids, for a total

123 of 19,656 9-mer MAPs (which represents 78% of MAPs in this dataset). We next created a decoy

124 dataset from transcripts that generated no MAPs, by randomly selecting 98,290 9-mers from these

125 transcripts. Finally, we used pyGeno [23] to extract MCCs flanking regions corresponding to both

126 hit and decoy MAPs, which constituted our final dataset for CAMAP. Of note, each sequence in

127 the final dataset is unique and derives from the canonical reading frame. In addition, in order to

128 investigate the relative importance of codon vs. amino acid usage in MAP biogenesis, we

129 generated a dataset of shuffled sequences (for both positive and negative datasets) in which original

130 codon sequences were randomly replaced by synonymous codons according to their usage

131 frequency in the dataset (Fig. 1B). This transformation was performed to ensure that both neural

132 networks received the same number of parameters as input, preventing the introduction of a

133 favorable bias for the codon network. The random shuffling causes any codon-specific feature to

134 be shared among synonyms, thereby causing the shuffled codon distribution to reflect the amino

135 acid usage (see Materials and Methods for more details). Indeed, codon distributions in the

136 shuffled datasets more closely reflected those of their corresponding amino acid than in the original

137 dataset (Supplementary Figure S1), with 92% of codons in the shuffled dataset showing a strong

138 correlating ($R^2 > 0.95$) with the amino acid distribution, compared to only 69% in the original

139 dataset ($p < 2x10^{-16}$, Supplementary Figure S2). Importantly, this shuffling does not affect the

140 resulting amino acid sequence thereby preserving all potential amino acid-related motifs.

141   Distributions of each codons in the original VS shuffled dataset and compared to its corresponding

142   amino acid can be found in Supplementary Figure S3.

143   **Figure 1. Construction of the dataset**. (a) Transcripts expressed in B cells from 18 subjects were

144   considered as source or non-source transcripts depending on their match with at least one MAP.

145   Because we were searching for features that might influence MAP generation and not the binding

146   of MAP to MHC-I, we focused our attention on mRNA sequences adjacent to the nine MCCs (i.e.

147   up to 162 nucleotides on each side of MCCs). (b) Creation of the shuffled dataset. Codons were

148   randomly replaced by a synonymous codon according to their respective frequencies (i.e. codon

149   usage) in the dataset. The random shuffling causes any codon-specific feature to be shared among

150   synonyms, thereby causing the shuffled codon distribution to reflect the amino acid usage.

151   Importantly, both the original sequence and its shuffled version translates into the same amino

152   acids.

153

154   **CAMAP links codon usage to MAP presentation**

155   To assess the importance of codon usage in MAP biogenesis, we reasoned that if codons bear

156   important information that is operative at the translational rather than the post-translational level,

157   then: (i) CAMAP trained to identify MCCs flanking regions should consistently perform better

158   when trained on original codon sequences than on shuffled codon sequences (reflecting amino acid

159   sequences), and (ii) synonymous codons should have different effects on the prediction. To test

160   these hypotheses, CAMAP received as inputs MCCs flanking regions from hit and decoy

161   sequences from either the original or shuffled datasets. It was then trained to predict the probability

162   that individual input sequences were MCCs flanking regions (i.e. hit) rather than sequences from

163   the negative dataset (Supplementary Figure S4A).

164   We compared CAMAP performance when predicting MAP presentation from original codon

165   sequences, versus shuffled sequences representing amino acid arrangement. To evaluate the

7

166 robustness of our approach, 12 different CAMAPs were trained in parallel, with different train-

167 validation-test splits of the dataset. Our results show that predictions were consistently better when

168 CAMAP received the original codons rather than the shuffled sequences (Fig. 2A). CAMAPs

169 receiving information from both pre-MCCs and post-MCCs sequences (i.e. whole MCC flanking

170 context) also performed better than when receiving only pre- or post-MCCs context (Fig. 2A and

171 Supplementary Figure S4B-C), suggesting that pre- and post-MCCs context are not redundant.

172 Indeed, we found a weak correlation between the prediction scores of CAMAPs trained only with

173 pre- or post-MCCs sequences (Supplementary Fig. S5). In addition, CAMAPs receiving longer

174 sequences performed better than those receiving shorter sequences (Fig. 2B). Because sequences

175 located far upstream and downstream of the MCCs (i.e. in ranges exceeding the direct influence

176 of proteases) are informative regarding MAP presentation, it supports the existence of factors

177 unrelated to protein degradation modulating MAP presentation.

178 **Figure 2. CAMAP predictions on MAP-flanking sequences**. (A) Area under the curve (AUC)
179 score for CAMAPs trained with whole MCCs context, versus CAMAPs trained with only pre- or
180 post-MCCs context. All CAMAPs presented here were trained with a context size of 162
181 nucleotides. (B) AUC for CAMAPs trained with codon context sizes of 9, 27, 81 and 162
182 nucleotides (context here refer to mRNA sequences flanking the MCCs).

183

184 Both MAP binding affinity to the MHC-I molecule and the level of gene expression are predictive

185 of MAP presentation [19]. Because codon usage has been shown to be different in highly expressed

186 genes, we wanted to verify whether the codon-specific rules captured by CAMAP were associated

187 with potential biases in our positive dataset, which is enriched in highly expressed genes. We first

188 show that there is no correlation between gene expression levels and CAMAP scores in both the

189 positive and negative datasets (R < 0.1, Fig. 3A). This was true for both average expression levels

190 across our samples (Fig. 3A), and for samples individually (see Supplementary Fig. S6). Secondly,

191 we trained CAMAP networks using a decoy dataset that mirrored the positive dataset gene

192 expression level (Supplementary Fig. S7A) and showed similar results: CAMAP trained on

193 original codon sequence performed better than CAMAP trained on shuffled sequences

194 (Supplementary Fig. S7B). These results show that the codon-specific rules captured by CAMAP

195 trained on original sequences are independent of gene expression levels.

196 **Figure 3. Correlation between CAMAP prediction score and (A) transcript expression levels**

197 **and (B) MAP binding affinity.** CAMAP used here was trained on original codon sequences using

198 a context size of 162 nucleotides (both pre- and post-MCCs context).

199

200 We stipulate that the presence of MHC-I binding motifs in the MCCs in the positive dataset might

201 be associated with biases in the MAP-flanking regions, which could also influence CAMAP

202 training. Therefore, to evaluate the presence of this potential bias, we first evaluated the correlation

203 between CAMAP scores and MAPs binding affinity. Again, our result showed no correlation

204 between CAMAP scores and MAP binding affinity, both when considering the minimal binding

205 affinity of each MAP to the MHC-I alleles contained in our dataset (Fig. 3B) or when considering

206 each allele individually (Supplementary Fig. S8). Secondly, we trained CAMAP networks using a

207 decoy dataset that mirrored the positive dataset MAP binding affinities (Supplementary Fig. S9A).

