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

Weipeng Hu*[1,2,3], Si Qiu*[2,3], Youping Li[2,3,4], Xinxin Lin[2,3], Le Zhang[2,3,4], Haitao Xiang[2], Xing Han[2,3], Lei Chen[2,3], Sha Li[2,3], Wenhui Li[2,3], Zhe Ren[2], GuiXue Hou[2], Zhilong Lin[2], Jianliang Lu[2], Geng Liu[2,3], Bo Li@[2,3,5], Leo J Lee@[2,3,6]

1. School of Biology and Biological Engineering, South China University of Technology, Guangzhou 510006, China
2. BGI-Shenzhen, Shenzhen 518083, China
3. BGI-GenoImmune, Wuhan 4300794, China
4. BGI Education Center, University of Chinese Academy of Sciences, Shenzhen 518083, China
5. Guangdong Enterprise Key Laboratory of Human Disease Genomics, Shenzhen 518083, China
6. 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 unbiased, high-throughput measurements of HLA associated peptides *in vivo*, could 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. Comparing the feature weights and biases across different HLA-I alleles and cell lines, we observed a universal relationship among these three variables. To confirm this, we built a single LR model by pooling PSSM scores, gene expression levels and peptide length features across different HLA alleles and cell lines, and compared its performance with the allele and cell line specific LR models. Indeed, the predictive powers had no significant differences across cell lines and HLA alleles, and both substantially outperformed predictions based on PSSM scores alone. Based on such a finding, we further built a universal LR model, termed Epitope Presentation Integrated prediCtion (EPIC), 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 EPIC 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 51.59%, compared to 36.98%, 36.41%, 24.67% and 23.39% 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 yet publicly available, on predicting epitope presentation and selecting immunogenic cancer neoantigens. However, the simplicity and flexibility of EPIC makes it much easier to be applied in diverse situations, especially when users would like to take advantage of emerging eluted ligand data for new HLA alleles. We demonstrated this by generating MS data for the HCC4006 cell line and adding the support of HLA-A*33:03, which has no previous MS or binding affinity data available, to EPIC. EPIC is publicly available at <https://github.com/BGI2016/EPIC>.

Conclusions: we have developed an easy to use, publicly available epitope prediction tool, EPIC, 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 traditional binding assay data [5,8]. For example, NetMHCpan-4.0 trained an ensemble of two hidden layer neural networks on both affinity and MS data, and was able to make predictions on both types of outputs. MHCflurry adopted a deep learning framework containing both locally connected and fully connected layers and was mainly trained on affinity data, but used MS data at the model selection stage, while MS data could play an even bigger role in its future releases. Other methods chose to rely on MS data alone [7,9,10], such as MixMHCpred [9,10], which built allele-specific position score specific matrixes (PSSMs) after deconvolving mixed-allele MS data, and the very recent EDGE method [7], which used a comprehensive, deep learning model that pooled together large sets of mono-allele and mixed-allele MS data as well as matching RNA-seq data. With the increasing amount of MS data becoming available, we think that it is better to build separate models for binding affinity (based on binding assay data) and epitope presentation (based on MS data) since the underlying data generation processes are different. These two related pieces of information could probably be combined at a later stage to contribute to the prediction of immunogenicity. The presentation of a peptide by MHC not only depends on information contained within the protein sequences but is also modulated by other factors such as protein abundance, localization and turnover [11,12]. Among these, the corresponding gene expression levels can be conveniently profiled by RNA-seq and have been shown to significantly improve prediction accuracy, first by the MS IntrinsicEC method on mono-allele cell lines [12], then more comprehensively by EDGE [7]. Unfortunately, both methods are not yet publicly available.

We set out to develop an effective, flexible and publicly available HLA-I epitope prediction method that takes advantage of large sets of MS and RNA-seq data. We first collected public MS and matching RNA-seq datasets for 21 HLA-I alleles, which frequently appear in the European or Asian population. By carefully examining the contributions of PSSM scores, gene expression levels and peptide length features to epitope presentation, we were able to effectively model these variables with a universal logistic regression (LR) model. We then trained such a LR model, termed Epitope Presentation Integrated prediCtion (EPIC), to support 66 common HLA-I alleles. In the remainder of this article, we first describe our design rationale in detail, followed by comprehensive evaluation of EPIC on large, independent MS and immunogenic datasets as well as a user case showing its extendibility, before concluding with some discussions on future improvements and research directions. We believe EPIC is a timely contribution to the research community of cancer immunotherapies.

Methods

Collection of MS and RNA-seq data

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

Construction of allele and tissue-specific EPIC (EPIC_s)

From the matching MS and RNA-seq data collected from public sources [12,13], we first built LR models to predict epitope presentation for 21 alleles. This included all 16 mono-allele data profiled in Abelin et al [12], as well as five more alleles (HLA-A*11:01, HLA-A*32:01, HLA-B*15:01, HLA-B*40:01, HLA-B*07:02) selected from Pearson et al [13], which frequently appear in the Chinese population [17]. For mono-allele MS data, we built PSSMs directly, as described in Liu et al [18]. For mixed-allele MS data, we performed motif deconvolution by GibbsCluster [19] followed by manual inspection to assign peptides to different HLA alleles, before building PSSMs. For each HLA allele, we also obtained an empirical probability mass function (pmf) to depict the length distribution of 8-15-mers based on the allele-specific MS data. All RNA-seq datasets were downloaded and re-processed with the same pipeline to ensure consistency. Specifically, raw fastq files were cleaned by fastp (version 0.18.0) [20] and aligned to human genome assembly GRCh38 and its associated transcriptome (Ensembl Release 92) with STAR (version 2.5.3a) [21]; gene expression (transcripts per million or TPM) was quantified by RSEM (version 1.3.0) [22] based on the alignment. After obtaining the allele and tissue-specific PSSM, peptide length distribution and gene expression, we built a LR model to predict peptide presentation, and the process is sketched in Figure 1A.

