EPIP: MHC-I epitope prediction integrating mass spectrometry derived motifs and tissue-specific expression profiles

Weipeng Hu[1,2], Si Qiu[1,2], Youping Li[1,2,3], Xinxin Lin[1,2,3], Le Zhang[1,2,3], Haitao Xiang[1], Xing Han[1,2], Sitao Zhu[2], Lei Chen[1], Sha Li[1,2], Wenhui Li[1,2], Zhe Ren[1], GuiXue Hou[1], Zhilong Lin[1], Jianliang Lu[1], Geng Liu[1,2], Bo Li@[1,2,4], Leo J Lee@[1,2,5]

1. BGI-Shenzhen, Shenzhen 518083, China
2. BGI-GenoImmune, Wuhan 4300794, China
3. BGI Education Center, University of Chinese Academy of Sciences, Shenzhen 518083, China
4. Guangdong Enterprise Key Laboratory of Human Disease Genomics, Shenzhen 518083, China
5. Department of Electrical and Computer Engineering, Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3G4, Canada

Correspondence: ljlee@psi.toronto.edu, libo@genomics.cn

Abstract

Background: Accurate prediction of epitopes presented by human leukocyte antigen (HLA) is crucial for personalized cancer immunotherapies targeting T cell epitopes. Mass spectrometry (MS) profiling of eluted HLA ligands, which provides high-throughput measurements of HLA associated peptides *in vivo*, can be used to faithfully model the presentation of epitopes on the cell surface. In addition, gene expression profiles measured by RNA-seq data in a specific cell/tissue type can significantly improve the performance of epitope presentation prediction. However, although large amount of high-quality MS data of HLA-bound peptides is being generated in recent years, few provide matching RNA-seq data, which makes incorporating gene expression into epitope prediction difficult.

Methods: We collected publicly available HLA peptidome and matching RNA-seq data of 34 cell lines derived from various sources. We built position score specific matrixes (PSSMs) for 21 HLA-I alleles based on these MS data, then used logistic regression (LR) to model the relationship among PSSM score, gene expression and peptide length to predict whether a peptide could be presented in each of the cell line. We further built a universal LR model, termed Epitope Presentation Integrated Prediction (EPIP), based on more than 180,000 unique HLA ligands collected from public sources and ~3,000 HLA ligands generated by ourselves, to predict epitope presentation for 66 common HLA-I alleles.

Results: When evaluating EPIP on large, independent HLA eluted ligand datasets, it performed substantially better than other popular methods, including MixMHCpred (v2.0), NetMHCpan (v4.0), and MHCflurry (v1.2.2), with an average 0.1% positive predictive value (PPV) of 52.01%, compared to 37.24%, 36.96%, 24.90% and 23.76% achieved by MixMHCpred,

NetMHCpan-4.0 (EL), NetMHCpan-4.0 (BA) and MHCflurry, respectively. It is also comparable to EDGE, a recent deep learning-based model that is not publicly available, on predicting epitope presentation and selecting immunogenic cancer neoantigens. However, the simplicity and flexibility of EPIP makes it easier to be applied in diverse situations, and we demonstrated this by generating MS data for the HCC4006 cell line and adding the support of HLA-A*33:03 to EPIP. EPIP is publicly available as a web tool < http://epip.genomics.cn/>.

Conclusions: we have developed an easy to use, publicly available epitope prediction tool, EPIP, that incorporates information from both MS and RNA-seq data, and demonstrated its superior performance over existing public methods.

Keywords: T cell epitope, MHC-I, HLA peptidome, Eluted ligand, Mass spectrometry, RNA-seq, Neoantigen, Cancer immunotherapy

Background

T cell epitopes, which are short peptides presented by major histocompatibility complex (MHC) molecules on the cell surface and recognized by T-cell receptors (TCRs), lie at the heart of the human immune system that can remove infected and malignant cells. With the rapid progress of cancer immunotherapies, the mechanistic understanding, computational prediction and clinical manipulation of T cell epitopes have gathered renewed and widespread interests [1-4]. It is well known that peptide binding to MHC molecules is the most selective step in the antigen presentation pathway [5], and intensive computational efforts have been exerted to predict this binding process based on peptide sequences (allele-specific models) or peptide and the corresponding human leukocyte antigen (HLA) sequences (pan-specific models). Traditionally, these models are trained on data accumulated from various types of *in vitro* binding assays, which can provide quantitative binding affinity measurements between pre-selected peptides and HLA-I or HLA-II molecules. However, the requirement of synthesizing peptides beforehand limited the ability of these assays to perform unbiased, high-throughput screening of the vast peptide space, and it also did not account for the antigen-loading process *in vivo* [6]. The recent advance of mass spectrometry (MS) profiling of HLA ligands, mainly for HLA-I alleles at the moment, overcomes most of these limitations by providing *in vivo* measurements of peptides presented by MHC molecules on the cell surface in a high-throughput manner, although the resulting measurements are only qualitative (binary).

The value of MS data as contributing to the ultimate prediction of peptide immunogenicity has been increasingly recognized [6, 7], and as a result, a number of recently developed peptide-MHC binding prediction methods incorporated some form of MS data, in addition to the

83 traditional binding assay data [5-8]. For example, NetMHCpan-4.0 trained an ensemble of two
84 hidden layer neural networks on both affinity and MS data, and was able to make predictions on
85 both types of outputs. MHCflurry adopted a deep learning framework containing both locally
86 connected and fully connected layers and was mainly trained on affinity data, but used MS data
87 at the model selection stage, while MS data may play an even bigger role in its future releases.
88 Other methods chose to rely on MS data alone [7, 9, 10], such as MixMHCpred [9, 10], which
89 built allele-specific position score specific matrixes (PSSMs) after deconvolving mixed-allele
90 MS data, and the recent EDGE method [7], which used a comprehensive, deep learning model
91 that pooled together large sets of mono-allele and mixed-allele MS data as well as matching
92 RNA-seq data. With the increasing amount of MS data becoming available, it is our view that it
93 might be better to build separate models for binding affinity (based on binding assay data) and
94 epitope presentation (based on MS data) since the underlying data generation processes are
95 different. These two related pieces of information could be further combined at a later stage to
96 contribute to the prediction of immunogenicity. The presentation of a peptide by MHC not only
97 depends on information contained within the protein sequences but is also modulated by other
98 factors such as protein abundance, localization and turnover [11, 12]. Among these, the
99 corresponding gene expression levels can be handily profiled by RNA-seq and have been shown
100 to significantly improve prediction accuracy, first by the MS IntrinsicEC method on mono-allele
101 cell lines [12], then more comprehensively by EDGE [7]. Unfortunately, both methods are not
102 publicly available.
103
104 We set out to develop an effective, flexible and publicly available HLA-I epitope prediction
105 method that takes advantage of large sets of MS and RNA-seq data. We first collected public MS

106 and matching RNA-seq datasets for 21 HLA-I alleles, which frequently appear in the European
107 or Asian population. By carefully examining the contributions of PSSM scores, gene expression
108 levels and peptide length features to epitope presentation, we were able to effectively model
109 these variables with a universal logistic regression (LR) model. We then trained such a LR
110 model, termed Epitope Presentation Integrated Prediction (EPIP), to support 66 common HLA-I
111 alleles. In the remainder of this article, we first describe our design rationale in detail, followed
112 by comprehensive evaluation of EPIP on large, independent MS and immunogenic datasets as
113 well as a user case showing its extendibility, before concluding with discussions on future
114 improvements and research directions. We believe EPIP is a timely contribution to the research
115 community of cancer immunotherapies and beyond.
116

117 ## Methods

118

119 ### Collection of MS and RNA-seq data

120

121 We collected matching MS and RNA-seq data from 16 mono-allele HLA-A and HLA-B cell
122 lines [12] and 18 mixed-allele cell lines [13] to build the EPIP model. We also collected MS data
123 of 15 mono-allele HLA-C cell lines from Di Marco et al [14], deconvolved MS data of 26 more
124 HLA alleles from SysteMHC [15] and a small portion of mixed-allele MS data released with
125 EDGE [7] to expand the allele coverage of EPIP. To perform independent evaluation of EPIP,
126 we further downloaded MS data from Trolle et al [16] and Bassani-Sternberg et al [11], in
127 addition to test data reserved from Pearson et al [13] and Bulik-Sullivan et al [7]. Detailed
128 information and usage of all collected MS data are provided in Table S1.