208 Again, CAMAPs trained on original codon sequence performed better than CAMAPs trained on

209 shuffled sequences (Supplementary Fig. S9B). These results show that codon-specific rules

210 captured by CAMAP trained on original sequences are independent of MAP binding affinities and

211 of potential biases in codon usage of MAP-flanking sequences associated with the presence of an

212 MHC-I binding motif in the MCCs.

213   We next evaluated the possibility of biases associated with many MAPs originating from

214   conserved regions (e.g., found in multiple domains of the same domain family such as zinc fingers

215   or kinases). We first evaluated MAPs that could originate from different transcripts within the

216   transcriptome (i.e. transcripts with sufficient expression levels detected by RNA sequencing) as

217   they are likely to represent conserved regions in the genome. While 79.9% of MAP originated

218   from unique contexts (Supplementary Fig. S10A), 2.1% of MAPs had more than 3 possible origins,

219   which represented 11.7% of the hit dataset (Supplementary Fig. S10B). These MAPs with several

220   possible origins preferentially derived from zinc finger proteins, which are known to share

221   homologous regions (Supplementary Fig. S11). We therefore trained CAMAPs with datasets

222   excluding entries encoding for MAPs that had >3 or >10 possible origins and compared their

223   performance with that of CAMAPs trained without excluding these MAPs. Our results show that

224   whatever the dataset used, CAMAP trained with original sequences always significantly

225   outperformed CAMAP trained with shuffled sequences (Supplementary Fig. S12). Taken together,

226   these results suggest that the codon-specific rules captured by CAMAP are independent of

227   potential homologies in the hit dataset, as they do not appear to influence CAMAP performance.

228   We next validated our CAMAP trained on 9-mer MAPs derived from B-LCL using 5 datasets

229   derived from different human and mouse cell types. All the validation datasets were described

230   through proteogenomic analyses similarly to our B-LCL training datasets. However, all the

231   validation datasets included MAPs of 8-11 mers, in contrast with the training dataset that contained

232   only 9-mer MAPs. The validation datasets consisted of (i) our B-LCL dataset, this time including

233   all peptide lengths [19,22], (ii) a dataset of human peripheral blood mononucleated cells or PBMCs

234   [24], (iii) a dataset of B-lymphoblastoid cells expressing unique HLA alleles (B721.221 [11]), (iv)

235   murine colon carcinoma cell line (CT26) and (v) a murine lymphoma cell line (EL4, [24,25]). For

236   all datasets, we created hit and decoy datasets of original and shuffled sequences using the same

237   approach described above but including MAPs of 8-11 amino acids. Notably, CAMAPs trained on

238   human sequences encoding 9-mers MAPs from one human cell type (i.e. B-LCL) could also

239   predict presentation of 8-11 mers MAPs in other human cell types (Fig. 4), as well as from mouse

240   cell lines, albeit with lower performances (Fig. 4). Here again, CAMAPs trained on original

241   sequences consistently outperformed CAMAPs trained on shuffled sequences (Fig. 4). These

242   results show that the codon-specific rules derived by CAMAPs to predict MAP presentation are

243   valid across different cell types, and can even be applied to another species, albeit with slightly

244   lower performances. These results support a role for codons in the modulation of MAP

245   presentation.

246   **Figure 4. Validation of CAMAP predictions on 5 datasets derived from human and murine**

247   **cell lines**. CAMAP prediction score for different datasets derived from humans (i.e. B-LCL,

248   PBMCs and B721.221) or mouse (i.e. CT26 and EL4) cells. Of note, all CAMAPs were trained on

249   B-LCL-derived sequences encoding for 9-mer MAPs only with a context size of 162 nucleotides.

250   Results are reported for 8 to 11-mer MAPs derived from the 5 datasets. In all panels, 12 CAMAPs

251   trained with original or shuffled synonymous sequences were compared (significance assessed

252   using Student T test).

253

254   The lower performances of CAMAP trained with shuffled sequences (representing amino acid

255   distribution) suggests that amino acids in MAP-flanking sequences are less informative than

256   codons regarding MAP presentation. We formally quantified this difference in information using

257   the Kullback-Leibler (KL) divergence (see Materials and Methods for more details). Most codons

258   (47/61, 77%) showed greater KL divergence in the original dataset than the shuffled dataset,

259   indicating that codon distributions contained more information with regards to MAP presentation

260 than amino acid distributions (Supplementary Fig. S13). These results suggest that codons in

261 MAP-flanking regions play a role that is non-redundant with amino acids in MAP biogenesis.

262 We wondered whether some regions were more influential on MAP presentation than others. To

263 address this question, we retrieved the model preferences for each codon at each position. The

264 preferences correspond to the prediction score of our best model (trained with original codon

265 sequences for a context size of 162 nucleotides) when a single codon at a single position is

266 provided as input (all other positions being set at [0,0] coordinates in the embedding space). The

267 model's preferences are therefore a measure of each individual codon's propensity to increase or

268 decrease the model's output probability as a function of its position relative to the MCCs. A value

269 of 0.5 denotes a neutral preference, while negative and positive preferences correspond to values

270 below and above 0.5, respectively. Preferences were obtained by feeding CAMAP sequences in

271 which all codon values were masked, except for a single position that received a non-null codon

272 label.

273 Interestingly, while codons closest to the MCCs were the most influential on CAMAP scores,

274 some synonymous codons showed opposite effects, further demonstrating that codon usage does

275 not recapitulate amino acid usage (Fig. 5A-B and Supplementary Fig. S14). The use of embeddings

276 to encode codons has the advantage of arranging them into a semantic space, wherein codons with

277 similar influences are positioned close to each other. Interestingly, most synonymous codons did

278 not form clusters, with a notable exception being proline codons (Fig. 5C). This finding indicates

279 that for some codons, their effect on CAMAP prediction score may be closer to that of a non-

280 synonymous codon than to that of one of its synonyms.

281 **Figure 5. CAMAP interpretation of codon impact on MAP biogenesis.** Preferences for a

282 network trained on a context of 162 nucleotides (54 codons) for (A) serine, proline and tyrosine

283   codons, and (B) leucine codons. (C) Learned codon embeddings. Some synonymous codons,

284   such as those encoding for Isoleucine (I), Cysteine (C) or Arginine (R) are located far from one

285   another, while others tend to cluster together (e.g. Proline [P] and Glutamic acid [E]).

286

287   CAMAP increases MAP prediction accuracy

288   We next compared MAP prediction capacities of CAMAPs scores to that of MAP predicted ligand

289   score (ranks as predicted by NetMHCpan4.0) and mRNA transcript expression levels. We used

290   ligand scores as predicted by NetMHCpan4.0, which was shown to possess the best predictive

291   capacities for naturally processed peptides compared to other predictive algorithms [26]. Because

292   MAP binding to the MHC molecule is essential for its presentation at the cell surface, we elected

293   to only compare hits and decoys encoding potential binders, i.e. with a minimal ligand score of

294   1% for at least one allele in the B-LCL dataset. Using a linear regression model, we compared the

295   predictive capacity of each single parameter using Matthews correlation coefficient, which

296   measures the quality of binary classifications [27]. Of note, only the predictions on the test set

297   were used to evaluate the Matthew correlation coefficient in our different models.

298   Because only potential binders were analyzed here, the mRNA expression level had the highest

299   predictive capacity, then followed by ligand scores (second) and CAMAP scores (third, Fig. 6A).