For each allele, length-specific PSSMs were built for those peptides that have enough MS data. To fully utilize the training data and avoid overfitting, we used the stacking method [23] to build PSSM (see supplementary materials, Fig S1). PSSM is a simple model that does not require much data to reach its optimum performance. To quantify this, we used subsampling to build PSSM models for different lengths of peptides to predict epitope presentation and tested their performance on various alleles. As shown in Figure 1B, the performances of PSSM models reach saturation when having 100-200 peptides in the training set. 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 this allowed us to make use of 74%-96% of MS profiled peptides across different alleles. The allele and tissue-specific EPIC (EPIC_s) also incorporated the peptide length distribution and the corresponding gene expression levels. The peptide length feature is the corresponding probability in the empirical pmf of peptide length distribution. Expression values of a peptide is calculated by summing up the expression levels of all transcripts containing the peptide (in TPM) before transformed by log2(TPM+1). For the mono-allele cell lines in Abelin et al [12], transcript expression levels are taken as the average of the four cell lines with RNA-seq data available; similarly, for the mixed-allele cell lines in Pearson et al [13] , transcript expression levels are averaged over the 10 cell lines with RNA-seq data available. All three features were further normalized to be zero mean, unit variance Gaussian before using as inputs to a LR model with weak L2 regularization implemented in scikit-learn 0.19.0 [24]. MS profiled peptides can only provide a positive training set for our model. To generate the negative set, we randomly sampled the human peptidome to obtain peptides of the matching length and remove those that overlap with the positive set. Adopting the same strategy as in NetMHCpan 4.0, we used more negative peptides than positive ones to train EPIC_s. For example, if the positive set contains 9-11-mers with the most abundant 9-mers having a size of *n*, we then generate a negative set with 10*n* peptides for each length (30*n* in total). To evaluate our model by 0.1% positive predictive value (PPV), we further generated 999-fold decoy peptides that overlapped with neither the positive nor the negative sets used in training. Five-fold cross validation was used to evaluate EPIC_s during training, and to compare with the full EPIC model described next.

Building the EPIC model

To build the full EPIC model for the above 21 alleles from different cell lines, we pooled all the PSSM scores, peptide length distributions and corresponding gene expression levels together to train a single LR model (Figure 1C). The same negative and 999-fold decoy peptide sets as described above were also used in training and evaluating EPIC. Five-fold cross validation was used to evaluate EPIC during training, while the full training set was used to build the final EPIC model and it was evaluated on independent test data. To expand the number of HLA alleles that EPIC can support, mono-allele HLA-C MS data from Di Marco et al [14], deconvolved MS data from SysteMHC [15] and some mixed-allele MS data released by EDGE [7] were used to learn PSSMs for these additional HLA alleles. We also generated additional MS-data for the HCC4006 cell line in order to support HLA-A*33:03 in EPIC, and the procedure is described as a user case in the Results section.

Immunogenicity evaluation

The immunogenic dataset used in this paper is the same as in Bulik-Sullivan et al. [7], which contains 2,023 assayed mutations from 17 patients with annotated HLA information from four previous studies [25–28]. In three of the four studies [25,26,28] , mutations were tested using 25mer tandem minigene (TMG) assays. For each mutation in these datasets, we used EPIC, MixMHCpred 2.0, NetMHCpan4.0-EL, NetMHCpan4.0-BA, MHCflurry 1.2.2 to make predictions with respect to the HLA alleles of each patient for overlapping 8-11mers that harbor the mutation, and the score of the mutation is takes as the best score among these peptides (EPIC: highest presentation probability; MixMHCpred 2.0: highest predicted score; NetMHCpan4.0-EL and NetMHCpn4.0-BA: lowest rank score; MHCflurry 1.2.2: minimum binding affinity). We then ranked the mutations according to the best score from each software. For the fourth study [27] , we ranked mutations by taking the best score from each software across all mutation-spanning peptides tested in the tetramer assays. To compare EPIC with EDGE on predicting immunogenic peptides, we excluded the peptides whose immunogenicity was undetermined as described in Bulik-Sullivan et al. We retrieved a total of 31 immunogenic epitopes from the original studies, which is slightly different from the number 29 reported in Bulik-Sullivan et al. Since the epitope prediction scores by EDGE was not available, we directly compared its immunogenic prediction results with EPIC.

MS profiling of HCC4006 HLA-I peptidome

HLA-I peptidome samples for HCC4006 were prepared according to the literature [11]. They were then analyzed by LC-MS/MS to obtain peptide MS spectra, and full details of the experimental procedure are provided in the Supplementary Material. We employed the MS-GF+ search engine version 2018.07.17 [29] to search the peak lists against the UniProt databases (161,521 entries for human as of December 2017) and a file containing 245 frequently observed contaminants such as human keratins, bovine serum proteins, and proteases. N-terminal acetylation (42.01 Da) and methionine oxidation (15.99 Da) were set as variable modifications. The enzyme specificity was set as unspecific. The initial allowed mass deviation of the precursor ion deviation was set to 10 ppm. Possible peptide lengths were restricted to 8 to 15. Possible precursor charges were restricted to 2 to 5. Range of allowed isotope peak errors was restricted to -1 to 2. Percolator version 3.02.0 was applied as a post-processing step for result filter [30]. The *q* value cutoff was set as 0.01. From the “pout.tab” output file produced by percolator, hits to the contaminants were eliminated.

Results

EPIC_s and EPIC achieve similar performance on epitope presentation prediction

After building the EPIC_s models for 21 HLA alleles, we first examined the distribution of weights and biases among these 21 models. As shown in Figure 2A, the corresponding weights and biases distributed quite tightly. Furthermore, when displaying these 21 sets of the 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. Therefore, we suspect that a single LR model might be able to capture the relationship among PSSM score, length feature and gene expression as well. To confirm this, we pooled the data from all 21 alleles to train a single LR model (EPIC) to compare with the EPIC_s models. As shown in Figure 3A, when evaluating across different alleles originating from the same cell line, the performances (0.1% PPV with 5-fold cross validation) of EPIC and EPIC_s are quite similar. The mean 0.1% PPV of EPIC is slightly better (0.5741 vs 0.5692), likely due to the more training data available to EPIC, but such a difference is not statistically significant (*P* = 0.1385, paired t-test), and both substantially outperformed predictions based on PSSM scores alone (the mean 0.1% PPV of PSSM is 0.4421).We also performed similar analyses for the same HLA alleles across different cell lines with diverse tissue origins and thus diverse gene expression profiles (Figure 3B). When comparing 0.1% PPV of EPIC and EPIC_s, EPIC is not worse than EPIC_s overall, and even notably better for some cell lines (Figure 3C-F), also likely due to the larger training set of EPIC. Based on these comparisons, we concluded that developing a universal LR model, EPIC, to account for the relationship among these three variables is a valid approach. This seemly simple strategy has substantial practical values, since the EPIC model trained on the selected 21 HLA alleles and cell lines from two different tissue origins can now be straightforwardly extended to previously unseen HLA alleles and tissues without adjusting the model parameters of LR, or requiring extra RNA-seq data for training. We did so for 44 more HLA alleles by incorporating public mono-allele [14] and deconvolved mixed allele [7,15] MS data. We also generated our own MS data for this purpose, and this will be demonstrated as a user case after evaluating EPIC against other methods.