129

130 **MS profiling of HCC4006 HLA-I peptidome**

131

132 HLA-I peptidome samples for HCC4006 were prepared according to Bassani-Sternberg et al
133 [11]. They were then analyzed by LC-MS/MS to obtain peptide MS spectra, and full details of
134 the experimental procedure are provided in the Supplementary Material. We employed the MS-
135 GF+ search engine version 2018.07.17 [17] to search the peak lists against the UniProt databases
136 (161,521 entries for human as of December 2017) and a file containing 245 frequently observed
137 contaminants such as human keratins, bovine serum proteins, and proteases. N-terminal
138 acetylation (42.01 Da) and methionine oxidation (15.99 Da) were set as variable modifications.
139 The enzyme specificity was set as unspecific. The initial allowed mass deviation of the precursor
140 ion deviation was set to 10 ppm. Possible peptide lengths were restricted to 8 to 15. Possible
141 precursor charges were restricted to 2 to 5. Range of allowed isotope peak errors was restricted
142 to -1 to 2. Percolator version 3.02.0 was applied as a post-processing step for result filtering [18].
143 The *q* value cutoff was set to 0.01. From the "pout.tab" output file produced by percolator, hits to
144 the contaminants were eliminated.

145

146 **Data and features of EPIP models**

147

148 From the matching MS and RNA-seq data collected from public sources [12, 13], we first built
149 separate LR models, EPIP_s, to predict epitope presentation for 21 alleles and tissues. This
150 included all 16 mono-allele data profiled in Abelin et al [12], as well as five more alleles (HLA-
151 A*11:01, HLA-A*32:01, HLA-B*15:01, HLA-B*40:01, HLA-B*07:02) selected from Pearson

152 et al [13], which frequently appear in the Chinese population [19]. All RNA-seq datasets were
153 downloaded and re-processed with the same pipeline to ensure consistency. Specifically, raw
154 fastq files were cleaned by fastp (version 0.18.0) [20]and aligned to human genome assembly
155 GRCh38 and its associated transcriptome (Ensembl Release 92) with STAR (version 2.5.3a)
156 [21]; gene expression (transcripts per million or TPM) was quantified by RSEM (version 1.3.0)
157 [22] based on the alignment.
158
159 EPIP_s uses three features to model epitope presentation: peptide motifs as modelled by PSSM,
160 peptide lengths and the corresponding gene expression levels. For mono-allele MS data , we built
161 PSSMs directly, as described in Liu et al [23]. For mixed-allele MS data, we first performed
162 motif deconvolution by GibbsCluster [24] before building PSSMs. Peptides from deconvolved
163 motifs that matched the HLA motifs of interest generated from SysteMHC data were retrieved.
164 For each allele, length-specific PSSMs were built for those that have at least 100 peptides, up to
165 the most abundant three lengths (9-11-mers or 8-10-mers, depending on the allele, details in
166 Table S1), and this allowed us to make use of 74%-96% of MS profiled peptides across different
167 alleles. To fully utilize the training data and avoid overfitting, we used the stacking method [25]
168 to build PSSM (see supplementary materials, Fig S1). For each HLA allele, we also obtained an
169 empirical probability mass function (pmf) to depict the length distribution of 8-15-mers based on
170 the allele-specific MS data, and the peptide length feature is the corresponding probability in the
171 empirical pmf of peptide length distribution. Expression values of a peptide was calculated by
172 summing up the expression levels of all transcripts containing the peptide (in TPM) before
173 transformed by log2(TPM+1). For the mono-allele cell lines in Abelin et al [12], transcript
174 expression levels were taken as the average of the four cell lines with RNA-seq data available;

175 similarly, for the mixed-allele cell lines in Pearson et al [13] , transcript expression levels were
176 averaged over the 10 cell lines with RNA-seq data available. All three features were further
177 normalized to be zero mean, unit variance Gaussian before using as inputs to a LR model with
178 weak L2 regularization implemented in scikit-learn 0.19.0 [26].
179
180 MS profiled peptides can only provide a positive training set for our model. To generate the
181 negative set, we randomly sampled the human peptidome to obtain peptides of the matching
182 length and remove those that overlap with the positive set. Adopting the same strategy as in
183 NetMHCpan 4.0, we used more negative peptides than positive ones to train EPIP_s. For
184 example, if the positive set contains 9-11-mers with the most abundant 9-mers having a size of n,
185 we then generate a negative set with $10n$ peptides for each length ($30n$ in total).
186
187 To build the full EPIP model for the above 21 alleles from different cell lines, we pooled all the
188 PSSM scores, peptide length distributions and corresponding gene expression levels together to
189 train a single LR model with the same training data as EPIP_s. Five-fold cross validation was
190 used to evaluate EPIP during training and to compare EPIP and EPIP_s, while the full training
191 set was used to build the final EPIP model and it was evaluated on independent test data. To
192 expand the number of HLA alleles that EPIP can support, mono-allele HLA-C MS data from Di
193 Marco et al [14], deconvolved MS data from SysteMHC [15] and some mixed-allele MS data
194 released by EDGE [7] were used to learn PSSMs for these additional HLA alleles. For the mono-
195 allele HLA-C MS data from Di Marco et al, we deconvolved the unfiltered list of peptides to
196 remove the B35 and C04 motifs through GibbsCluster [24] before using them to train PSSMs.

197 We further generated additional MS-data from the HCC4006 cell line so that EPIP can support
198 HLA-A*33:03, a high frequency allele with 8.03% prevalence in the Chinese population [19].
199
200 **Immunogenicity evaluation**
201
202 The immunogenic dataset used in this paper is the same as in Bulik-Sullivan et al. [7], which
203 contains 2,023 assayed mutations from 17 patients with annotated HLA information from four
204 previous studies [27-30]. In three of the four studies [27, 28, 30] , mutations were tested using
205 25mer tandem minigene (TMG) assays. For each mutation in these datasets, we used EPIP,
206 MixMHCpred 2.0, NetMHCpan4.0-EL, NetMHCpan4.0-BA, MHCflurry 1.2.2 to make
207 predictions with respect to the HLA alleles of each patient for overlapping 8-11mers that harbor
208 the mutation, and the score of the mutation is taken as the best score among these peptides
209 (EPIP: highest presentation probability; MixMHCpred 2.0: highest predicted score;
210 NetMHCpan4.0-EL and NetMHCpn4.0-BA: lowest rank score; MHCflurry 1.2.2: minimum
211 binding affinity). We then ranked the mutations according to the best scores from each software.
212 For the fourth study [29] , we ranked mutations by taking the best scores from each software
213 across all mutation-spanning peptides tested in the tetramer assays. To compare EPIP with
214 EDGE on predicting immunogenic peptides, we excluded the peptides whose immunogenicity
215 was undetermined as described in Bulik-Sullivan et al. We retrieved a total of 31 immunogenic
216 epitopes from Tran et al., 216 Gros et al. and Stronen et al., which were not provided by Bulik-
217 Sullivan et al and slightly different from the number 29 reported in Bulik-Sullivan et al[1]. Since

[1] The number of immunogenic epitopes is the same as in a more recent publication [31]retrieving data from the same source.

218 the epitope prediction scores by EDGE was not available, we directly compared its immunogenic
219 prediction results with EPIP.
220

221 # Results

222

223 **Construction and analysis of EPIP models**

224

225 After collecting matching MS and RNA-seq data from public sources, we first developed
226 separate LR models for each of the 21 tissues and alleles (termed EPIP_s) to predict epitope
227 presentation by combining motif, peptide length and gene expression features. Full details of the
228 datasets and features are described in Methods, and the overall process is sketched in Figure 1A.
229 According to previous studies [12, 32-34], about one in a thousand peptides generated from the
230 human proteome can be presented by MHC. Therefore, we further generated 999-fold decoy
231 peptides that are not in the training set to evaluate our models. Similar to Abelin et al [12], the
232 main evaluation criterion used throughout the paper is 0.1% positive predictive value (PPV),
233 which measures the proportion of truly presented peptides in the top-scoring 0.1% ones found by
234 our model. The 0.1% PPV is both a more realistic measure for immunopeptidome prediction and
235 a more stringent one than many other popular metrics, such as ROC-AUC and F1 scores, for this
236 highly biased classification task.
237

238 We used PSSM, a simple model that can learn patterns in biological sequences, to model peptide
239 motifs since it has previously been widely used in modelling MHC binding peptides [35-37], is
240 easy to visualize, and seems to have adequate representational power. It is also a data efficient

model due to its simplicity. To show this, we used subsampling to build PSSM models for different lengths of peptides to predict epitope presentation and tested their performances (0.1% PPV) on various alleles. As shown in Figure 1B, the performances of PSSM models reached saturation when having 100-200 peptides in the training set. Therefore, we selected 100 as a cutoff to build length-specific PSSM models for up to three most abundant lengths in each allele (9-11-mers or 8-10-mers, depending on the allele, details in Table S1), and the model would also be restricted to make predictions on these peptide lengths.