300   As expected due to the multiplicative relationship between MAP ligand score and expression levels

301   in predicting naturally processed MAPs [11], combining both variables greatly increased

302   prediction performances (Fig. 6B). Importantly, adding CAMAP scores to the regression model

303   further increased predictive performances (Fig. 6B). We next computed how many predicted

304   peptides would need to be tested to capture 1, 5 or 10% of hits in the B-LCL dataset. Results

305   presented in Table 1 show that using only NetMHCpan4.0 ligand scores (ranks) leads to a very

306   high false positive rate (FPR) at 72.1% when targeting the top 1%. Adding the expression levels

307   greatly increased prediction accuracy and decreased the FPR to 32.8% for the top 1% hits. When

308   adding CAMAP scores as a third variable, the number of peptides needed to capture 1% of hits

309   greatly decreased, resulting in a very low FPR at 1.1%. Similar trends were observed when

310   targeting 5 or 10% of hits, although with higher FPR (see Table 1). Similarly, adding CAMAP

311   scores to expression levels and ligand scores also ameliorated prediction accuracies for the two

312   other human datasets introduced above (B721.221 and PBMCs, see Supplementary Table S2).

313   These results show that combining CAMAP scores with the MAP's ligand score (ranks) and its

314   corresponding transcript expression level significantly improves prediction of MAP and facilitate

315   identification of relevant epitopes through more accurate predictions.

316

317   **Figure 6. CAMAP prediction score contributes to the prediction of MAPs.** (A) Matthews

318   correlation coefficient for MAP prediction using a single variable. (B) Matthews correlation

319   coefficient for MAP prediction using multivariable regression models. The B-LCL dataset (all

320   MAP lengths) was filtered for MAP with a minimal ligand score (rank) of 1% (NetMHCpan4.0).

321

322   **Table 1. Number of peptides needed to capture 1%, 5% or 10% of epitopes detected by mass**

323   **spectrometry.** The lower the number of peptides needed to capture the respective number of

324   epitopes, the better the performance of the prediction model. This is also illustrated by the

325   percentage of false identification (false positive rate, FPR) reported here. Peptides were rank-

326   ordered according to regression scores, for a total of 490,297 unique peptides and 8,991 hits. Of

327   note, only the maximal regression score was kept for peptides with multiple potential origins.

Regression model	1% hits (n=90)	5% hits (n=450)	10% hits (n=899)

	n	FPR	n	FPR	n	FPR
NetMHCpan4.0	322 ± 18	72.05%	2183 ± 69	79.39%	4927 ± 136	81.75%
NetMHCpan4.0 + expression	134 ± 6	32.84%	601 ± 10	25.12%	1211 ± 16	25.76%
NetMHCpan4.0 + expression + CAMAP	**91 ± 2**	**1.11%**	**524 ± 13**	**14.12%**	**1170 ± 18**	**23.16%**

328

Codon usage can modulate MAP presentation

330 To evaluate whether changing the codon arrangement in a MAP-coding sequence might directly

331 lead to modulation of MAP presentation, we generated three variants of the chicken ovalbumin

332 (OVA) protein containing the model MAP SIINFEKL [28]. One construct encoded the wild type

333 OVA (OVA-WT). For the other two constructs, we used CAMAP (trained on original human B-

334 LCL sequences; Fig. 2) to generate two OVA variants *in silico*, both encoding for the same OVA

335 protein but using different synonymous codons: one predicted to enhance SIINFEKL presentation

336 (OVA-EP), the other predicted to reduce it (OVA-RP). Accordingly, the respective CAMAP

337 scores for OVA-RP, OVA-WT and OVA-EP were: 0.03, 0.65, and 0.96 (Fig. 7A). All variants

338 encoded the same amino acid sequence but used different synonymous codons. Notably, the sole

339 difference between the three constructs were the 162 nucleotides flanking each side of the

340 SIINFEKL-coding codons (i.e. the RNA sequences coding for $OVA_{202-256}$ and $OVA_{265-319}$,

341 Supplementary Table S1 and Supplementary Figure S15).

342 **Figure 7. Codon usage in MAP-flanking mRNA sequences can influence antigen**

343 **presentation and translation efficiency.** (A) Design of the inducible Translation Reporter (iTR-

344 OVA) constructs and CAMAP scores for OVA-WT, OVA-EP and OVA-RP sequences. (B)

345 Schematic representation of possible translation events. When mRNA codon usage leads to

346 efficient (uninterrupted) translation, similar amounts of eGFP and Ametrine proteins would be

347 synthesized. When codon usage in the MAP-flanking regions enhances the frequency of

15

348   translation interruption, a lower Ametrine/eGFP ratio would be observed. (C) Kinetics of

349   SIINFEKL MAP presentation following induction of iTR-OVA constructs expression by

350   doxycycline, measured in a T-cell activation assay. To remove the influence of differential

351   expression levels on antigenic presentation and of varying proportion of transduced cells

352   between samples, T-cell activation levels were normalized to the average Ametrine fluorescence

353   intensity and to the proportion of eGFP+ cells (i.e. cells expressing the construct). (D)

354   Translation efficiency as measured by Ametrine/eGFP ratio following iTR-OVA construct

355   induction. For C and D, results are normalized over the WT sample from the same experiment

356   (n=4). Statistical differences at each time point were determined using bilateral paired Student T

357   tests. Significance for the comparison against WT are indicated with *, while comparison of EP

358   vs RP is indicated with †. N.B.: Each replicate is shown with a dot, while the line and shaded

359   area represent the average and 95% confidence interval, respectively.

360

361   Because codon usage affects translation efficiency, theoretically leading to DRiP formation

362   through premature translation arrest [20,21], we expected the variable regions of our construct to

363   affect both translation rates and SIINFEKL presentation in our variants. Therefore, each construct

364   also coded for two other proteins, eGFP and Ametrine, placed upstream and downstream of the

365   OVA coding sequence, respectively (Fig. 7A). While the Ametrine fluorescence intensity reflected

366   the translation rate of the whole construct, the ratio of Ametrine/eGFP fluorescence intensity was

367   informative regarding the translation efficiency of the whole construct. Indeed, efficient translation

368   of the full-length construct should produce equivalent quantities of Ametrine and eGFP proteins,

369   while inefficient/interrupted translation of the construct (i.e. leading to DRiP formation) should

370   decrease the Ametrine/eGFP ratio (Fig. 7B). The three protein coding sequences were separated

371   with P2A self-cleaving peptides [29], therefore allowing the co-synthesis of three separate

372   proteins, controlled by the doxycycline-inducible Tet-On promoter. Importantly, the three proteins

373   were tightly co-expressed because of the presence of only one start codon at the 5' end of the GFP

16

374   protein, as shown by the very high correlation between eGFP and Ametrine fluorescence for each

375   construct (R>0.97, see Supplementary Figure S16). As we assumed that CAMAP scores reflected

376   the probability of DRiP generation leading to increased MAP presentation, we expected the OVA-

377   RP construct to show both reduced SIINFEKL presentation and enhanced translation efficiency

378   compared to the OVA-EP and OVA-WT constructs. However, as both the OVA-EP and OVA-

379   WT have CAMAP scores above the neutral threshold of 0.5 and closer to one another (0.98 and

380   0.65, respectively) compared to the OVA-RP construct (0.03), we expected OVA-EP and OVA-

381   WT to behave more similarly.

382   We then used a SIINFEKL-H2-K^b specific T-cell activation assay [30] to measure SIINFEKL

383   presentation at the cell surface following doxycycline induction. Results for the T-cell activation

384   assay were normalized by both the Ametrine mean fluorescence intensity and the percentage of

385   transduced (eGFP+) cells in each specific sample, so that any difference in T-cell activation

386   observed between our constructs could only be ascribed to synonymous codon variants in the

387   SIINFEKL-flanking OVA codons. Two main findings emerged from our analyses. First, in

388   accordance with CAMAP predictions, variation in codon usage led to a 2.3-fold difference in

389   SIINFEKL presentation between the OVA-EP and OVA-RP variants, with OVA-WT in between

390   (Fig. 7C). Second, translation efficiency (Ametrine/eGFP ratio) was higher with OVA-RP than

391   with OVA-EP or OVA-WT, while OVA-EP showed similar translation efficiency compared to

392   ONA-WT (Fig. 7D). Hence, synonymous codon variations led to slightly divergent outcomes in

393   OVA-EP and OVA-RP: they modulated the levels of SIINFEKL presentation in both constructs,

394   but enhanced translation efficiency could only be detected for OVA-RP. These data show that

395   codon arrangement can modulate MAP presentation strength without any changes in the amino

17

396 acid sequence and support a role for translation efficiency and DRiP formation in the modulation

397 of MAP presentation.