EPIC significantly outperforms other existing methods

We evaluated EPIC against MixMHCpred (v2.0), NetMHCpan-4.0 (EL), NetMHCpan-4.0 (BA), and MHCflurry (v1.2.2) on independent validation MS datasets collected from public sources, which contained 30 MS datasets and was not used in the training of EPIC. EPIC significantly outperformed other methods with mean 0.1% PPV of 51.59%, compared to 36.98%, 36.41%, 24.67% and 23.39% achieved by MixMHCpred, NetMHCpan(EL), NetMHCpan (BA) and MHCflurry, with highly significant p-values (*P* = 1.263e-11, 1.075e-11, 5.267e-15, 1.659e-14, respectively). The advantage of EPIC persists after applying additional expression value cutoffs to other methods and the differences are still highly significant, as shown in Figure 4A. Full comparisons for each validation dataset are provided in Figure S2. The recently developed EDGE method has also been shown to excel at epitope presentation prediction comparing to existing methods, but since it is not publicly available, we were not able to do a fair comparison. However, we made some qualitative comparisons to show that the performances of EPIC and EDGE are comparable, and more details are provided in the Discussion section. Overall, the results on independent MS validation sets showed the clear advantage of EPIC in predicting epitope presentation.

EPIC is valuable to immunogenicity prediction

For cancer immunotherapies, as well as other clinical applications, the ultimate task is usually to predict the immunogenicity of an epitope. While this remains an elusive goal in general [31], it has recently been shown that accurate prediction of epitope presentation can significantly improve immunogenicity prediction [7]. Therefore, we downloaded the same dataset as used in Bulik-Sullivan et al [7], which was collected from four previous publications [25–28] that consists of 26 SNVs and 31 neoantigens with pre-existing T-cell responses among 2,023 assayed single nucleotide variants (SNVs) from 17 patients, to evaluate our methods against others. We first asked if the increased accuracy of EPIC over other publicly available methods on epitope presentation could translate into better immunogenicity prediction. To quantify this, we plotted the precision recall curves (PRCs) by ranking all SNVs scored by different methods, where the score of a SNV is the maximum score of all possible epitopes generated by the SNV. EPIC' area under PRC (AUPRC) was more than twice that of other methods, and kept advantages when we applied expression value threshold of 1 or 2 TPM to other methods' prediction (Figure 4B and Figure S3). We also compared EPIC with EDGE at the level of ranking mutations and minimal epitopes (the recognized 8-11-mers overlapping the mutation) as originally done in Bulik-Sullivan et al [7]. EPIC is very comparably to EDGE on these two evaluations, especially for the top 10 and top 20 ranked mutations and epitopes. When prioritized the mutations, the number of immunogenic SNVs ranked in the top 20, 10, and 5 were 18 (69.23%), 16 (61.54%%) and 11 (42.31%) for EPIC, and 19 (73.08%), 18 (69.23%) and 12 (46.15%) for EDGE (Figure 4C). When prioritized the minimal epitopes, EPIC ranked 15 (48.39%), 12 (38.71%) and 5 (16.13%) CD8+ recognized epitopes in the top 20, 10 and 5. Therefore, despite being a much simpler model trained with less data comparing to EDGE, EPIC is also quite valuable for immunogenicity prediction in neoantigen-based cancer immunotherapies and is a significant improvement over other publicly available methods.

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

Having established the superiority of EPIC over other publicly available methods, we now provide a detailed example of how to incorporate new MS data into EPIC to expand its supported HLA alleles. In this example, the allele that we are particularly interested in is HLA-A*33:03, which appears often in the Chinese population but with no MS or binding affinity data available prior to this publication. 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 [32] analysis for the peptides in this cluster, together with peptides of HLA-*A02:01 and HLA-A*11:01 in EPIC'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 belongs to a different supertype. Based on these evidences, we confidently assigned this cluster to HLA-A*33:03. By inputting these HLA-A*33:03-specific peptides to EPIC, it can learn the corresponding PSSM and length distribution and add HLA-A*33:03 to its supported alleles. The overall work flow is illustrated in Figure 5. Similarly, we also took advantage of the MS data recently released by the EDGE paper [7] to add three more supported alleles (HLA-A*33:01, HLA-B*13:02, HLA-B*15:03) to EPIC, making the total number of supported HLA alleles to be 66.

Discussions

Traditionally, mixed-allele MS data has been deconvolved based on affinity predictions by pan-specific methods such as NetMHCpan, with slightly varying strategies adopted by different groups [12,13,33]. A notable drawback of such an approach is that it depends on the accuracy of affinity prediction to different HLA alleles, which could be particularly imprecise for those with limited or no training data. Unfortunately, this often coincides with the previously unsupported alleles that need to be added to EPIC, such as the A33:03 allele in the above example. Furthermore, peptides presented by MHC molecules *in vivo* may have different sequence properties than those measured by binding assays *in vitro*, thus some informative MS profiled peptides could be excluded by affinity-based deconvolution. We chose to use GibbsCluster, an unsupervised learning method that can handle variable length peptides, in order to provide unbiased training data to EPIC. To confirm that GibbsCluster performed adequately for this purpose, we carried out several analyses on the mixed-allele MS data in Pearson et al [13], using the mono-allele MS data in Abelin et al [12] as the reference. For a number of alleles in Pearson et al [13] that has corresponding mono-allele data in Abelin et al [12], we used both affinity-based method as described in Pearson et al [13] and GibbsCluster [19](with default parameters) to do deconvolution. As shown in Figure S5, both deconvolution methods generated motifs that were very similar to those obtained from mono-allele data, as measured by motif distances defined by Bassani-Sternberg et al [9] . By using both versions of deconvolved MS data to train simple PSSM models and tested on the corresponding mono-allele data, they also achieved very similar 0.1% PPVs (Figure S6). However, in all cases, GibbsCluster was able to retain more peptides in each cluster, sometimes considerably more (Table S2), and this facilitates training more accurate epitope presentation models. The deep learning-based EDGE method was also trained on a mixture of mono-allele and mixed-allele MS data, but it did not explicitly perform deconvolution before model training. Instead, it incorporated the allele information in an integrated model and implicitly performed “soft” clustering during training. While such an approach is undoubtedly more sophisticated, deconvolution as an intermediate step helps to provide intuitions on allele-specific binding properties. We believe this could be an advantage of EPIC, especially since the performances of EPIC and EDGE don’t differ that much, as discussed next.