Although it is reasonable to assume that the relationship among motif, peptide length and gene expression is a general one across different tissues (cell lines) and alleles so that data from these different sources could be pooled together to train a general LR model (as outlined in Figure 1C), building EPIP_s first allowed us to rigorously analyze and quantify this. We first examined the distribution of optimized model parameters (weights and biases) among these 21 EPIP_s models. As shown in Figure 2A, the corresponding model parameters distributed quite tightly. Since the features have been normalized prior to model training, the relative magnitudes of model parameters also displayed the contribution of each feature, where PSSM played the most important role, followed by expression levels and peptide lengths (the average weights of PSSM, expression level and peptide lengths are 4.30, 1.37, 0.37, respectively). Furthermore, when displaying these 21 sets of weights and biases as hive plots (Figure 2B), they share very similar inter-relationships and seem to be mostly differ by a scaling factor, which is likely caused by somewhat different influences of L2 regularization in each model. This provided us with a strong motivation to build a universal LR model, EPIP, to account for the contribution of PSSM,

263 peptide length and gene expression across these 21 tissues and alleles, to compare against the
264 EPIP_s models, which will be described next.
265
266 **EPIP and EPIP_s achieved similar performance on epitope presentation prediction**
267
268 Based on five-fold cross validation, we performed comprehensive evaluations between EPIP and
269 EPIP_s in terms of 0.1% PPV. As shown in Figure 3A, when evaluating across different alleles
270 originating from the same cell line, the performances of EPIP and EPIP_s are quite similar. The
271 mean 0.1% PPV of EPIP is slightly better (0.5741 vs 0.5692), likely due to the more training
272 data available to EPIP, but such a difference is not statistically significant (P = 0.1385, paired t-
273 test), and both substantially outperformed predictions based on PSSM scores alone (the mean
274 0.1% PPV of PSSM is 0.4421).We also performed similar analyses for the same HLA alleles
275 across different cell lines with diverse tissue origins and thus diverse gene expression profiles
276 (Figure 3B). When comparing 0.1% PPV of EPIP and EPIP_s, EPIP is not worse than EPIP_s
277 overall, and even notably better for some cell lines (Figure 3C-F), also likely due to the larger
278 training set of EPIP. Based on these comparisons, we concluded that developing a universal LR
279 model, EPIP, to account for the relationship among these three variables is a valid approach. This
280 seemly simple strategy has substantial practical values, since the EPIP model trained on the
281 selected 21 HLA alleles and cell lines from two different tissue origins can now be
282 straightforwardly extended to previously unseen HLA alleles and tissues without adjusting the
283 model parameters of LR, or requiring extra RNA-seq data for training. We did so for 44 more
284 HLA alleles by incorporating public mono-allele [14] and deconvolved mixed allele [7-15] MS

285 data. We also generated our own MS data to demonstrate as a user case, after evaluating EPIP
286 against other methods.
287
288 **EPIP significantly outperforms other existing methods**
289
290 We evaluated EPIP against MixMHCpred (v2.0), NetMHCpan-4.0 (EL), NetMHCpan-4.0 (BA),
291 and MHCflurry (v1.2.2) on independent validation MS datasets collected from public sources,
292 which contained 29 MS datasets and was not used in the training of EPIP. EPIP significantly
293 outperformed other methods with mean 0.1% PPV of 52.01%, compared to 37.24%, 36.96%,
294 24.90% and 23.76% achieved by MixMHCpred, NetMHCpan(EL), NetMHCpan (BA) and
295 MHCflurry, with highly significant p-values (P = 2.819e-11, 3.960e-11, 1.738e-14, 6.826e-14,
296 respectively). Since other methods in the comparison do not make use of gene expression values,
297 we further combined different gene expression value cutoffs with their predictions. The
298 advantage of EPIP persists and the differences are still highly significant, as shown in Figure 4A.
299 Full comparisons for each validation dataset are provided in Figure S2. The recently developed
300 EDGE method has also been shown to excel at epitope presentation prediction comparing to
301 existing methods, but we were not able to do a fair comparison since it is not publicly available.
302 Nevertheless, we will discuss this further with some qualitative comparisons in the Discussion
303 section to show that the performances of EPIP and EDGE are comparable. Overall, the results on
304 independent MS validation sets showed the clear advantage of EPIP in predicting epitope
305 presentation.
306
307 **EPIP is valuable to immunogenicity prediction**

308

309 For cancer immunotherapies, as well as other clinical applications, the ultimate task is usually to
310 predict the immunogenicity of an epitope. While this remains an elusive goal in general [38], it
311 has recently been shown that accurate prediction of epitope presentation can significantly
312 improve immunogenicity predictions [7]. Therefore, we downloaded the same dataset as used in
313 Bulik-Sullivan et al [7], which was collected from four previous publications [27-30] that
314 consisted of 26 SNVs and 31 neoantigens with pre-existing T-cell responses among 2,023
315 assayed single nucleotide variants (SNVs) from 17 patients, to evaluate our method against
316 others. We first asked if the increased accuracy of EPIP over other publicly available methods on
317 epitope presentation could translate into better immunogenicity prediction. To quantify this, we
318 plotted the precision recall curves (PRCs) by ranking all SNVs scored by different methods,
319 where the score of a SNV is the maximum score of all possible epitopes generated by the SNV.
320 EPIP's area under PRC (AUPRC) was more than twice that of other methods, and kept the
321 advantages when we applied expression value threshold of 1 or 2 TPM to other methods'
322 predictions (Figure 4B and Figure S3). We also compared EPIP with EDGE at the level of
323 ranking mutations and minimal epitopes (the recognized 8-11-mers overlapping the mutation) as
324 originally done in Bulik-Sullivan et al [7]. EPIP is very comparably to EDGE on these two
325 evaluations, especially for the top 10 and top 20 ranked mutations and epitopes. When
326 prioritizing the mutations, the number of immunogenic SNVs ranked in the top 20, 10, and 5
327 were 18 (69.23%), 16 (61.54%) and 11 (42.31%) for EPIP, and 19 (73.08%), 18 (69.23%) and 12
328 (46.15%) for EDGE (Figure 4C). When prioritizing the minimal epitopes, EPIP ranked 15
329 (48.39%), 12 (38.71%) and 5 (16.13%) CD8+ recognized epitopes in the top 20, 10 and 5.
330 Therefore, despite being a much simpler model trained with less data comparing to EDGE, EPIP

is also quite valuable for immunogenicity prediction in neoantigen-based cancer immunotherapies and is a significant improvement over other publicly available methods.

EPIP can be extended to support more alleles by incorporating new MS data

Due to its simplicity and modular structure, EPIP can be straightforwardly extended to support more HLA alleles by incorporating new MS data. An allele that we are particularly interested in is HLA-A*33:03, which appears often in the Chinese population but with very limited MS or binding affinity data prior to this publication (more details in the supplementary material). We first performed MS profiling of eluted HLA ligands on the HCC4006 lung cancer cell line, which is homozygous on all three HLA alleles (HLA-A*33:03, HLA-B*44:03 and HLA-C*07:06, respectively). Upon deconvolution of MS data by GibbsCluster and removing the trash cluster, we were only able to obtain two clusters, likely due to the low expression of HLA-C alleles. We were able to assign one of the clusters to HLA-B*44:03 by comparing to the known motif of HLA-B*44:03 obtained from mono-allelic HLA-B*44:03 cell line [12]. We believe the other cluster should belong to HLA-A*33:03, since the motif is quite similar to the A33:03 motif predicted by NetMHCpan-4.0 in its pan-specific mode (although binding affinity has not been directly profiled for this allele) and ~80% of the peptides in this cluster has binding affinity < 500nm to HLA-A*33:03 as predicted by NetMHCpan-4.0. To further check on this, we also performed t-SNE [39] analysis for the peptides in this cluster, together with peptides of HLA-*A02:01 and HLA-A*11:01 in EPIP's training set. As shown in Figure S4, peptides from this cluster is much closer to those from HLA-A*11:01 than HLA-A*02:01, consistent with the fact that HLA-A*33:03 and HLA-A*11:01 belong to the same HLA supertype while HLA-A*02:01

354 belongs to a different supertype. Based on these evidences, we confidently assigned this cluster
355 to HLA-A*33:03. By inputting these HLA-A*33:03-specific peptides to EPIP, where a dedicated
356 module has been built to accept new allele-specific MS data, it can learn the corresponding
357 PSSM and length distribution and add HLA-A*33:03 to its supported alleles. The overall work
358 flow is illustrated in Figure 5.
359

360 **Discussions**

361

362 Besides making use of recently available, large-scale MS immunopeptidome data, our method
363 clearly demonstrated the advantage of integrating gene expression levels in a principled way
364 when modelling epitope presentation, but the use of gene expression levels could present its own
365 challenges. Even when discovering neoantigens for cancer patients where RNA-seq is routinely
366 preformed, this information could be missing or inaccurate. In this case, since it is well-known
367 that gene expression levels are largely preserved in the same tissue [40], one could make use of
368 the gene expression levels quantified in a different sample from a similar tissue origin. For
369 example, we showed that when replacing the expression levels of train_sample55 with HELA,
370 which are both originated from ovary tumors, the performance of EPIP on HLA-A*2402 only
371 dropped a little and still significantly outperformed other public methods (Figure S6). In other
372 applications of epitope prediction, such as those originating from pathogens, the definition of
373 gene expression levels could be problematic, but since virus employ the human cell machinery to
374 produce their protein, the virus epitopes may share the same processing disciplines with human
375 epitopes. As more sophisticated RNA-seq analysis methods are being developed to analyze the

transcriptome of virus infected cells and calculate the TPM/RPKM of specific virus proteins [41], EPIP could be used to predict virus epitopes in these cells as well.