398

399 **Discussion**

400 Our analyses of large datasets using artificial neural networks and other bioinformatics approaches

401 provide compelling evidence that codon usage regulates MAP biogenesis via both short- and long-

402 range effects. While most MAP predictive approaches focus on MAP sequences *per se*, CAMAP's

403 novelty is that it only receives the MAP-flanking mRNA sequences as input, and no information

404 on the MAP itself, thereby providing completely independent information for MAP prediction.

405 The better prediction accuracy of CAMAPs trained with original codons rather than with shuffled

406 synonyms supports the role of codon usage in modulating MAP biogenesis (Fig. 2). In addition,

407 we demonstrated that the codon-specific signal that is captured by CAMAP was independent of

408 transcript expression levels and MAP ligand scores, thereby providing complementary and non-

409 overlapping information regarding MAP presentation. Additionally, while CAMAP preferences

410 were more influential for codons located close to the MCCs (Fig. 5), the better performance of

411 CAMAP trained with longer context size pointed toward a long-range impact of codon usage on

412 MAP presentation.

413 The functional link between codon arrangement and MAP biogenesis was illustrated by our *in*

414 *vitro* analyses of SIINFEKL biogenesis, in which we were able to modulate SIINFEKL

415 presentation solely by substituting synonymous codons in mRNA regions flanking SIINFEKL

416 codons, without changing the protein sequence. While the experimental data derives from a single

417 model thus limiting the interpretability of our results, this points nonetheless to an interesting

418   mechanism that could be exploited to enhance antigenic presentation in peptide-bas4ed

419   immunotherapy (i.e. dendritic cells modified to express a specific MAP).

420   Further analyses will be needed to assess the full extent of codon arrangement's impact on both

421   classic MAPs (i.e. derived from canonical reading frames of coding sequences) and cryptic MAPs

422   (i.e. derived from non-canonical reading frames and non-coding sequences) [31,32], as well as the

423   potential contribution of codons in non-coding regions (e.g. 5'- or 3'-UTRs) on the regulation of

424   MAP presentation. However, our results show that the integration of CAMAP scores to the two

425   best predictive factors for naturally processed MAPs led to a significant increase in prediction

426   accuracy. Indeed, our regression model combining only transcript expression levels to MAP ligand

427   scores (ranks as predicted by NetMHCpan4.0), showed that a total of 134 peptides would need to

428   be tested in order to capture 1% of all presented MAPs (hits), leading to a false positive rate of

429   32.8%. In contrast, the addition of CAMAP to this model decreased the false positive rate to only

430   1.1%, leading to 90 correct identifications out of 91 MAPs tested. Although predictions were not

431   as accurate for the two other human datasets, adding CAMAP scores always resulted in improved

432   prediction accuracy. Our results therefore support the combined use of ligand scores, transcript

433   expression levels and CAMAP scores in MAP predictive algorithms. These results have important

434   practical implications for cancer immunotherapy and peptide-based vaccines, where discovery of

435   suitable target antigens remains a formidable challenge to this day [33,34].

436

437   **Materials and methods**

438   **Dataset generation**

19

439   We analyzed a previously published dataset consisting of MAPs presented on B lymphocytes by

440   a total of 33 MHC-I alleles from 18 subjects [19,22]. Since this dataset was assembled using older

441   versions of MHC-I binding prediction algorithms (i.e. using a combination of NetMHC3.4 for

442   common alleles and NetMHCcons1.1 for rare alleles), we verified that the majority of MAPs in

443   this dataset would also be predicted as binders using more recent algorithms (i.e. a rank $\leq 2.0\%$

444   using NetMHC4.0 or NetMHCpan4.0). We found an overlap of >92% between these methods (see

445   Supplementary Fig. S17), thereby validating this dataset for further analysis. In addition, we

446   reasoned that a transcript should be considered as a genuine positive or negative regarding MAP

447   biogenesis only if it was expressed in the cells. We therefore excluded from the dataset all

448   transcripts with very low expression (<1st percentile in terms of FPKM).

449   To facilitate data analysis and interpretation, we only included transcripts coding for MAPs with

450   a length of 9 amino acids, for a total of 19,656 9-mer MAPs (which represents 78% of MAPs in

451   this dataset). We then used pyGeno [23] to extract the mRNA sequences of transcripts coding for

452   these 9-mer MAPs, which constituted our source-transcripts (Fig. 1A). We next created a negative

453   (non-source) dataset from transcripts that generated no MAPs. Importantly, transcripts that

454   encoded for MAPs of any length (i.e. 8 to 11-mer) were excluded from the negative dataset. We

455   then randomly selected 98,290 non-MAP 9-mers from this negative dataset, and extracted their

456   coding sequences using pyGeno. Of note, both positive and negative datasets were derived from

457   the canonical reading frame of non-redundant transcripts.

458   We analyzed only the MAP context and excluded the MCCs *per se* from our positive (hits) and

459   negative (decoys) sequences (Fig. 1A). We limited our analyses of flanking sequences to 162

460   nucleotides (54 codons) on each side of MCCs, because longer lengths would entail the exclusion

461   of >25% of transcripts (Supplementary Fig. S18).

Short title: Codon arrangement modulates MHC-I peptide presentation.

Creation of the shuffled synonymous codon dataset

462

463 To create the shuffled synonymous codon dataset, each sequence was re-encoded by replacing

464 each codon with itself or with a random synonym according to the human transcriptome usage

465 frequencies. These frequencies were calculated using the annotations provided by *Ensembl* for the

466 human reference genome GRCh37.75. Thus, all codon-specific features differing between the

467 positive and negative datasets was removed from the shuffled datasets. Because codons were

468 replaced by their synonymous codons, the shuffled sequences directly reflected amino acid usage

469 in the positive and negative datasets.

CAMAP architecture, sequence encoding and training

470

471 The first (input) layer received either MCCs flanking regions from the hit dataset or sequences of

472 the same length contained in the decoy dataset (Fig. 1A). The second layer (Supplementary Fig.

473 4A) was a codon embedding layer similar to that introduced for a neural language model [35].

474 Embedding is a technique used in natural language processing to encode discrete words, and has

475 been shown to greatly improve performances [36]. With this technique, the user defines a fixed

476 number of dimensions in which words should be encoded. When the training starts, each word

477 receives a random vector-valued position (its embedding coordinates) in that space. The network

478 then iteratively adjusts the words' embedding vectors during the training phase and arranges them

479 in a way that optimizes the classification task. Notably, embeddings have been shown to represent

480 semantic spaces in which words of similar meanings are arranged close to each other [36]. In the

481 present work, we treated codons as words: each codon received a set of random 2D coordinates

482 that were subsequently optimized during training. The third (output) layer delivered the probability

483 that the input sequence was a MCCs flanking region (rather than a sequence from the negative

484 dataset).

