Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size
   Describe how sample size was determined.
   No statistical methods were used to predetermine sample sizes for the mass spectrometry and prospective NSCLC T-cell sections. Sample sizes were based on availability of clinical specimens and lab processing throughput. The retrospective T-cell response data were compiled from earlier publications.

2. Data exclusions
   Describe any data exclusions.
   No exclusion was applied

3. Replication
   Describe whether the experimental findings were reliably reproduced.
   No replicates were performed for sequencing and mass spectrometry portion. Replicate T-cell assays were performed when sufficient cells were available. All attempts at replication were successful. Methods for data analysis are described and can be reproduced.

4. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups.
   The experiments were not randomized. Randomization and control of covariates are not relevant to this study, as this is not an attempt at causal inference, but rather a study developing a classifier.

5. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   The study is not blinded. Samples were not allocated into groups; instead, we aimed to predict presentation or T-cell recognition of HLA peptides for each sample.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.
6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

<table>
<thead>
<tr>
<th>n/a</th>
<th>Confirmed</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>The exact sample size ((n)) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)</td>
</tr>
<tr>
<td>✗</td>
<td>A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly</td>
</tr>
<tr>
<td>✗</td>
<td>A statement indicating how many times each experiment was replicated</td>
</tr>
<tr>
<td>✗</td>
<td>The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)</td>
</tr>
<tr>
<td>✗</td>
<td>A description of any assumptions or corrections, such as an adjustment for multiple comparisons</td>
</tr>
<tr>
<td>✗</td>
<td>The test results (e.g. (P) values) given as exact values whenever possible and with confidence intervals noted</td>
</tr>
<tr>
<td>✗</td>
<td>A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)</td>
</tr>
<tr>
<td>✗</td>
<td>Clearly defined error bars</td>
</tr>
</tbody>
</table>

*See the web collection on [statistics for biologists](https://www.nature.com/articles/nbt0210-1285) for further resources and guidance.*

7. Software

Describe the software used to analyze the data in this study.

All software used is publicly available and described in the methods section. For NGS analysis: BWA-MEM 0.7.13-r1126, STAR 2.5.1b, RSEM 1.2.31, Picard 2.7.1, GATK 3.5-0, FreeBayes 1.0.2, MuTect 1.3-0, SnpEff 4.2, Optitype 3.1-0. For mass spectrometry analysis: Crux 3.1 Comet 2016.01 rev. 1, Crux 3.1 Percolator 3.01.nightly-13-655e4c7. For machine learning: Scikit-Learn 0.18.0, Theano 0.9.0, Keras 2.0.4. EDGE model code is provided in the supplementary information. For HLA-peptide binding affinity prediction: MHCflurry 1.2.0. For binding motifs, weblogolib Python API 3.5.0.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods guidance for providing algorithms and software for publication* provides further information on this topic.
8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

- Custom-made, recombinant lyophilized peptides (JPT Peptide Technologies (Berlin, Germany) or Genscript (Piscataway, NJ, USA)
- Cryopreserved HLA-typed PBMCs from healthy donors (Precision for Medicine, Gladstone, NJ, USA or Cellular Technology, Ltd., Cleveland, OH, USA)
- Fresh blood samples (Research Blood Components, Boston, MA, USA)
- Leukopaks (AllCells, Boston, MA, USA)
- Patient PBMCs were processed at local clinical processing centers according to local clinical standard operating procedures (SOPs) and IRB approved protocols.

Normal/germline DNA derived from blood were isolated using Qiagen DNeasy columns (Hilden, Germany) following manufacturer recommended procedures. DNA and RNA from tissue specimens were isolated using Qiagen Allprep DNA/RNA isolation kits following manufacturer recommended procedures. The DNA and RNA were quantitated by Picogreen and Ribogreen Fluorescence (Molecular Probes). DNA sequencing libraries were generated by acoustic shearing (Covaris, Woburn, MA) followed by DNA Ultra II (NEB, Beverly, MA) library preparation kit following the manufacturers recommended protocols. Tumor RNA sequencing libraries were generated by heat fragmentation and library construction with RNA Ultra II (NEB). The resulting libraries were quantitated by Picogreen (Molecular Probes).

Exon enrichment for both DNA and RNA sequencing libraries was performed using xGEN Whole Exome Panel (Integrated DNA Technologies). The captured libraries were minimally amplified by PCR and quantitated by NEBNext Library Quant Kit (NEB). Captured libraries were pooled at equimolar concentrations and clustered using the c-bot (Illumina) and sequenced at 75 base paired-end on a HiSeq4000 (Illumina).

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

- IFNgamma Enzyme Linked Immunospot (ELISpot) assay
  - Mabtech, Cincinatti, OH, USA
  - Anti-human IFN-γ mAb 1-D1K, purified, catalog number 3420-3-1000
    - Lots: 93.2 and 93.3
  - Anti-human IFN-γ mAb 7-B6-1, biotinylated, catalog number 3420-6-1000
    - Lot: 50.1 and 52.1
- Granzyme B ELISA
  - Granzyme B DuoSet® ELISA (R & D Systems, Minneapolis, MN)
    - Kit lot no: P136555
- MSD multiplex assay
  - MSD U-PLEX Biomarker assay (catalog number K15067L-2)
  - Antibody lots:
    - MSD capture antibodies
    - TNFa: B00U0004A
    - IL-2: 3-3035-021003-A^SET00010A
    - IL-5: 3-3035-021011-A^SET00018A
  - MSD detection antibodies:
    - TNFa: D00U0004A
    - IL-2: 3-3035-021003-A^SET00010A
    - IL-5: 3-3035-021011-A^SET00018A

Anti-human HLA-A,B,C clone W632 from BioLegend based on catalog number 311402. Special bulk order Lot No. B233326 at a final concentration of 5mg/mL. Validation performed by vendor.
10. Eukaryotic cell lines
   a. State the source of each eukaryotic cell line used.
   H128, H122, H2009, H2126, Colo829 and their normal donor matched control cell lines BL128, BL2122, BL2009, BL2126 and Colo829BL were all purchased from ATCC (Manassas, VA).
   b. Describe the method of cell line authentication used.
   Cell lines were purchased from ATCC and verified by checking the label on the frozen vial against the order and vendor Certificate of Analysis.
   c. Report whether the cell lines were tested for mycoplasma contamination.
   Cell lines tested negative for mycoplasma contamination.
   d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.
   No commonly misidentified cell lines were used.

Arrows and human research participants
Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals
   Provide details on animals and/or animal-derived materials used in the study.
   No animals were used in the study.

Policy information about studies involving human research participants

12. Description of human research participants
   Describe the covariate-relevant population characteristics of the human research participants.
   Patient tumor specimens used for model training (Supplementary Data 1) came predominantly from Caucasian patients with resected CRC, NSCLC, or OC. Patient tumor samples and PBMCs for T cell studies (Supplementary Data 5) came from 9 patients aged 51-90 (7 White, 1 Hispanic, 1 African American) with advanced lung cancer undergoing systemic therapy.