The incorporation of gene expression levels in EPIP is similar in principle to what has been done by EDGE [7]. Although EDGE is neither publicly available nor straightforward to be reproduced, some comparisons with EDGE on epitope presentation prediction would still be informative. To achieve this, we used the same evaluation criteria as in EDGE to compare against other affinity-based methods, i.e., to further reduce the positive to negative ratio, or prevalence, to 1:2,500 and 1:10,000 and checked the PPVs of EPIP at 40% recall. Although it is difficult to justify the optimal threshold of prevalence for neoantigen prediction, further reducing this threshold does have the side effect of enlarging the performance gap between EPIP and other methods. As shown in Figure S5, EPIP reached a more than 10-fold improvement over MHCflurry with 1:10,000 prevalence, similar to what has been achieved by EDGE. Combined with the fact that EPIP and EDGE performed similarly on the same immunogenicity test data, we concluded that the performances of EPIP and EDGE should be quite comparable. Since EPIP is a much simpler model comparing to EDGE, this also implies that the performance gain of EDGE is to a larger degree contributed by its large-scale training data than the more sophisticated deep learning framework.

One clear future work for EPIP is to add more allele support as more MS data becomes available, both from public sources and our own experiments. Enabled by EPIP's simple and modular structure, doing so is quite straightforward as demonstrated by adding the support of HLA-A*33:03, and we will keep doing so and update our web tool over time. We chose to adopt simple machine learning techniques, i.e., PSSMs and LR, for this initial version of EPIP for easy analysis and development. More recent deep learning techniques for sequence analysis, mostly

399 developed in the field of natural language processing [42, 43], can make better use of large
400 datasets and model different contributing factors more effectively, and could lead to more
401 powerful variable length, pan-specific models. This is another promising direction for future
402 versions of EPIP. Moreover, immunogenicity prediction in a clinical setting depends on more
403 than epitope presentation [44-46], and we believe it is important to account for the extra factors
404 contributing to immunogenicity as well. To that end, community efforts to generate more
405 immunogenicity data and improved experimental techniques to profile TCR and peptide-MHC
406 interactions [47, 48] might lead to larger datasets and deeper biological understanding that can
407 eventually enable accurate predictions of immunogenicity.
408

409 ## Conclusion

410

411 In this paper, we have developed a user-friendly, publicly available epitope prediction tool,
412 EPIP, that incorporates information from both MS and RNA-seq data. It significantly
413 outperforms other publicly available tools and can be easily extended to support more HLA
414 alleles when new MS data becomes available. We believe EPIP is a timely contribution to the
415 community to advance cancer immunotherapies and beyond.
416

417 **List of abbreviations**

418

419 HLA: human leukocyte antigen
420 MS: Mass spectrometry
421 PSSMs: position score specific matrices

422 LR: logistic regression

423 EPIP: Epitope Presentation Integrated Prediction

424 MHC: major histocompatibility complex

425 TCRs: T-cell receptors

426 PMF: probability mass function

427 PPV: positive predictive value

428 TMG: tandem minigene

429 PRCs: precision recall curves

430

431 **Declarations**

432

433 **Ethics approval and consent to participate**

434

435 Not applicable

436

437 **Consent for publication**

438

439 Not applicable

440

441 **Availability of data and materials**

442

443 The MS data of HCC4006 generated during the current study are available in the CNGB

444 Nucleotide Sequence Archive (CNSA: https://db.cngb.org).

Competing interests

The authors declare that they have no competing interests

Funding

This project is supported by National Natural Science Foundation of China (grant numbers 81702826 and 81772910), Science, Technology and Innovation Commission of Shenzhen Municipality (grant no. JCYJ20170303151334808) and Shenzhen Municipal Government of China (grant no. 20170731162715261)

Authors' contributions

SQ, LJL, WH, BL designed the project. WH, XH, SZ, GL, JL implemented the EPIP model. WH and YL collected the data for training and evaluation of the model. XL, LC, SL cultured the HCC4006 cell line. XL, LC, SL, GH, ZL, YR generated the MS data of HCC4006. LZ, YL, WH, WL, HX, ZR analyzed the MS data of HCC4006. LJL, WH, SQ, YL, XL, LZ, BL wrote the manuscript. LJL, SQ, BL supervised the project. All authors read and approved the final manuscript.

Acknowledgements

468 We wish to thank Dian Dong for helping with formatting the paper and references.

469

470 **References**

471

472 1. U. Sahin, E. Derhovanessian, M. Miller, B. P. Kloke, P. Simon, M. Lower, et al.
473 Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity
474 against cancer. Nature. 2017; 547: 222-226. doi:10.1038/nature23003.
475 2. P. A. Ott, Z. Hu, D. B. Keskin, S. A. Shukla, J. Sun, D. J. Bozym, et al. An immunogenic
476 personal neoantigen vaccine for patients with melanoma. Nature. 2017; 547: 217-221.
477 doi:10.1038/nature22991.
478 3. N. Hilf, S. Kuttruff-Coqui, K. Frenzel, V. Bukur, S. Stevanovic, C. Gouttefangeas, et al.
479 Actively personalized vaccination trial for newly diagnosed glioblastoma. Nature. 2019;
480 565: 240-245. doi:10.1038/s41586-018-0810-y.
481 4. D. B. Keskin, A. J. Anandappa, J. Sun, I. Tirosh, N. D. Mathewson, S. Li, et al.
482 Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial.
483 Nature. 2019; 565: 234-239. doi:10.1038/s41586-018-0792-9.
484 5. V. Jurtz, S. Paul, M. Andreatta, P. Marcatili, B. Peters and M. Nielsen. NetMHCpan-4.0:
485 Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and
486 Peptide Binding Affinity Data. J Immunol. 2017; 199: 3360-3368.
487 doi:10.4049/jimmunol.1700893.
488 6. D. Gfeller and M. Bassani-Sternberg. Predicting Antigen Presentation-What Could We
489 Learn From a Million Peptides? Front Immunol. 2018; 9: 1-17.
490 doi:10.3389/fimmu.2018.01716.

491 7. B. Bulik-Sullivan, J. Busby, C. D. Palmer, M. J. Davis, T. Murphy, A. Clark, et al. Deep
492 learning using tumor HLA peptide mass spectrometry datasets improves neoantigen
493 identification. Nat Biotechnol. 2018. doi:10.1038/nbt.4313.
494 8. T. J. O'Donnell, A. Rubinsteyn, M. Bonsack, A. B. Riemer, U. Laserson and J.
495 Hammerbacher. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction.
496 Cell Syst. 2018; 7: 129-132 e124. doi:10.1016/j.cels.2018.05.014.
497 9. M. Bassani-Sternberg, C. Chong, P. Guillaume, M. Solleder, H. Pak, P. O. Gannon, et al.
498 Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and
499 identifies allostery regulating HLA specificity. PLoS Comput Biol. 2017; 13: e1005725.
500 doi:10.1371/journal.pcbi.1005725.
501 10. D. Gfeller, P. Guillaume, J. Michaux, H.-S. Pak, R. T. Daniel, J. Racle, et al. Peptide
502 length distribution and multiple specificity in naturally presented HLA-I ligands. 2018.
503 doi:10.1101/335661.
504 11. M. Bassani-Sternberg, S. Pletscher-Frankild, L. J. Jensen and M. Mann. Mass
505 spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of
506 protein abundance and turnover on antigen presentation. Mol Cell Proteomics. 2015; 14:
507 658-673. doi:10.1074/mcp.M114.042812.
508 12. J. G. Abelin, D. B. Keskin, S. Sarkizova, C. R. Hartigan, W. Zhang, J. Sidney, et al. Mass
509 Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables
510 More Accurate Epitope Prediction. Immunity. 2017; 46: 315-326.
511 doi:10.1016/j.immuni.2017.02.007.