21

485   CAMAPs were trained on sequences resulting from the concatenation of pre- and post-MCCs

486   regions. Before presenting sequences to our CAMAPs, we associated each codon to a unique

487   number ranging from 1 to 64 (we reserved 0 to indicate a null value) and used this encoding to

488   transform every sequence into a vector of integers representing codons. Neural networks were built

489   using the Python package Mariana [37] [https://www.github.com/tariqdaouda/Mariana]. The

490   *Embedding* layer of Mariana was used to associate each label superior to 0 to a set of 2D trainable

491   parameters; the 0 label represents a *null* (masking) embedding fixed at coordinates (0,0). As an

492   output layer, we used a *Softmax* layer with two outputs (positive / negative). Because negative

493   sequences are more numerous than positive ones, we used an oversampling strategy during

494   training. At each epoch, CAMAPs were randomly presented with the same number of positive and

495   negative sequences. All CAMAPs in this work share the same architecture (Supplementary Fig.

496   4A), number of parameters and hyper-parameter values: learning rate: 0.001; mini-batch size: 64;

497   embedding dimensions: 2; linear output without offset on the embedding layer; *Softmax* non-

498   linearity without offset on the output layer.

499   For each condition (e.g. context size), the positive and negative datasets were randomly divided

500   into three non-redundant subsets: (i) the training subsets containing 60% of the positive and

501   negative transcripts, (ii) the validation and (iii) the test subsets each containing 20% of the positive

502   and negative transcripts. Transcripts were assigned through a sequence redundancy removal

503   algorithm, thereby ensuring that no transcript was assigned to multiple subsets. We used an early

504   stopping strategy on validation sets to prevent over-fitting and reported average performances

505   computed on test sets. We trained 12 CAMAPs for each combination of conditions, each one using

506   a different random split of train/validation/test sets. To mask sequences either before or after the

507   MCCs, we masked either half with *null* value.

508 **Kullback-Leibler divergence**

509 The Kullback-Leibler (KL) divergence computes how well a given distribution is approximated

510 by another distribution. Its value can be either positive or 0, a null value indicating that the two

511 distributions are identical (see Materials and Methods for more details). Accordingly, a higher KL

512 divergence for codon distributions vs. amino acid distributions would indicate that codon

513 variations are not entirely accounted for by amino acid variations. KL divergence is not a metric,

514 as it is neither symmetric nor does it satisfy the triangle inequality. It is nevertheless an accurate

515 and most common way of comparing two probability distributions.

516 We defined the probability of having codon c at position i as a function of the number of

517 occurrences of c at position i, divided by the total number of occurrences of that same codon:

518
$$Q_{(c,y,s)}(i) = \frac{N_{c,y,s}(i)}{\sum_j N_{c,y,s}(j)}$$

519 Here Q is a probability, N is a number of occurrences, c is a codon, y is a class (positive or

520 negative), s indicates if codons have been randomized (true or false), i is a position in sequence.

521 For the remainder of the text we will use the following abbreviations:

522
$$P_c(i) = Q_{c,y=positive,s=false}(i)$$

523
$$D_c(i) = Q_{c,y=negative,s=false}(i)$$

524
$$PS_c(i) = Q_{c,y=positive,s=true}(i)$$

525
$$DS_c(i) = Q_{c,y=negative,s=true}(i)$$

526 We then used the KL divergence to compute how well P_c distributions approximate D_c

527 distributions and PS_c distributions approximate DS_c distributions.

23

528 The KL divergence was defined as:

529
$$D_{KL}(P||Q) = \sum_i P(i)\log\left(\frac{P(i)}{Q(i)}\right)$$

530 We performed this calculation for both the original and the shuffled dataset, which we then

531 compared together. If codons and amino acid distributions were equivalent, KL divergence

532 between hits and decoys would be the same for both original and shuffled sequences, and codons

533 would cluster along the diagonal.

534 **Predicting MAP presentation with linear regressions**

535 The prediction capacity of CAMAP, NetMHCpan-4.0 ligand score and transcription expression

536 (TPM) was tested in different combinations of those parameters (Ligand Score + Expression,

537 Ligand score + Expression + CAMAP score) using the *LogisticRegressionCV* function from the

538 python package *sklearn* (*sklearn.linear_mode*l, v0.22.1). In each case, the dataset containing hits

539 and decoy sequences was split into train and test datasets with a ratio of 0.7 to 0.3, respectively.

540 Values for CAMAP score, Ligand Score and TPM were each scaled to a range of 0-1 in the train

541 set using MinMaxScaler from *sklearn.preprocessing* and the same scaling model was applied to

542 the test set afterwards. Regression analysis was performed using *LogisticRegressionCV* with a 10x

543 cross-validation using the *lbfgs* solver with 1000 iterations. MCC scores were calculated using

544 *matthews_corrcoef* from *sklearn.metrics*. When a peptide had multiple sources (multiple

545 transcripts or genes), only the maximum value from its regression scores was kept.

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

547 An inducible translation reporter was generated by flanking the truncated chicken ovalbumin

548 (OVA) cDNA (amino acids 144-386) with EGFP-P2A (in 5') and P2A-Ametrine (in 3') cDNA

549 sequences. MCCs flanking contexts for the EP and RP construct were synthesized as gBlocks

24

550   (purchased from Integrated DNA Technologies). The fragments were amplified by PCR and joined

551   by Gibson assembly under a doxycycline-inducible Tet-ON promoter in a pCW backbone.

552   Synthetic variants of the OVA coding sequence were generated in silico by varying synonymous

553   codon usage in the MAP context regions (i.e. 162 nucleotides pre- and post-MCCs). Importantly,

554   the amino acid sequence was preserved between the different variants; only nucleotide sequences

555   in the MAP context (162 nucleotides on either side) differed. The sequences with the highest (EP)

556   and the lowest (RP) prediction scores were selected for further in vitro validation and swapped

557   into the iTR-OVA plasmid by Gibson assembly [38]. OVA-EP and OVA-RP sequences can be

558   found in Supplementary Table 1.

559   Important features of our inducible translation reporter construct and T cell activation assay were:

560   (i) No changes in amino acid sequence between the three variants: only co-translational events can

561   differ between the three variants, post-translational events being equivalent for the three

562   constructs; (ii) Only one start codon, at the beginning of the eGFP coding sequence: this is

563   important for the translation reporter aspect of our construct (i.e. Ametrine/eGFP ratio), to ensure

564   that translation can only start at the 5'-end of the whole construct, and not at the beginning of the

565   OVA or Ametrine coding sequences; (iii) Separation of the three proteins using P2A peptide:

566   allows the inducible synthesis of three separate proteins in a highly correlated manner; also, the

567   degradation of one protein will be independent from the others. As we hypothesized that codon

568   usage might lead to DRiP formation, we did not want the degradation of OVA-derived polypeptide

569   to induce degradation of attached eGFP or Ametrine, which would affect our translation reporter

570   assay (Ametrine/eGFP ratio); (iv) Because transcript expression level impacts MAP presentation,

571   we normalized T-cell activation results by both the number of transduced cells present in the

572   samples (% of eGFP+ cells) and the Ametrine mean fluorescence intensity of eGFP+ cells

573   (representing whole construct expression level). Because of these four features, any difference

574   between the three constructs could be ascribed solely to synonymous codon variants in the

575   SIINFEKL-flanking OVA codons.

576   **Stable cell line generation**

577   Wildtype and transduced Raw-K^b cells [39] were cultured in DMEM supplemented with 10% Fetal

578   Bovine Serum (FBS), penicillin (100 units/ml), and streptomycin (100mg/ml). B3Z cells [40] were

579   maintained in RPMI medium supplemented with 5% FBS, penicillin (100 units/ml), and

580   streptomycin (100mg/ml).

