Widespread immunogenic poly-epitope frameshift mutations in microsatellite unstable tumors

Vladimir Roudko[1,5 †], Cansu Cimen Bozkus[1 †], Theofano Orfanelli[3], Stephanie V. Blank[3], Benjamin Greenbaum[1,2,4,5 ‡], Nina Bhardwaj[1 ‡ *].

[1]*Department of Hematology and Medical Oncology, Icahn School of Medicine at Mount Sinai Hospital, New York, New York, USA;*

[2]*Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai Hospital, New York, New York, USA;*

[3]*Department of Obstetrics, Gynecology and Reproductive Science, Icahn School of Medicine at Mount Sinai Hospital, New York, New York, USA;*

[4]*Department of Pathology, Icahn School of Medicine at Mount Sinai Hospital, New York, New York, USA;*

[5]*Center for Computational Immunology, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai Hospital, New York, New York, USA.*

[†]*These authors contributed equally to the work*

[‡]*These authors shared the last authorship*

[*]*Correspondence: nina.bhardwaj@mssm.edu (N.B.)*

Manuscript

Abstract

Microsatellite instability-high (MSI-H) tumors are an important model system for evaluating neoantigen-based immunotherapies given their high tumor mutation burden and response to checkpoint blockade. We identified tumor-specific, frameshift peptides, encoding multiple epitopes that originated from indel mutations shared among patients with MSI-H endometrial, colorectal and stomach cancers. Epitopes derived from these shared frameshifts have high population occurrence rates, wide presence in many tumor subclones and are predicted to bind to the most frequent HLA alleles in the TCGA MSI-H patient cohorts. Neoantigens arising from these mutations are more dissimilar to both self and viral antigens, indicating the creation of peptides, that, when translated, can present truly novel antigens to the immune system. Finally, we validated the immunogenicity of common frameshift peptides from MSI-H endometrial patients in an array of T cell stimulation experiments, using peripheral blood mononuclear cells isolated from healthy donors. Our study describes for the first time the widespread occurrence and strong immunogenicity of tumor-specific antigens, derived from shared frameshift mutations in MSI-H cancer and Lynch syndrome patients, suitable for the design of common preventive "off-the-shelf" cancer vaccines.

Manuscript

Introduction

Genetic alterations in tumor genomes that encode novel stretches of amino acids compared to normal cells are a potential source of immunogenic tumor-specific epitopes, commonly referred to neoantigens. Total neoantigen burden, (the sum of neoantigens predicted to be expressed by a tumor), has been demonstrated to be an independent proxy for response to anti-programmed cell death protein 1 (PD-1) immunotherapy[1–4]. However, determining neoepitopes in individual tumor sample remains fraught with uncertainties, such as the lack of congruence between selection of somatic variants and immunogenic neoantigens. Due to the presence of high loads of tumor-specific antigens, and strong effector T cell infiltration, microsatellite instability-high (MSI-H) tumors are an important model system for evaluating neoantigen-based immunotherapy in therapeutic and protective settings. The MSI-H tumor phenotype arises from defective DNA repair mechanisms due to a loss of the mismatch repair (MMR) activity. MSI-H is typically characterized by the variation of DNA length in microsatellite loci – units of one to ten mono-, di-, tri-, or tetra-nucleotides repeated multiple times[5]. In healthy cells these unpaired nucleotides are recognized and excised by MMR, but in MSI-H tumors they remain unrepaired. Some of these microsatellite regions are located in coding regions, where their destabilization may cause frameshift (fs-) mutations and thereby create a large somatic mutation burden, possessing a huge source of tumor-specific neoantigens.[4,5].

Inactivation of several MMR genes plays a key role in the acquisition of the MSI-H phenotype in hypermutated tumors[6–9]. This often happens at the later stages in tumorigenesis, and it originates by either deleterious somatic mutagenesis or epigenetic inactivation of MMR genes. This sporadic type of MSI-H tumors occurs in 10-40% of colorectal and endometrial cancers and is