512 13. H. Pearson, T. Daouda, D. P. Granados, C. Durette, E. Bonneil, M. Courcelles, et al.
513 MHC class I-associated peptides derive from selective regions of the human genome. J
514 Clin Invest. 2016; 126: 4690-4701. doi:10.1172/JCI88590.
515 14. M. Di Marco, H. Schuster, L. Backert, M. Ghosh, H. G. Rammensee and S. Stevanovic.
516 Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides
517 and Generation of Binding Prediction Matrices. J Immunol. 2017; 199: 2639-2651.
518 doi:10.4049/jimmunol.1700938.
519 15. W. Shao, P. G. A. Pedrioli, W. Wolski, C. Scurtescu, E. Schmid, J. A. Vizcaino, et al.
520 The SysteMHC Atlas project. Nucleic Acids Res. 2018; 46: D1237-D1247.
521 doi:10.1093/nar/gkx664.
522 16. T. Trolle, C. P. McMurtrey, J. Sidney, W. Bardet, S. C. Osborn, T. Kaever, et al. The
523 Length Distribution of Class I–Restricted T Cell Epitopes Is Determined by Both Peptide
524 Supply and MHC Allele–Specific Binding Preference. The Journal of Immunology.
525 2016; 196: 1480-1487. doi:10.4049/jimmunol.1501721.
526 17. S. Kim and P. A. Pevzner. MS-GF+ makes progress towards a universal database search
527 tool for proteomics. Nat Commun. 2014; 5: 5277. doi:10.1038/ncomms6277.
528 18. M. The, M. J. MacCoss, W. S. Noble and L. Kall. Fast and Accurate Protein False
529 Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. J Am Soc
530 Mass Spectrom. 2016; 27: 1719-1727. doi:10.1007/s13361-016-1460-7.
531 19. F. Zhou, H. Cao, X. Zuo, T. Zhang, X. Zhang, X. Liu, et al. Deep sequencing of the
532 MHC region in the Chinese population contributes to studies of complex disease. Nat
533 Genet. 2016; 48: 740-746. doi:10.1038/ng.3576.

534 20. S. Chen, Y. Zhou, Y. Chen and J. Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor.
535 Bioinformatics. 2018; 34: i884-i890. doi:10.1093/bioinformatics/bty560.
536 21. C. A. D. Alexander Dobin, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha,
537 Philippe Batut, Mark Chaisson and Thomas R. Gingeras. STAR: ultrafast universal RNA-
538 seq aligner. Bioinformatics Advance Access. 2012.
539 22. C. N. D. Bo Li. RSEM: Accurate transcript quantification from RNA-seq data with or
540 without a reference genome. Bioinformatics: The Impact of Accurate Quantification on
541 Proteomic and Genetic Analysis and Research. . 2014. doi:10.1201/b16589.
542 23. G. Liu, D. Li, Z. Li, S. Qiu, W. Li, C. C. Chao, et al. PSSMHCpan: a novel PSSM-based
543 software for predicting class I peptide-HLA binding affinity. Gigascience. 2017; 6: 1-11.
544 doi:10.1093/gigascience/gix017.
545 24. M. Andreatta, B. Alvarez and M. Nielsen. GibbsCluster: unsupervised clustering and
546 alignment of peptide sequences. Nucleic Acids Res. 2017; 45: W458-W463.
547 doi:10.1093/nar/gkx248.
548 25. L. BREIMAN. Stacked Regressions. Mach Learn. 1996. doi:10.1007/BF00117832.
549 26. A. G. N. Pedregosa FABIANPEDREGOSA F, Michel V, Thirion BERTRANDTHIRION
550 B, Grisel O, Blondel M. Scikitlearn: Machine Learning in Python Gaël Varoquaux. J
551 Mach Learn Res. 2011. doi:10.1007/s13398-014-0173-7.2.
552 27. A. M. Tran E, Lu YC, Gros A, Turcotte S, Robbins PF, et al. Immunogenicity of somatic
553 mutations in human gastrointestinal cancers. Science (80-). 2015; 350: 1387-1390.
554 doi:10.1126/science.aad1253
555

556 28. A. Gros, M. R. Parkhurst, E. Tran, A. Pasetto, P. F. Robbins, S. Ilyas, et al. Prospective
557 identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma
558 patients. Nat Med. 2016; 22: 433-438. doi:10.1038/nm.4051.
559 29. T. M. Strønen E, Kelderman S, Van Buuren MM, Yang W, Van Rooij N, et al. Targeting
560 of cancer neoantigens with donor-derived T cell receptor repertoires. Science. 2016.
561 doi:10.1126/science.aaf2288.
562 30. N. Zacharakis, H. Chinnasamy, M. Black, H. Xu, Y. C. Lu, Z. Zheng, et al. Immune
563 recognition of somatic mutations leading to complete durable regression in metastatic
564 breast cancer. Nat Med. 2018; 24: 724-730. doi:10.1038/s41591-018-0040-8.
565 31. J. Wu, W. Wang, J. Zhang, B. Zhou, W. Zhao, Z. Su, et al. DeepHLApan: A Deep
566 Learning Approach for Neoantigen Prediction Considering Both HLA-Peptide Binding
567 and Immunogenicity. Front Immunol. 2019; 10: 2559. doi:10.3389/fimmu.2019.02559.
568 32. H. R. Hunt DF, Shabanowitz J, Sakaguchi K, Michel H, Sevilir N, et al. .
569 Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass
570 spectrometry. . Science (80-). 1992. doi:10.1126/science.1546328.
571 33. F. T. Rammensee HG, Stevanović S. MHC ligands and peptide motifs: first listing.
572 Immunogenetics. 1995. doi:10.1007/BF00172063.
573 34. R. Vita, J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, et
574 al. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2015; 43: D405-412.
575 doi:10.1093/nar/gku938.
576 35. G. J. Reche PA, Reinherz EL. Prediction of MHC class I binding peptides using profile
577 motifs. Hum Immunol. 2002; 63: 701–709. doi:10.1016/S0198-8859(02)00432-9.

578 36. H. Zhang, O. Lund and M. Nielsen. The PickPocket method for predicting binding
579 specificities for receptors based on receptor pocket similarities: application to MHC-
580 peptide binding. Bioinformatics. 2009; 25: 1293-1299.
581 doi:10.1093/bioinformatics/btp137.
582 37. B. Schubert and O. Kohlbacher. Designing string-of-beads vaccines with optimal spacers.
583 Genome Med. 2016; 8: 9. doi:10.1186/s13073-016-0263-6.
584 38. The problem with neoantigen prediction. Nat Biotechnol. 2017; 35: 97.
585 doi:10.1038/nbt.3800.
586 39. H. G. Van Der Maaten L. Visualizing Data using t-SNE. J Mach Learn Res. . 2008.
587 doi:10.1007/s10479-011-0841-3.
588 40. D. Brawand, M. Soumillon, A. Necsulea, P. Julien, G. Csardi, P. Harrigan, et al. The
589 evolution of gene expression levels in mammalian organs. Nature. 2011; 478: 343-348.
590 doi:10.1038/nature10532.
591 41. R. J. Peng, B. W. Han, Q. Q. Cai, X. Y. Zuo, T. Xia, J. R. Chen, et al. Genomic and
592 transcriptomic landscapes of Epstein-Barr virus in extranodal natural killer T-cell
593 lymphoma. Leukemia. 2019; 33: 1451-1462. doi:10.1038/s41375-018-0324-5.
594 42. A. M. Gehring J, Grangier D, Yarats D, Dauphin YN. Convolutional sequence to
595 sequence learning. . 34th International Conference on Machine Learning, ICML 2017.
596 2017.
597 43. C. M.-W. Devlin J, Lee K, Toutanova K. . Bert: Pre-training of deep bidirectional
598 transformers for language understanding. 2018. doi:arXiv Prepr arXiv181004805.