581   Lentiviral particles were produced from HEK293T cells by co-transfection of iTR-OVA WT, EP

582   or RP along with pMD2-VSVG, pMDLg/pRRE and pRSV-REV plasmids. Viral supernatants

583   were used for Raw-K^b transduction. Raw-K^b OVA-WT, Raw-K^b OVA-EP were sorted on

584   Ametrine and GFP double positive population after 24h of doxycycline treatment (1 mg/ml).

585   **T-cell activation assay**

586   Raw-K^b OVA-EP, OVA-RP and OVA-WT cells were plated at a density of 250,000 cells/well in

587   24 well-plates 24h prior to doxycycline treatment (1 mg/ml). After the corresponding treatment

588   duration, cells were harvested and fixed using PFA 1% for 10 minutes at room temperature and

589   washed using DMEM 10% FBS. Raw-K^b were then co-cultured (37°C, 5% CO_2) in triplicates with

590   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

591   Raw-K^b) in 96 well-plates. Cells were lysed for 20 minutes at room temperature using 50 µl/well

592   of lysis solution (25mM Tris-Base, 0.2 mM CDTA, 10% glycerol, 0.5% Triton X-100, 0.3mM

593   DTT; pH 7.8). 170 µl/well CPRG buffer was added (0.15mM chlorophenol red-β-d-

594   galactopyranoside (Roche), 50mM Na_2HPO_4•$7H_2O$, 35mM NaH_2PO_4•H_2O, 9mM KCl, 0.9mM

595    $MgSO_4$•$7H_2O$). β-galactosidase activity was measured at 575 nm using SpectraMax® 190

26

596 Microplate Reader (Molecular Devices). In parallel, cells were analyzed by flow cytometry using

597 a BD FACS CantoII for eGFP and Ametrine fluorescence.

598 **Data Availability**

599 The datasets analyzed for this study can be found:

600 • Human B-LCL: RNA-Seq data can be accessed on the NCBI Bioproject database

601 (http://www.ncbi.nlm.nih.gov/bioproject/; accession PRJNA286122).

602 • Human PBMC: RNA-sequencing data for human PBMC were extracted from healthy

603 donors in Zucca et al (2019) [41] and can be accessed under the GEO accession number

604 GSE106443 and GSE115259, while MAPs were extracted from Murphy et al (2017) [24].

605 • Human B721.221: The B721.221 dataset was retrieved from Abelin et al (2017) [11]; RNA

606 sequencing data can be accessed under the GEO accession number GSE93315.

607 • Murine CT26: RNA-Seq data can be accessed under the GEO accession number

608 GSE111092. Mass spectrometry data can be found on the ProteomeXchange Consortium

609 via the PRIDE partner repository (human B-LCL: PXD004023 and murine CT26:

610 PXD009065 and 10.6019/PXD009065).

611 • Murine EL4: MAP dataset was extracted from Murphy et al (2017) [24] and EL4 RNA

612 sequencing dataset was extracted from Sidoli et al (2019) [42] and can be accessed under

613 the GEO accession number GSE125384.

614 All figures were generated using R's package "ggplot2". Source code for pyGeno

615 (https://github.com/tariqdaouda/pyGeno, doi: 10.12688/f1000research.8251.2) and Mariana

616 (https://github.com/tariqdaouda/Mariana, doi: [to be provided after acceptance]) are freely

617 available online.

27

618

Acknowledgements

628

Author contributions

TD designed all computational experiments. TD and AF performed computational experiments. TD wrote pyGeno and Mariana, 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, 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 viruses 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

640 assay. PT reviewed the manuscript. Y.Bengio reviewed and contributed to the manuscript. SL and

641 CP contributed to study design, reviewed and contributed to the manuscript. All co-authors

642 reviewed the manuscript.

643 The authors declare no competing interests.

644

References

646 1. Granados DP, Laumont CM, Thibault P, Perreault C. The nature of self for T cells—a
647 systems-level perspective. Curr Opin Immunol. 2015;34: 1–8. doi:10.1016/j.coi.2014.10.012

648 2. Caron E, Espona L, Kowalewski DJ, Schuster H, Ternette N, Alpízar A, et al. An open-source
649 computational and data resource to analyze digital maps of immunopeptidomes. Chakraborty
650 AK, editor. eLife. 2015;4: e07661. doi:10.7554/eLife.07661

651 3. Davis MM, Krogsgaard M, Huse M, Huppa J, Lillemeier BF, Li Q. T Cells as a Self-
652 Referential, Sensory Organ. Annu Rev Immunol. 2007;25: 681–695.
653 doi:10.1146/annurev.immunol.24.021605.090600

654 4. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Science. 2015;348:
655 69–74. doi:10.1126/science.aaa4971

656 5. Caron E, Vincent K, Fortier M-H, Laverdure J-P, Bramoullé A, Hardy M-P, et al. The MHC
657 I immunopeptidome conveys to the cell surface an integrative view of cellular regulation.
658 Mol Syst Biol. 2011;7: 533. doi:10.1038/msb.2011.68

659 6. Neefjes J, Jongsma MLM, Paul P, Bakke O. Towards a systems understanding of MHC class
660 I and MHC class II antigen presentation. Nat Rev Immunol. 2011;11: 823–836.
661 doi:10.1038/nri3084

662 7. Bassani-Sternberg M, Gfeller D. Unsupervised HLA Peptidome Deconvolution Improves
663 Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide–HLA Interactions.
664 J Immunol. 2016;197: 2492–2499. doi:10.4049/jimmunol.1600808

665 8. Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I
666 molecules integrating information from multiple receptor and peptide length datasets.
667 Genome Med. 2016;8. doi:10.1186/s13073-016-0288-x

668 9. Tenzer S, Peters B, Bulik S, Schoor O, Lemmel C, Schatz MM, et al. Modeling the MHC
669 class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC
670 class I binding. Cell Mol Life Sci CMLS. 2005;62: 1025–1037. doi:10.1007/s00018-005-
671 4528-2

672   10. Nielsen M, Lundegaard C, Lund O, Keşmir C. The role of the proteasome in generating
673   cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal
674   cleavage. Immunogenetics. 2005;57: 33–41. doi:10.1007/s00251-005-0781-7

675   11. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass
676   Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More
677   Accurate   Epitope   Prediction.   Immunity.   2017;46:   315–326.
678   doi:10.1016/j.immuni.2017.02.007

679   12. Capietto A-H, Jhunjhunwala S, Delamarre L. Characterizing neoantigens for personalized
680   cancer immunotherapy. Curr Opin Immunol. 2017;46: 58–65. doi:10.1016/j.coi.2017.04.007

681   13. Antón LC, Yewdell JW. Translating DRiPs: MHC class I immunosurveillance of pathogens
682   and tumors. J Leukoc Biol. 2014;95: 551–562. doi:10.1189/jlb.1113599

683   14. Wei J, Kishton RJ, Angel M, Conn CS, Dalla-Venezia N, Marcel V, et al. Ribosomal Proteins
684   Regulate MHC Class I Peptide Generation for Immunosurveillance. Mol Cell. 2019;73:
685   1162-1173.e5. doi:10.1016/j.molcel.2018.12.020

686   15. Yewdell JW, Antón LC, Bennink JR. Defective ribosomal products (DRiPs): a major source
687   of antigenic peptides for MHC class I molecules? J Immunol. 1996;157: 1823–1826.