The recently developed EDGE method has been shown to be almost an order of magnitude better than affinity-based models (mainly MHCflurry v1.2.0) on epitope presentation prediction, but it should be noted that the evaluation criterion of Bulik-Sullivan et al [7] was different from what we used here. While we used the more popular 0.1% PPV, i.e., positive to negative ratio or prevalence of 1:1,000, the EDGE paper used prevalence of 1:2,500 to 1:10,000 instead. In order to make a better comparison, we also evaluated EPIC on PPV at 40% recall with a 1:2,500 to 1:10,000 prevalence using our test data. As shown in Figure S7, the advantage of EPIC becomes more substantial and it achieved a more than 10-fold improvement over MHCflurry with 1:10,000 prevalence. Combined with the fact and EPIC and EDGE performed similarly on the same immunogenicity test data, we concluded that the performances of EPIC and EDGE should be quite comparable, although we were not able to run EDGE directly. EPIC is a much simpler model comparing to EDGE, but they are both trained on large-scale MS and RNA-seq datasets. This also implies that the performance gain achieved by EDGE should to a large degree attributed to the large-scale training data, which might be even more important than its deep learning framework. Nevertheless, using more advanced machine learning techniques to better take advantage of even larger training sets and account for more contributing factors to epitope presentation is clearly an important future research direction of EPIC. However, immunogenicity prediction in a clinical setting depends on more than epitope presentation [34–36], and we believe it is important to account for the extra factors contributing to immunogenicity as well. To that end, community efforts to generate more immunogenicity data and improved experimental techniques to profile TCR and peptide-MHC interactions [37,38] might lead to larger datasets and deeper biological understanding that can eventually enable accurate predictions of immunogenicity.

Conclusion

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

List of abbreviations

HLA: human leukocyte antigen
MS: Mass spectrometry
PSSMs: position score specific matrixes
LR: logistic regression
EPIC: Epitope Presentation Integrated prediCtion
MHC: major histocompatibility complex
TCRs: T-cell receptors
PMF: probability mass function
PPV: positive predictive value
TMG: tandem minigene
PRCs: precision recall curves

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The MS data of HCC4006 generated during the current study are available in the CNGB Nucleotide Sequence Archive (CNSA: https://db.cngb.org), accession number: XXXXX

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, GL, JL implemented the EPIC 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

None

Reference

1. Sahin U, Derhovanessian E, Miller M, Kloke BP, Simon P, Löwer M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature. Nature Publishing Group; 2017;547:222–6.
2. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature. 2017. p. 217–21.
3. Hilf N, Kuttruff-Coqui S, Frenzel K, Bukur V, Stevanović S, Gouttefangeas C, et al. Actively personalized vaccination trial for newly diagnosed glioblastoma. Nature [Internet]. 2018; Available from: http://www.nature.com/articles/s41586-018-0810-y
4. Keskin DB, Anandappa AJ, Sun J, Tirosh I, Mathewson ND, Li S, et al. Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. Nature. 2018;
5. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol [Internet]. 2017;ji1700893. Available from: http://www.jimmunol.org/lookup/doi/10.4049/jimmunol.1700893
6. Gfeller D, Bassani-Sternberg M. Predicting antigen presentation-What could we learn from a million peptides? Front Immunol. 2018;9:1–17.

7. Bulik-Sullivan B, Busby J, Palmer CD, Davis MJ, Murphy T, Clark A, et al. Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. Nat Biotechnol 2018 [Internet]. Nature Publishing Group; 2018; Available from: https://www.nature.com/articles/nbt.4313
8. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. Cell Syst. Elsevier Inc.; 2018;7:129–132.e4.
9. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO, et al. Deciphering HLA motifs across HLA peptidomes correctly predicts neo-antigens and identifies allostery in HLA specificity. bioRxiv. 2017;1–30.
10. Gfeller D, Guillaume P, Michaux J, Pak H-S, Daniel RT, Racle J, et al. Peptide length distribution and multiple specificity in naturally presented HLA-I ligands. bioRxiv. 2018;1–36.
11. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M. Mass Spectrometry of Human Leukocyte Antigen Class I Peptidomes Reveals Strong Effects of Protein Abundance and Turnover on Antigen Presentation. Mol Cell Proteomics. 2015;14:658–73.
12. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. Immunity. 2017;46:315–26.
13. Pearson H, Daouda T, Granados DP, Durette C, Bonneil E, Courcelles M, et al. MHC class I – associated peptides derive from selective regions of the human genome. J Clin Invest. 2016;126:1–12.
14. Di Marco M, Schuster H, Backert L, Ghosh M, Rammensee H-G, Stevanović S. Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction Matrices. J Immunol [Internet]. 2017;ji1700938. Available from: http://www.jimmunol.org/lookup/doi/10.4049/jimmunol.1700938
15. Shao W, Pedrioli PGA, Wolski W, Scurtescu C, Schmid E, Vizcaíno JA, et al. The SysteMHC Atlas project. Nucleic Acids Res. 2017;1–11.
16. Trolle T, McMurtrey CP, Sidney J, Bardet W, Osborn SC, Kaever T, et al. The Length Distribution of Class I–Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele–Specific Binding Preference. J Immunol. 2016;196:1480–7.
17. Zhou F, Cao H, Zuo X, Zhang T, Zhang X, Liu X, et al. Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. Nat Genet. 2016;
18. Liu G, Li D, Li Z, Qiu S, Li W, Chao C, et al. PSSMHCpan: a novel PSSM based software for predicting class I peptide-HLA binding affinity. Gigascience. 2017;1–26.
19. Andreatta M, Alvarez B, Nielsen M. GibbsCluster: Unsupervised clustering and alignment of peptide sequences. Nucleic Acids Res. 2017;45:W458–63.
20. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018.
21. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;
22. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. Bioinforma Impact Accurate Quantif Proteomic Genet Anal Res. 2014.
23. Breiman L. Stacked regressions. Mach Learn [Internet]. 1996;24:49–64. Available from: http://link.springer.com/10.1007/BF00117832