599 44. S. J. Assarsson E, Oseroff C, Pasquetto V, Bui H-H, Frahm N,et al. A Quantitative
600 Analysis of the Variables Affecting the Repertoire of T Cell Specificities Recognized
601 after Vaccinia Virus Infection. 2007. doi:178/12/7890 [pii].
602 45. J. J. Calis, M. Maybeno, J. A. Greenbaum, D. Weiskopf, A. D. De Silva, A. Sette, et al.
603 Properties of MHC class I presented peptides that enhance immunogenicity. PLoS
604 Comput Biol. 2013; 9: e1003266. doi:10.1371/journal.pcbi.1003266.
605 46. F. A. McGranahan N, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, et al. . Clonal
606 neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint
607 blockade. . Science (80-). 2016; 351: 1463-1469. doi:10.1126/science.aaf1490.
608 47. A. K. Bentzen, A. M. Marquard, R. Lyngaa, S. K. Saini, S. Ramskov, M. Donia, et al.
609 Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled
610 with DNA barcodes. Nat Biotechnol. 2016; 34: 1037-1045. doi:10.1038/nbt.3662.
611 48. A. K. Bentzen, L. Such, K. K. Jensen, A. M. Marquard, L. E. Jessen, N. J. Miller, et al. T
612 cell receptor fingerprinting enables in-depth characterization of the interactions
613 governing recognition of peptide-MHC complexes. Nat Biotechnol. 2018.
614 doi:10.1038/nbt.4303.
615

616 **Figure legend**

617 Fig. 1 Construction of EPIP_s and EPIP. a) Construction of allele and tissue-specific EPIP
618 (EPIP_s). EPIP_s was built for 21 alleles that have matching expression profile.
619 b) Saturation analysis of PSSM models. The PSSM models were trained and tested on 9-
620 11mer peptides independently for six HLA alleles; evaluated with 0.1% PPV, the performances

621 of PSSM improved with increasing training data, and reached a plateau with about 100-200
622 peptides, depending on the specific alleles.
623 c) Construction of EPIP. To build the full EPIP model, we pooled peptides from 21 alleles
624 encoded by PSSM, expression and length features to fit a single LR model
625
626 Fig. 2 Relationship among weights and biases of 21 EPIP_s models. a) Boxplot for feature
627 weights and biases of EPIP_s models: PSSMs contributed most to the prediction, followed by
628 expressions and peptide lengths.
629 b) Hive plots demonstrating the similar inter-relationships among the feature weights and
630 biases of different EPIP_s models, where the negative biases have been inverted for plotting.
631
632 Fig. 3 Comparison of prediction performances between EPIP and EPIP_s. a) Performance
633 comparison of EPIP and EPIP_s across different alleles originating from the same B cell line.
634 Error bars indicate the standard deviations across five-fold cross validation. The performance
635 difference between EPIP_s and EPIP across 21 alleles is not statistically significant ($P = 0.1385$,
636 paired t-test)
637 b) Heatmap demonstrates diverse expression profiles of different cell lines originating from
638 different tissues. Each row represents a gene that presents peptides in at least one cell line, a total
639 of 10068 genes was selected. Each column represents a cell line that used in this study, the
640 originated tissue is indicated below the cell line. The expression values are transformed by
641 log2(TPM+1).
642 c-f) Performance comparison of EPIP and EPIP_s for the same allele originating from
643 different cell lines. Error bar indicate the standard deviations across five-fold cross validation.

644

645 Fig 4. Comparison of prediction performances between EPIP and other methods. a)
646 Performances of EPIP and other publicly available methods on independent MS dataset. The
647 evaluation dataset was obtained from four publications that contained a total of 30 MS datasets,
648 including 16 different tissue origins and 17 HLA alleles. The mean 0.1% PPV of these 30
649 datasets were displayed. EPIP significantly outperformed other methods with different
650 expression filtering thresholds (TPM > 0, > 1, > 2), and pairwise P-values (paired t-test) for
651 TPM > 2 were shown in the plot.
652 b). Precision-recall curves of EPIP and other publicly available methods on T cell response
653 data. Other methods' predictions were made with expression threshold of TPM > 0. Comparisons
654 of EPIP and other methods with TPM > 1 and TPM > 2 were displayed in Fig S3.
655 c) Prediction performance of EPIP and EDGE on immunogenic SNVs. EDGE's performance
656 was recalculated according to the supplementary data 3c of the original paper.
657 d) Prediction performance of EPIP and EDGE on minimal epitopes. EDGE's performance
658 was taken from Figure 4 of the original paper.

659

660

661 Fig 5. The workflow of adding new allele HLA-A*33:03 to EPIP's supported allele repository.
662 HCC4006 is homozygous on all three HLA alleles (HLA-A*33:03, HLA-B*44:03 and HLA-
663 C*07:06, respectively). Presented peptides from HCC4006 were identified by MS (see Methods).
664 The derived peptidome was then deconvolved using GibbsCluster2.0. Apart from a trash cluster
665 of 67 peptides, two peptide clusters were identified. One cluster matched the sequence motif of
666 B4403 (derived from HLA-B*44:03 mono-allelic dataset [12]), and the other matched the

sequence motif of A33:03 derived from NetMHCpan4.0 online motif reviewer. Noted that due to the lack of training data, NetMHCpan4.0 generated the A33:03 motif based on training data from other HLA alleles in a pan-specific mode, which could explain the slight difference between the motif we discovered and the NetMHCpan-4.0 provided. We then inputted the HLA-A*33:03-specific peptides to EPIP, which could learn PSSM and length distributions automatically, and resulted in an updated version of EPIP that could support the newly added allele.

Figure 1

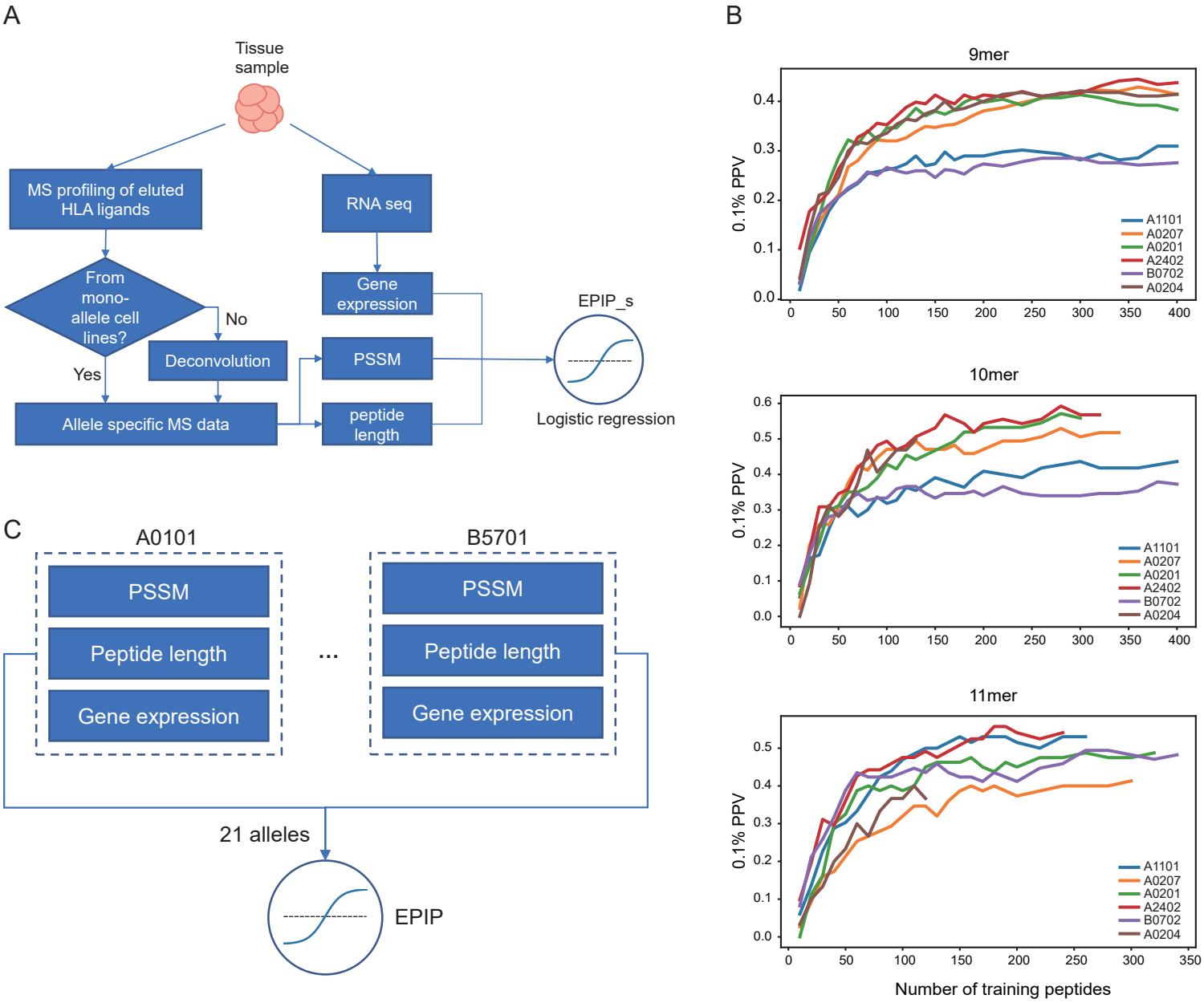
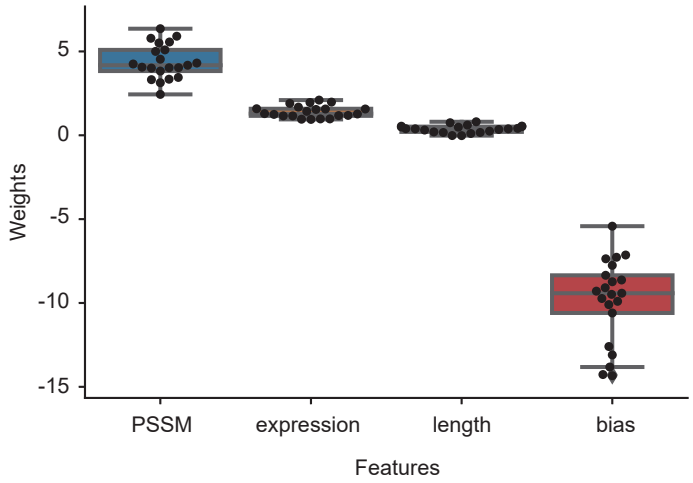