688   16. Croft NP, Smith SA, Wong YC, Tan CT, Dudek NL, Flesch IEA, et al. Kinetics of antigen
689   expression and epitope presentation during virus infection. PLoS Pathog. 2013;9: e1003129.
690   doi:10.1371/journal.ppat.1003129

691   17. Milner E, Barnea E, Beer I, Admon A. The turnover kinetics of major histocompatibility
692   complex peptides of human cancer cells. Mol Cell Proteomics MCP. 2006;5: 357–365.
693   doi:10.1074/mcp.M500241-MCP200

694   18. Hassan C, Kester MGD, de Ru AH, Hombrink P, Drijfhout JW, Nijveen H, et al. The human
695   leukocyte antigen-presented ligandome of B lymphocytes. Mol Cell Proteomics MCP.
696   2013;12: 1829–1843. doi:10.1074/mcp.M112.024810

697   19. Pearson H, Daouda T, Granados DP, Durette C, Bonneil E, Courcelles M, et al. MHC class
698   I–associated peptides derive from selective regions of the human genome. J Clin Invest.
699   2016;126: 4690–4701. doi:10.1172/JCI88590

700   20. Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, et al. A Role
701   for   Codon   Order   in   Translation   Dynamics.   Cell.   2010;141:   355–367.
702   doi:10.1016/j.cell.2010.02.036

703   21. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon
704   bias. Nat Rev Genet. 2011;12: 32–42. doi:10.1038/nrg2899

705   22. Granados DP, Rodenbrock A, Laverdure J-P, Côté C, Caron-Lizotte O, Carli C, et al.
706   Proteogenomic-based discovery of minor histocompatibility antigens with suitable features

707   for immunotherapy of hematologic cancers. Leukemia. 2016;30: 1344–1354.
708   doi:10.1038/leu.2016.22

709 23. Daouda T, Perreault C, Lemieux S. pyGeno: A Python package for precision medicine and
710   proteogenomics. F1000Research. 2016;5: 381. doi:10.12688/f1000research.8251.2

711 24. Murphy JP, Konda P, Kowalewski DJ, Schuster H, Clements D, Kim Y, et al. MHC-I Ligand
712   Discovery Using Targeted Database Searches of Mass Spectrometry Data: Implications for
713   T-Cell   Immunotherapies.   J   Proteome   Res.   2017;16:   1806–1816.
714   doi:10.1021/acs.jproteome.6b00971

715 25. Laumont CM, Vincent K, Hesnard L, Audemard É, Bonneil É, Laverdure J-P, et al.
716   Noncoding regions are the main source of targetable tumor-specific antigens. Sci Transl Med.
717   2018;10: eaau5516. doi:10.1126/scitranslmed.aau5516

718 26. Paul S, Croft NP, Purcell AW, Tscharke DC, Sette A, Nielsen M, et al. Benchmarking
719   predictions of MHC class I restricted T cell epitopes in a comprehensively studied model
720   system. PLoS Comput Biol. 2020;16: e1007757. doi:10.1371/journal.pcbi.1007757

721 27. Powers D. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness,
722   Markedness & Correlation. Mach Learn Technol. 2008;2.

723 28. Dersh D, Yewdell JW, Wei J. A SIINFEKL-Based System to Measure MHC Class I Antigen
724   Presentation Efficiency and Kinetics. Methods Mol Biol Clifton NJ. 2019;1988: 109–122.
725   doi:10.1007/978-1-4939-9450-2_9

726 29. Kim JH, Lee S-R, Li L-H, Park H-J, Park J-H, Lee KY, et al. High Cleavage Efficiency of a
727   2A Peptide Derived from Porcine Teschovirus-1 in Human Cell Lines, Zebrafish and Mice.
728   PLOS ONE. 2011;6: e18556. doi:10.1371/journal.pone.0018556

729 30. Shastri N, Gonzalez F. Endogenous generation and presentation of the ovalbumin peptide/Kb
730   complex to T cells. J Immunol Baltim Md 1950. 1993;150: 2724–2736.

731 31. Laumont CM, Daouda T, Laverdure J-P, Bonneil É, Caron-Lizotte O, Hardy M-P, et al.
732   Global proteogenomic analysis of human MHC class I-associated peptides derived from non-
733   canonical reading frames. Nat Commun. 2016;7: 10238. doi:10.1038/ncomms10238

734 32. Zanker DJ, Oveissi S, Tscharke DC, Duan M, Wan S, Zhang X, et al. Influenza A Virus
735   Infection Induces Viral and Cellular Defective Ribosomal Products Encoded by Alternative
736   Reading   Frames.   J   Immunol   Baltim   Md   1950.   2019;202:   3370–3380.
737   doi:10.4049/jimmunol.1900070

738 33. Ehx G, Perreault C. Discovery and characterization of actionable tumor antigens. Genome
739   Med. 2019;11: 29. doi:10.1186/s13073-019-0642-x

740 34. Sette A, Fikes J. Epitope-based vaccines: an update on epitope identification, vaccine design
741   and delivery. Curr Opin Immunol. 2003;15: 461–470. doi:10.1016/s0952-7915(03)00083-9

742   35. Bengio Y, Ducharme R, Vincent P, Jauvin C. A Neural Probabilistic Language Model. J
743   Mach Learn Res. 2003;3: 1137–1155.

744   36. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521: 436–444.
745   doi:10.1038/nature14539

746   37. Daouda T. Mariana: The Cutest Deep learning Framework. 2015. Available:
747   https://www.github.com/tariqdaouda/Mariana

748   38. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison Iii CA, Smith HO. Enzymatic
749   assembly of DNA molecules up to several hundred kilobases. Nat Methods. 2009;6: 343–
750   345. doi:10.1038/nmeth.1318

751   39. Bell C, English L, Boulais J, Chemali M, Caron-Lizotte O, Desjardins M, et al. Quantitative
752   Proteomics Reveals the Induction of Mitophagy in Tumor Necrosis Factor-α-activated
753   (TNFα) Macrophages. Mol Cell Proteomics MCP. 2013;12: 2394–2407.
754   doi:10.1074/mcp.M112.025775

755   40. Karttunen J, Sanderson S, Shastri N. Detection of rare antigen-presenting cells by the lacZ
756   T-cell activation assay suggests an expression cloning strategy for T-cell antigens. Proc Natl
757   Acad Sci U S A. 1992;89: 6020–6024.

758   41. Zucca S, Gagliardi S, Pandini C, Diamanti L, Bordoni M, Sproviero D, et al. RNA-Seq
759   profiling in peripheral blood mononuclear cells of amyotrophic lateral sclerosis patients and
760   controls. Sci Data. 2019;6: 190006. doi:10.1038/sdata.2019.6

761   42. Sidoli S, Lopes M, Lund PJ, Goldman N, Fasolino M, Coradin M, et al. A mass spectrometry-
762   based assay using metabolic labeling to rapidly monitor chromatin accessibility of modified
763   histone proteins. Sci Rep. 2019;9: 13613. doi:10.1038/s41598-019-49894-4

764

765

Supporting information captions

Supplementary Figures

Supplementary Figure S1. Codon distribution in the shuffled datasets more closely resembles that of amino acids, compared to the original datasets. (A) Pearson correlation (R^2) factors and (b) Kullback-Leibler (KL) divergence between positional distribution of codons and their corresponding amino acid in the shuffled (y axis) VS original (x axis) datasets. For all codons, the shuffled dataset showed greater correlations (A) and smaller KL divergence to their respective amino acid distributions than the original datasets ($p < 1 \times 10^{-8}$, assessed using unilateral paired Student T test).