24. Pedregosa FABIANPEDREGOSA F, Alexandre Gramfort N, Michel V, Thirion BERTRANDTHIRION B, Grisel O, Blondel M, et al. Scikitlearn: Machine Learning in Python Gaël Varoquaux. J Mach Learn Res. 2011;
25. Gros A, Parkhurst MR, Tran E, Pasetto A, Robbins PF, Ilyas S, et al. Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. Nat Med [Internet]. Nature Publishing Group; 2016;22:433–8. Available from: http://dx.doi.org/10.1038/nm.4051
26. Tran E, Ahmadzadeh M, Lu YC, Gros A, Turcotte S, Robbins PF, et al. Immunogenicity of somatic mutations in human gastrointestinal cancers. Science (80-). 2015;350:1387–90.
27. Strønen E, Toebes M, Kelderman S, Van Buuren MM, Yang W, Van Rooij N, et al. Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. Science (80-). 2016;
28. Zacharakis N, Chinnasamy H, Black M, Xu H, Lu YC, Zheng Z, et al. Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. Nat Med. 2018;24:724–30.
29. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun. 2014;
30. The M, MacCoss MJ, Noble WS, Käll L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. J Am Soc Mass Spectrom. 2016;
31. The problem with neoantigen prediction. Nat Biotechnol [Internet]. 2017;35:97–97. Available from: http://www.nature.com/doifinder/10.1038/nbt.3800
32. Van Der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;
33. Caron E, Espona L, Kowalewski DJ, Schuster H, Ternette N, Alpízar A, et al. An open-source computational and data resource to analyze digital maps of immunopeptidomes. Elife. 2015;4:1–17.
34. Assarsson E, Sidney J, Oseroff C, Pasquetto V, Bui H-H, Frahm N, et al. A quantitative analysis of the variables affecting the repertoire of T cell specificities recognized after vaccinia virus infection. J Immunol. 2007;
35. Calis JJA, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, et al. Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. PLoS Comput Biol. 2013;
36. McGranahan N, Furness AJS, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science (80-). 2016;
37. Bentzen AK, Marquard AM, Lyngaa R, Saini SK, Ramskov S, Donia M, et al. Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. Nat Biotechnol. 2016;
38. Bentzen AK, Such L, Jensen KK, Marquard AM, Jessen LE, Miller NJ, et al. T cell receptor fingerprinting enables in-depth characterization of the interactions governing recognition of peptide–MHC complexes. Nat Publ Gr. 2018;

Figure legend

Fig. 1 Construction of EPIC_s and EPIC. a) Construction of allele and tissue-specific EPIC (EPIC_s).

b) Saturation analysis of PSSM models.

c) Construction of EPIC.

Fig. 2 Relationship among weights and biases of 21 EPIC_s models. a) Boxplot for feature weights and biases of EPIC_s models: PSSMs contributed most to the prediction, followed by expressions and peptide lengths.

b) Hive plots demonstrating the similar inter-relationships among the feature weights and biases of different EPIC_s models, where the negative biases have been inversed for plotting.

Fig. 3 Predictive performance comparison of EPIC and EPIC_s. a) Performance comparison of EPIC and EPIC_s across different alleles originating from the same B cell line.

b) Gene expression profiles of different cell lines used in EPIC and EPIC_s; genes presenting peptides in at least one cell line were selected.

c-f) Performance comparison of EPIC and EPIC_s for the same allele originating from different cell lines.

Fig 4. Predictive performance comparison of EPIC and other methods. a) Predictive performance of EPIC and other publicly available methods on independent MS dataset. The evaluation dataset was obtained from four publications that contained a total of 30 MS datasets, including 16 different tissue origins and 17 HLA alleles. The mean 0.1% PPV of these 30 datasets were displayed. EPIC significantly outperformed other methods with different expression filtering thresholds (TPM > 0, > 1, > 2), and pairwise P-values (paired t-test) for TPM > 2 were shown in the plot.

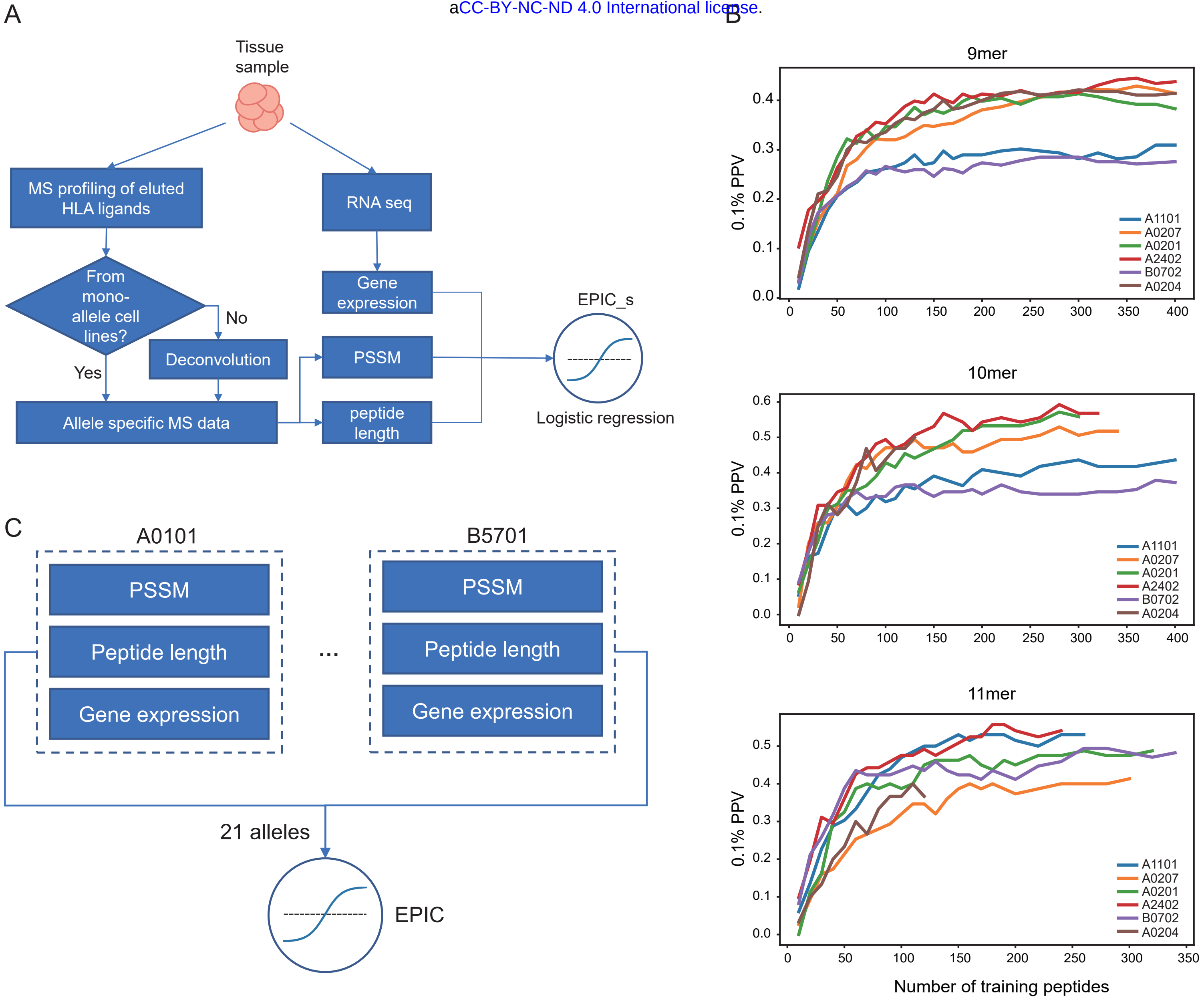
b). Precision-recall curves of EPIC and other publicly available methods on T cell responses data. Other methods' predictions were made with expression threshold of TPM > 0. Comparisons of EPIC and other methods with TPM > 1 and TPM > 2 were displayed in Fig S2.