A
Tissue sample
MS profiling of eluted HLA ligands
RNA seq
From mono-allele cell lines?
No
Yes
Deconvolution
Gene expression
PSSM
peptide length
Allele specific MS data
EPIP_s
Logistic regression
B
9mer
10mer
11mer
0.1% PPV
A1101
A0207
A0201
A2402
B0702
A0204
Number of training peptides
C
A0101
B5701
PSSM
Peptide length
Gene expression
...
21 alleles
EPIP

Figure 2

A


5
0
-5
-10
-15
Weights
PSSM
expression
length
bias
Features

B

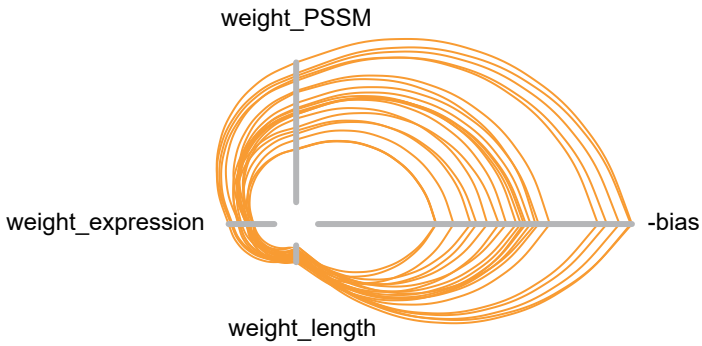

weight_PSSM
weight_expression
-bias
weight_length

Figure 3

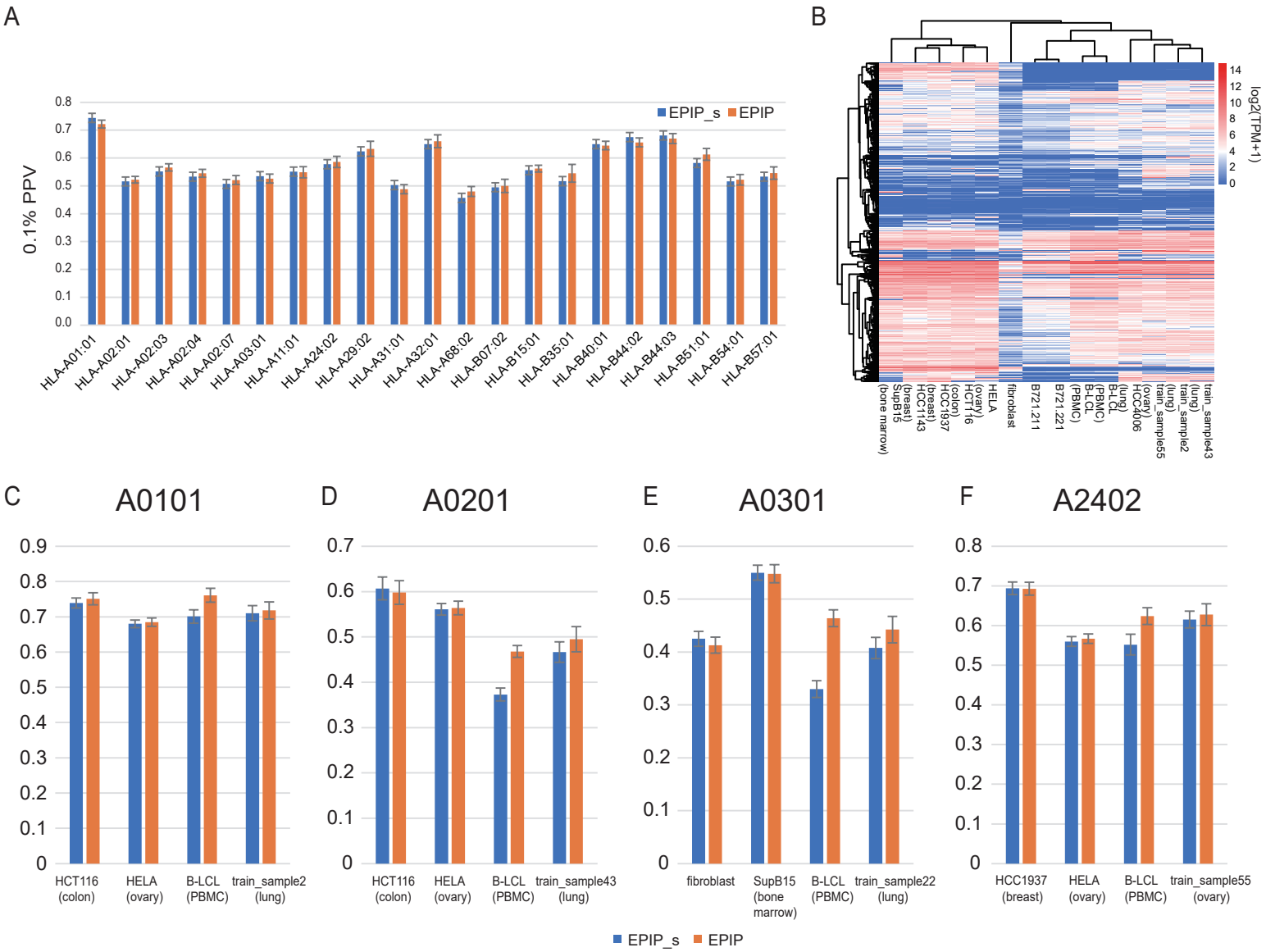

A
0.1% PPV
EPIP_s
EPIP
HLA-A01:01
HLA-A02:01
HLA-A02:03
HLA-A02:04
HLA-A02:07
HLA-A03:01
HLA-A11:01
HLA-A24:02
HLA-A29:02
HLA-A31:01
HLA-A32:01
HLA-A68:02
HLA-B07:02
HLA-B15:01
HLA-B35:01
HLA-B40:01
HLA-B44:02
HLA-B44:03
HLA-B51:01
HLA-B54:01
HLA-B57:01
B
log2(TPM+1)
SupB15 (bone marrow)
HCC1143 (breast)
HCC1937 (breast)
HCT116 (colon)
HELA (ovary)
fibroblast
B721.211
B721.221
B-LCL (PBMC)
HCC4006 (lung)
train_sample55 (ovary)
train_sample2 (lung)
train_sample43 (lung)
C
A0101
HCT116 (colon)
HELA (ovary)
B-LCL (PBMC)
train_sample2 (lung)
D
A0201
HCT116 (colon)
HELA (ovary)
B-LCL (PBMC)
train_sample43 (lung)
E
A0301
fibroblast
SupB15 (bone marrow)
B-LCL (PBMC)
train_sample22 (lung)
F
A2402
HCC1937 (breast)
HELA (ovary)
B-LCL (PBMC)
train_sample55 (ovary)
EPIP_s
EPIP