Supplementary Figure S2. Distribution of Pearson's correlation factors calculated between codons and amino acids positional distributions in the original (green) and shuffled (coral) datasets. 92% of codons in the shuffled dataset reflecting the amino acids distribution with a R2 > 0.95, compared to only 69% in the original dataset ($p < 5 \times 10^{-5}$).

Supplementary Figure S3. Distribution of amino acid and codon usage per position in the original VS shuffled datasets. (A) Alanine – A. (B) Cysteine – C. (C) Aspartic acid – D. (D) Glutamic acid – E. (E) Phenylalanine – F. (F) Glycine – G. (G) Histidine – H. (H) Isoleucine – I. (I) Lysine – K. (J) Leucine – L. (K) Asparagine – N. (L) Proline – P. (M) Glutamine – Q. (N) Arginine – R. (O) Serine – S. (P) Threonine – T. (Q) Valine – V. (R) Tyrosine – Y.

Supplementary Figure S4. CAMAP architecture and detailed predictions. (A) Architecture of the ANN used in this work. (B) Results for the AUC on all train, validation and test subsets. Grey areas represent the 95% confidence intervals. (C) Distributions of output probabilities of CAMAPs used to calculate correlations in Supplementary Figure S5.

33

792 **Supplementary Figure S5**. Correlation between CAMAP prediction score trained only with pre-

793 MCC or post-MCC sequences. For each sequence in the test set we calculated the average

794 prediction score given by CAMAPs in each condition, and calculated the Pearson correlation using

795 the R software. Densities were calculated on all points and drawn using ggplot2. Only a random

796 subset of the points is represented in the figures to limit their size.

797

798 **Supplementary Figure S6.** Absence of correlation between CAMAP prediction score and

799 transcript expression levels in 4 individual B-LCL samples (each derived from a different subject).

800

801 **Supplementary Figure S7. Training of CAMAP on dataset selected to reflect positive**

802 **dataset's distribution in expression levels.** (A) Distribution of transcript expression levels for

803 normal datasets (related to Figure 2) and the dataset used here to retrain CAMAP. As shown in

804 this figure, the decoy dataset was selected to mirror the distribution of transcript expression level

805 in the hit dataset. (B) CAMAP performance (measured by the AUC) when trained using the decoy

806 dataset that mirrors the transcript expression levels of the hit dataset. Significance was assessed

807 using bilateral paired Student T test ($p = 5.36 \times 10^{-7}$).

808

809 **Supplementary Figure S8. Absence of correlation between CAMAP prediction score and**

810 **binding affinities for individual alleles for decoys (A) and hits (B).**

811

812 **Supplementary Figure S9. Training of CAMAP on dataset selected to reflect positive**

813 **dataset's distribution in binding affinities.** (A) Distribution of binding affinities for normal

814 datasets (related to Figure 2) and the corrected dataset used to retrain CAMAP. As shown in this

815 figure, the decoy dataset was selected to mirror the distribution of binding affinities in the hit

816 dataset. (B) CAMAP performance (measured by the AUC) when trained using the decoy dataset

817 that mirrors the binding affinities of the hit dataset. Significance was assessed using bilateral paired

818 Student T test ($p = 1.21 \times 10^{-9}$).

34

819

820 **Supplementary Figure S10. Evaluation of homology in hit dataset and its impact of CAMAP**

821 **performance.** (A) Proportion of unique MAPs that can be ascribed to a single origin, 2-3, or >10

822 possible origins. (B) Proportion of entries in the hit dataset that encode for MAPs with a single

823 origin, 2-3, 4-10 or >10 possible origins

824

825 **Supplementary Figure S11. Gene families overrepresented in hits with >3 possible origins.**

826

827 **Supplementary Figure S12. CAMAP performance (AUC) when trained using either all hits**

828 **(left), hits with 10 possible origins or less (center) or hits with 3 possible origins or less (right).**

829

830 **Supplementary Figure S13**. Kullback-Leibler divergence between hit and decoy datasets in

831 original codon (y-axis) or shuffled synonymous codon sequences (x-axis). Shuffled sequences

832 represent amino acid usage, as codon-specific information are removed with synonymous codon

833 shuffling.

834

835 **Supplementary Figure S14. Preferences per position for all codons for CAMAP trained with**

836 **original sequences.** See Materials and Methods for more details.

837

838 **Supplementary Figure S15.** OVA-construct alignment, showing point mutations (red lines) in

839 the mRNA sequences flanking the SIINFEKL MCC. (A) Comparison of the OVA-EP nucleotide

840 sequence to the wildtype OVA sequence. The OVA-EP and OVA-WT sequences have 93.3%

841 nucleotide identity for a total of 78 modified nucleotides. (B) Comparison of the OVA-RP

842 nucleotide sequence to the wildtype OVA sequence. The OVA-EP and OVA-WT sequences have

843 92.6% nucleotide identity, for a total of 86 modified nucleotides. Mutations, shown in red, are

844   located only in the 162 nucleotide regions flanking the SIINFEKL coding codons. Of note, the

845   SIINFEKL coding codons (nucleotides 772-799) were not modified between the 3 constructs.

846

847   **Supplementary Figure S16.** Correlations between eGFP and Ametrine fluorescence intensity at

848   the single cell level. Single cell eGFP and Ametrine fluorescence intensities measured at 10 hours

849   post-induction are shown for the OVA-WT (A), OVA-EP (B) and OVA-RP (C) constructs. N.B.:

850   only transduced cells are shown (eGFP+ cells).

851

852   **Supplementary Figure S17.** Validation of MHC-I associated peptides (MAP) dataset from

853   Pearson H. *et al*. (2016) using the new versions of MAP binding affinity prediction algorithm

854   NetMHC4.0 (A) and NetMHCpan4.0 (B).

855

856   **Supplementary Figure S18.** Percentage of transcript ineligibility as a function of context size.

857   Transcript length corresponds to $C \times 2 + 27$, where C is the context size in nucleotides and 27 the

858   length of the MCCs. Related to Figure 1A.

859

860

861   **Supplementary Tables**

862   **Supplementary Table S1. Nucleotide sequences of the EP and RP constructs.** SIINFEKL

863   MCCs are shown in bold, while the variant regions (pre- and post-MCCs flanking sequences,

864   context size of 162-nucleotides) are in blue and italics. Related to Fig. 7.

865

866   **Supplementary Table S2. Number of peptides needed to capture 1%, 5 and 10% of epitopes**

867   **detected by mass spectrometry in B721.221 and PBMC cell lines.** The lower the number of

868   peptides needed to capture the respective number of epitopes, the better the performance of the

869   prediction model. This is also illustrated by the percentage of false identification (false positive

870   rate, FPR) reported here. Peptides were rank-ordered according to regression scores. Of note, only

871   the maximal regression score was kept for peptides with multiple potential origins.

872

873