c) Predictive performance of EPIC and EDGE on immunogenic SNVs. EDGE's performance was recalculated according to the supplementary data 3c of the original paper.

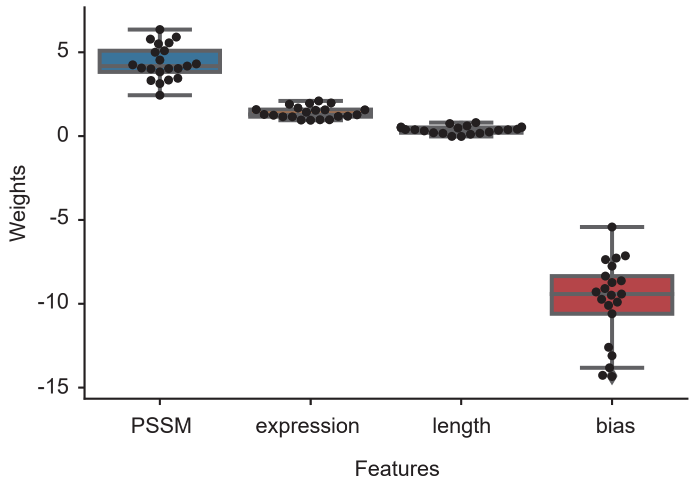
d) Predictive performance of EPIC and EDGE on minimal epitopes. EDGE's performance was taken from Figure 4 of the original paper.

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

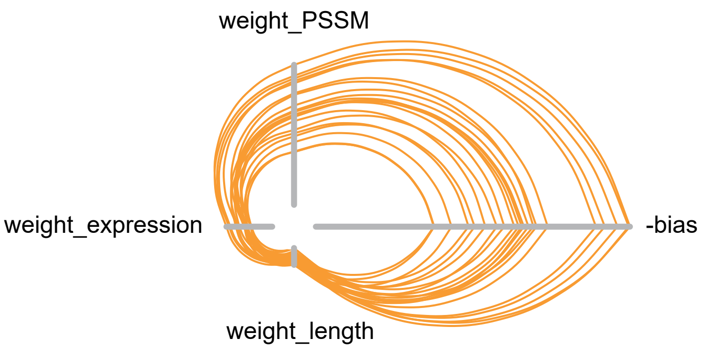
specific peptides to EPIC, which could learn PSSM and length distributions automatically, and resulted in an updated version of EPIC that could support the newly added allele.