Figure 4

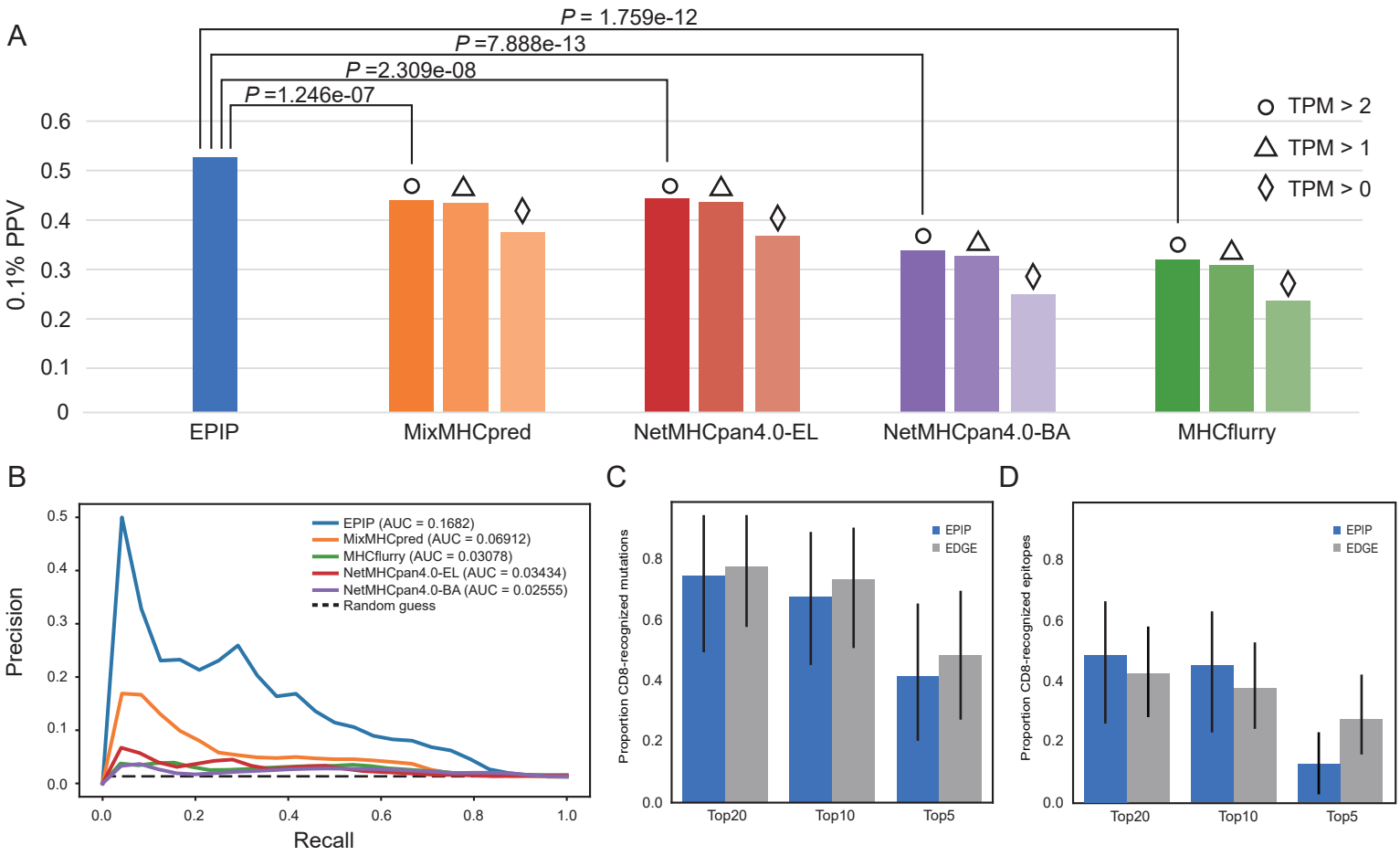
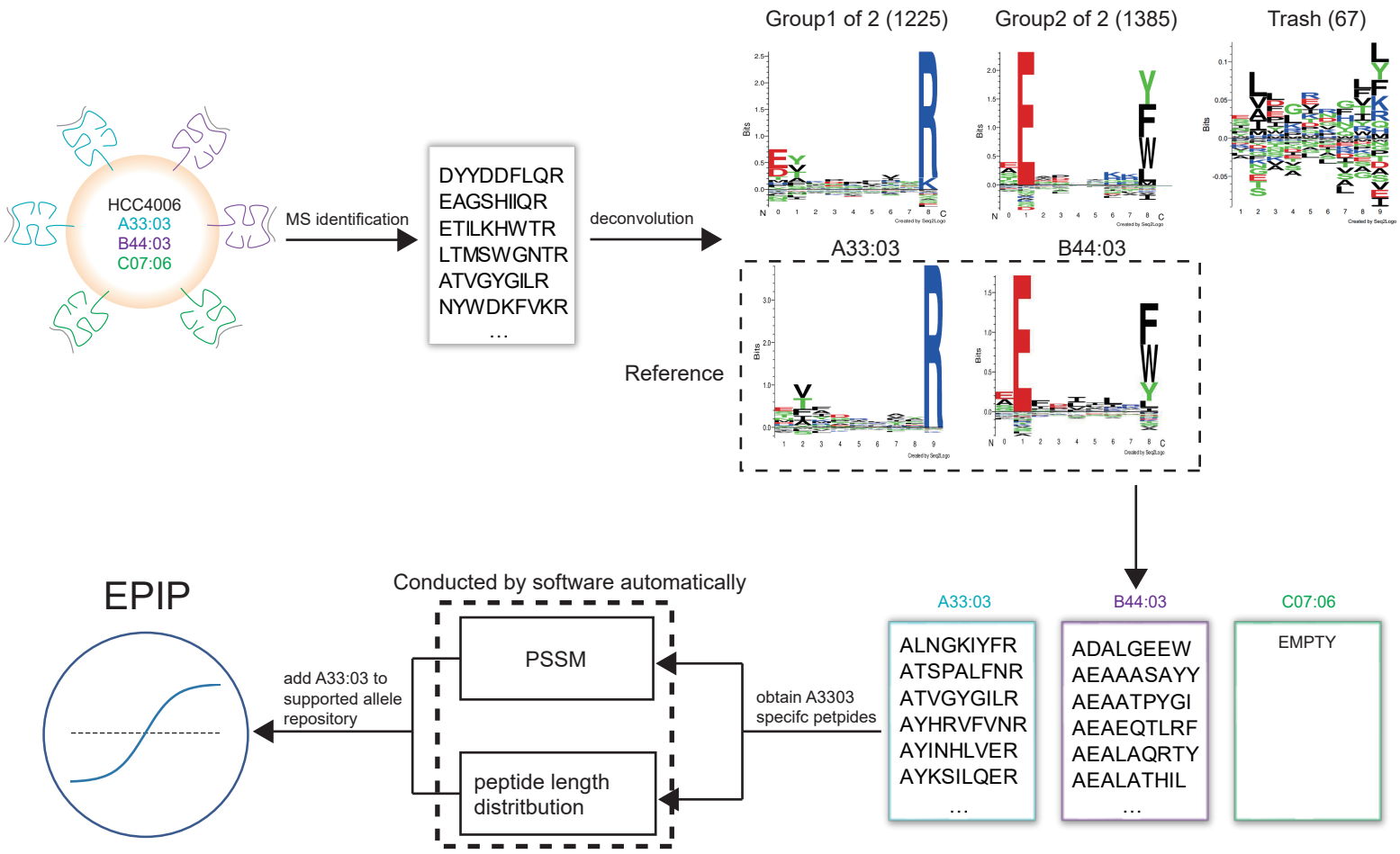

A
P = 1.759e-12
P =7.888e-13
P =2.309e-08
P =1.246e-07
0.1% PPV
0.6
0.5
0.4
0.3
0.2
0.1
0
TPM > 2
TPM > 1
TPM > 0
EPIP
MixMHCpred
NetMHCpan4.0-EL
NetMHCpan4.0-BA
MHCflurry
B
Precision
Recall
EPIP (AUC = 0.1682)
MixMHCpred (AUC = 0.06912)
MHCflurry (AUC = 0.03078)
NetMHCpan4.0-EL (AUC = 0.03434)
NetMHCpan4.0-BA (AUC = 0.02555)
Random guess
C
Proportion CD8-recognized mutations
EPIP
EDGE
Top20
Top10
Top5
D
Proportion CD8-recognized epitopes
EPIP
EDGE
Top20
Top10
Top5

Figure 5


HCC4006
A33:03
B44:03
C07:06
MS identification
DYYDDFLQR
EAGSHIIQR
ETILKHWTR
LTMSWGNTR
ATVGYGILR
NYWDKFVKR
…
deconvolution
Group1 of 2 (1225)
Group2 of 2 (1385)
Trash (67)
A33:03
B44:03
Reference
A33:03
B44:03
C07:06
ALNGKIYFR
ATSPALFNR
ATVGYGILR
AYHRVFVNR
AYINHLVER
AYKSILQER
…
ADALGEEW
AEAAASAYY
AEAATPYGI
AEAEQTLRF
AEALAQRTY
AEALATHIL
…
EMPTY
obtain A3303
specifc petpides
Conducted by software automatically
PSSM
peptide length
distritbution
add A33:03 to
supported allele
repository
EPIP