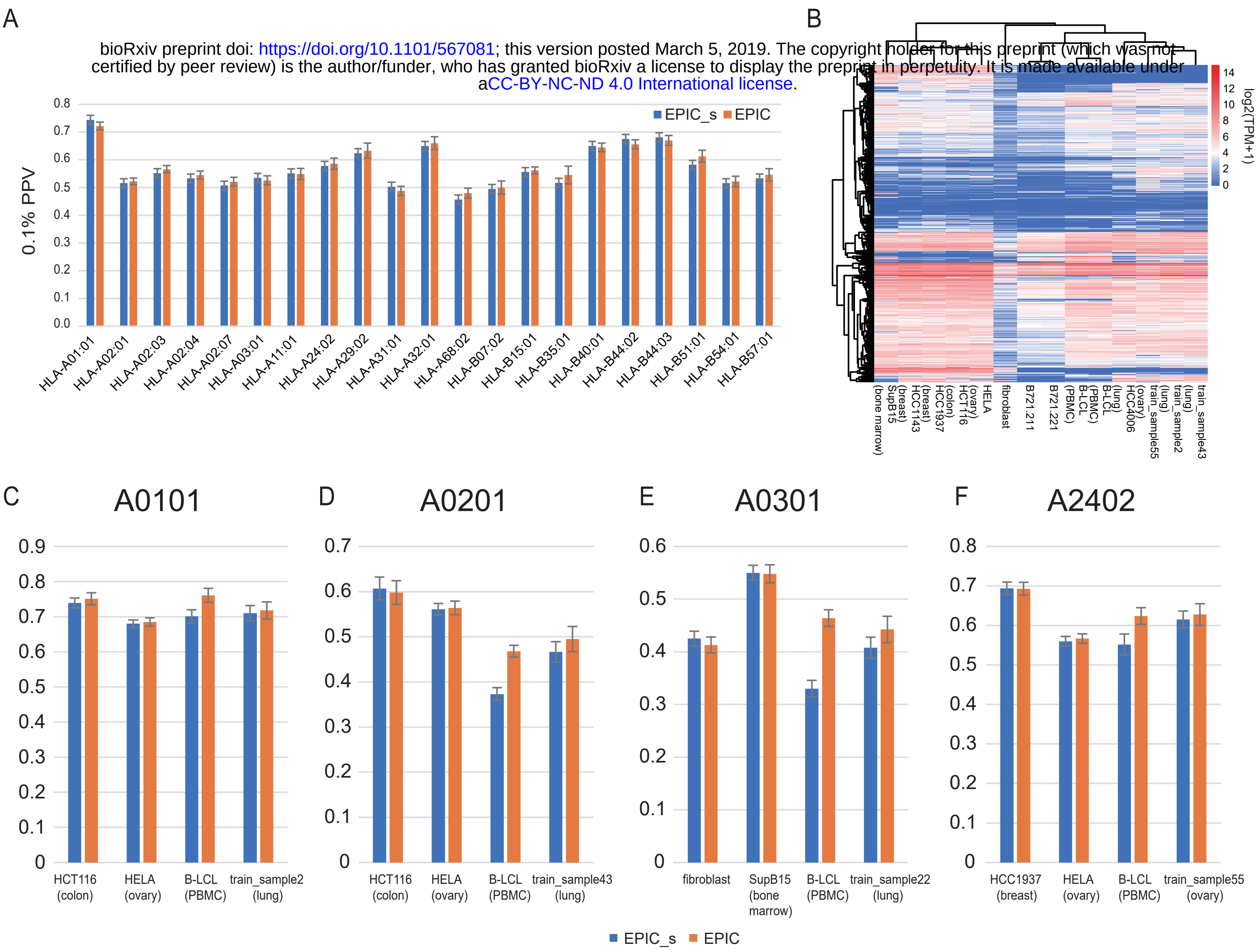

A
Tissue sample
MS profiling of eluted HLA ligands
RNA seq
From mono-allele cell lines?
No
Yes
Deconvolution
Allele specific MS data
Gene expression
PSSM
peptide length
EPIC_s
Logistic regression
B
9mer
10mer
11mer
0.1% PPV
Number of training peptides
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
200
250
300
350
400
A1101
A0207
A0201
A2402
B0702
A0204
C
A0101
B5701
PSSM
Peptide length
Gene expression
...
21 alleles
EPIC

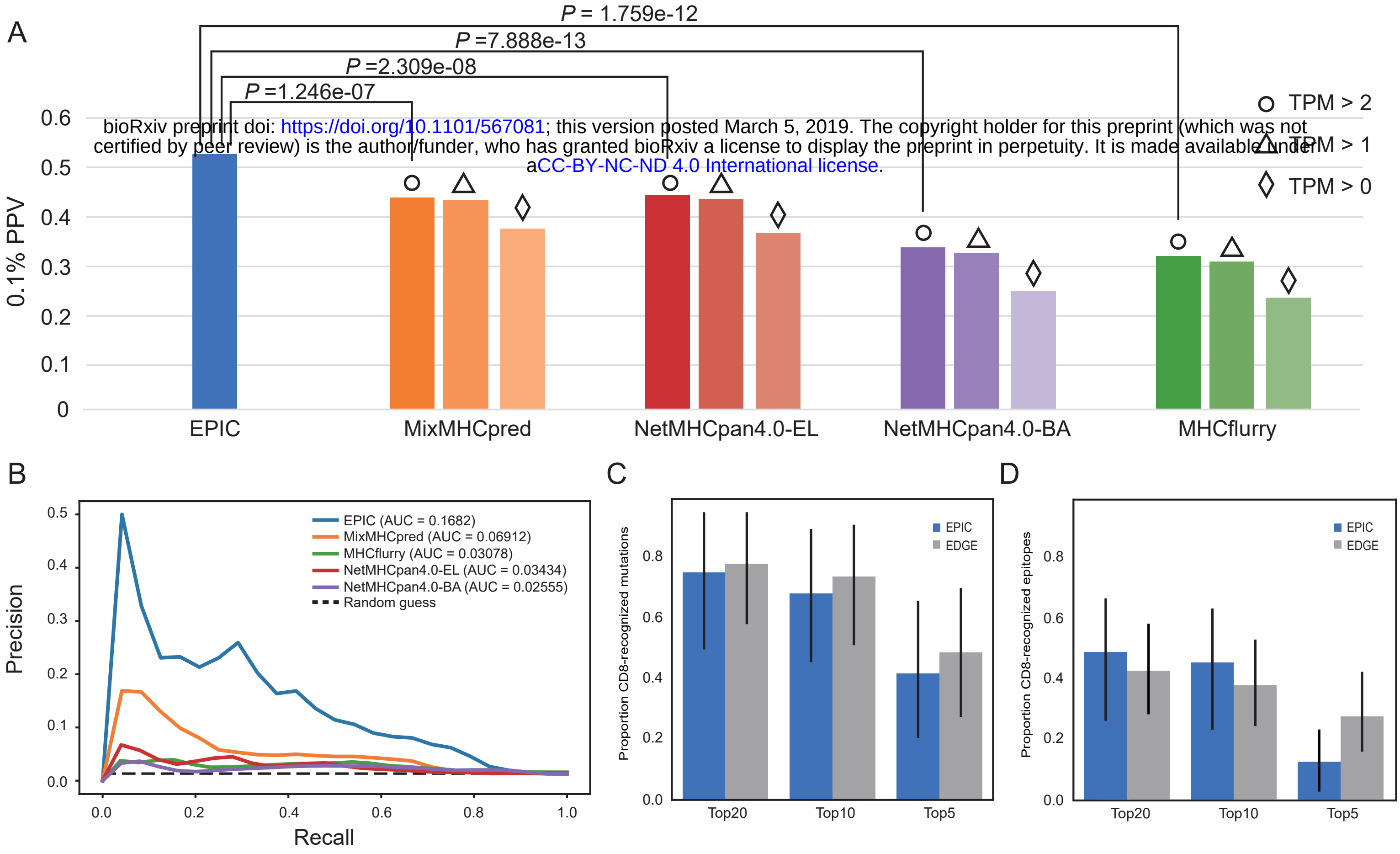
A

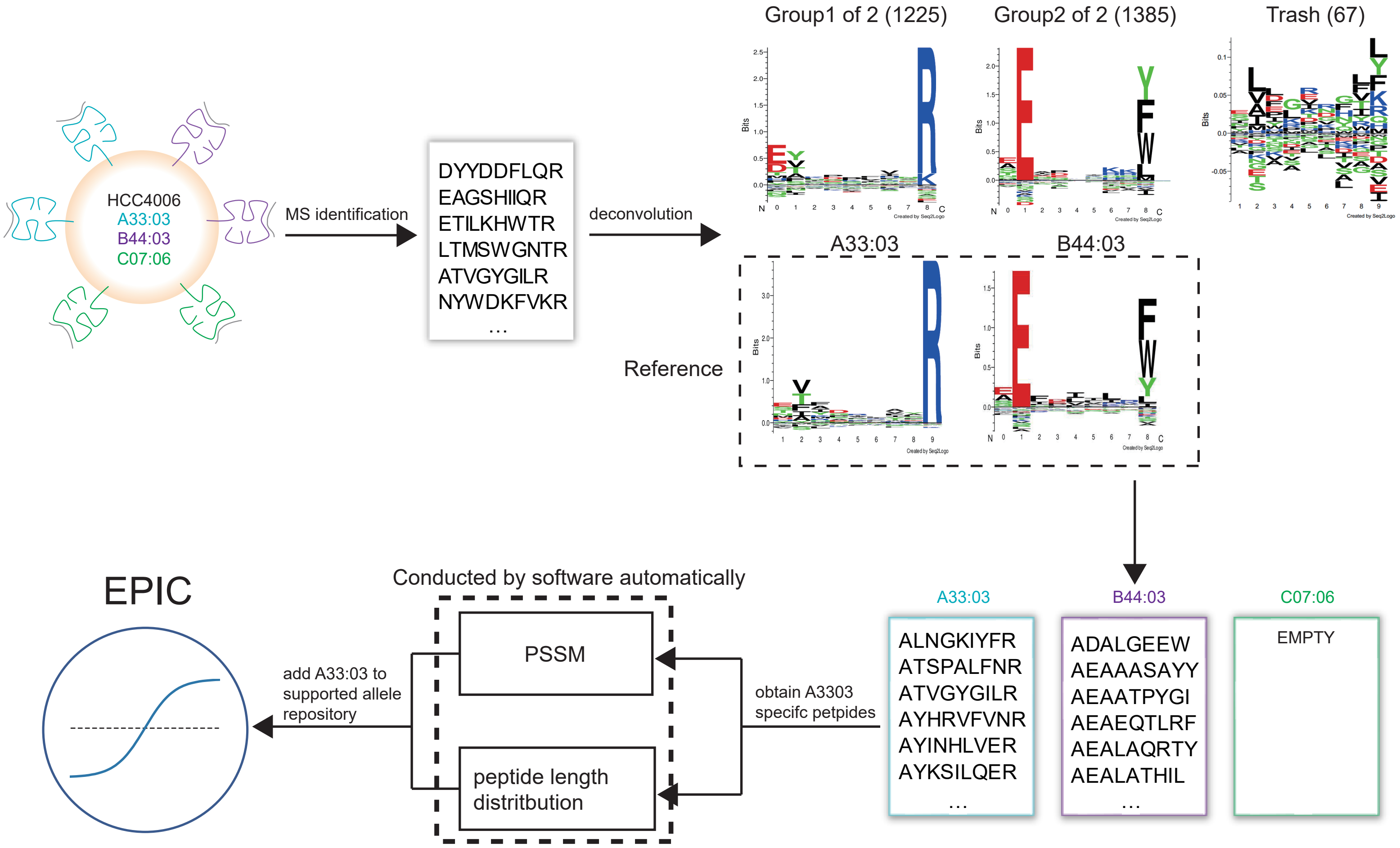

5
0
-5
-10
-15
Weights
PSSM
expression
length
bias
Features

B


weight_PSSM
weight_expression
-bias
weight_length


A
B
0.1% PPV
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
EPIC_s
EPIC
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
log2(TPM+1)
14
12
10
8
6
4
2
0
SupB15 (bone marrow)
HCC1143 (breast)
HCC1937 (breast)
HCT116 (colon)
HELA (ovary)
fibroblast
B721.211
B721.221
B-LCL (PBMC)
B-LCL (PBMC)
HCC4006 (lung)
train_sample55 (ovary)
train_sample2 (lung)
train_sample43 (lung)
C
A0101
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
HCT116 (colon)
HELA (ovary)
B-LCL (PBMC)
train_sample2 (lung)
D
A0201
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
HCT116 (colon)
HELA (ovary)
B-LCL (PBMC)
train_sample43 (lung)
E
A0301
0.6
0.5
0.4
0.3
0.2
0.1
0
fibroblast
SupB15 (bone marrow)
B-LCL (PBMC)
train_sample22 (lung)
F
A2402
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
HCC1937 (breast)
HELA (ovary)
B-LCL (PBMC)
train_sample55 (ovary)
EPIC_s
EPIC


A
P = 1.759e-12
P =7.888e-13
P =2.309e-08
P =1.246e-07
○ TPM > 2
△ TPM > 1
◊ TPM > 0
0.1% PPV
0.6
0.5
0.4
0.3
0.2
0.1
0
EPIC
MixMHCpred
NetMHCpan4.0-EL
NetMHCpan4.0-BA
MHCflurry
B
Precision
Recall
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.2
0.4
0.6
0.8
1.0
EPIC (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
0.8
0.6
0.4
0.2
0.0
EPIC
EDGE
Top20
Top10
Top5
D
Proportion CD8-recognized epitopes
0.8
0.6
0.4
0.2
0.0
EPIC
EDGE
Top20
Top10
Top5


Group1 of 2 (1225)
Group2 of 2 (1385)
Trash (67)
HCC4006
A33:03
B44:03
C07:06
MS identification
DYYDDFLQR
EAGSHIIQR
ETILKHWTR
LTMSWGNTR
ATVGYGILR
NYWDKFVKR
…
deconvolution
A33:03
B44:03
Reference
Bits
Created by Seq2Logo
EPIC
Conducted by software automatically
PSSM
peptide length distritbution
add A33:03 to supported allele repository
obtain A3303 specifc petpides
A33:03
ALNGKIYFR
ATSPALFNR
ATVGYGILR
AYHRVFVNR
AYINHLVER
AYKSILQER
…
B44:03
ADALGEEW
AEAAASAYY
AEAATPYGI
AEAEQTLRF
AEALAQRTY
AEALATHIL
…
C07:06
EMPTY