Manuscript

70 mainly caused by biallelic hypermethylation of the MLH1 promoter[14,15]. Apart from somatic
71 inactivation of MMR genes during tumorigenesis, germline mutations within the same genes are
72 also found. Lynch syndrome, sometimes referred to hereditary nonpolyposis colorectal cancer
73 (HNCC), is an inherited, autosomal-dominant disorder characterized by germline non-
74 synonymous mutations in MRR genes, accounts for 3-5% of all colorectal and endometrial
75 cancers. The majority of patients have germline mutations in the MSH2 (~30%) and PMS2
76 (~70%) genes[12]. Subjects with Lynch syndrome suffer from a high predisposition to develop
77 cancer early life, with an 80% life time risk for colorectal or endometrial MSI-H cancers. Also
78 estimates suggest as many as 1 in every 300 people may carry Lynch syndrome-associated
79 germline alterations[13,14,15]. Epithelial tissues are primarily at risk of tumorigenesis, with
80 colorectal cancers being the most common (80% of HNCC patients). The second most common
81 cancer with a significant percentage of sporadic and hereditary-predisposed MSI-H type is
82 endometrial carcinoma (60% of HNCC patients). The MSI-H group accounts for up to 28.6% of
83 low-grade and 54.3% of high-grade endometrioid cancers. Other cancers such as bladder, gastric,
84 ovarian, small bowel and renal are also somewhat predisposed[16].
85
86 Most neoantigens that are predicted from point mutations are derived from patient-specific
87 passenger mutations, with most shared driver mutations creating rarely presented peptides[17].
88 Consequently, the former are preferentially applied in cancer vaccine platforms. However, the
89 MSI-H phenotype has several features which can be leveraged for common "off the shelf"
90 vaccine design: (1) a high-mutational burden in well-defined, limited sequence space –
91 microsatellite regions; (2) a restricted pattern of the mutations due to nucleotide insertions or
92 deletions; and (3) a high probability of shared indel mutations in protein coding genes, which

Manuscript

may induce formation of frameshift (fs-) peptide, which encodes multiple MHC-I restricted epitopes (poly-epitope fs-peptide) that are shared among multiple patients[18,19]. Based on this premise, we investigated fs-mutations in tumor genomes of MSI-H patients with colorectal, stomach and endometrial carcinomas and identified a high frequency of broadly shared, immunogenic, poly-epitope fs-peptides.

Results

MSI-H colorectal, stomach and endometrial patient cohorts have a high fs-load

Though tumor evolution is primarily viewed as driven by a random mutational process, there is accumulated evidence that some mutations are acquired non-randomly[40–43]. Considering the existing skewness in mutational process and high load of indel mutations in microsatellite regions within open reading frames of MSI-H tumors[23], we hypothesized MSI-H patients may share mutational events. While missense somatic mutations share limited similarity across multiple tumors[24,25], frameshift (fs-) mutations in microsatellite (MS) unstable regions are indeed likely to generate common fs-peptides (**Supplement Figure 1A**). Though previous studies have annotated somatic mutation load as well as the distribution and frequency of insertions/deletions on a pan-cancer scale[17,18,26] (**Supplement Figure 1B**), little is known about the frequency of shared mutational events. We examined mutational load in cancer cell lines from Cancer Cell Line Encyclopedia[27,28] (CCLE) and found that fs-mutations are more frequenty shared among multiple cancer cell lines, than missense mutations (**Supplement Figure 1C**).

The majority of TCGA tumors are microsatellite stable or their status is unknown. However, ~20%-30% of endometrial, colorectal and stomach adenocarcinomas are diagnosed as

Manuscript

116 microsatellite unstable, MSI-H (UCEC, COAD, STAD respectively, **Figure 1A**). In total, the
117 MSI-H population in TCGA accounts for 338 patients (**Figure 1B**). The fs-load, as determined
118 by fs-mutation count per each patient, is particularly high in a subset of UCEC, COAD and
119 STAD patients (**Figure 1C**). Consistent with previous studies, the majority of frameshifts stems
120 from nucleotide deletions, as determined by correlating patient's fs-load with insertion-to-
121 deletion ratio (**Figure 1D**). Finally, the majority of fs-enriched patients are segregated according
122 to the clinical MSI-H biomarker, indicating nearly perfect specificity/selectivity of this
123 biomarker in detecting indel-enriched tumor types (**Figure 1E**).
124

125 **Colorectal, stomach and endometrial MSI-H adenocarcinomas are enriched in shared poly-**
126 **epitope fs-peptides**

127 We then hypothesized that de-novo translated peptide segments – fs-peptides – within a newly-
128 defined codon sequence frame (either +1 or +2) downstream of the indel mutation, may possess
129 a significant source of immunogenic class I neoepitopes. If such fs-derived neoepitopes exist,
130 they may also exhibit a high rate of turnover and processivity: mRNAs that encode frameshift
131 mutations may be rapidly degraded through the nonsense-mediated decay (NMD) pathway,
132 which is accompanied by nascent peptide decay on the 80S ribosome[29–33]. While the expression
133 of fs-genes may be downregulated, the translated product is destabilized, potentially producing
134 short, presentable peptides at a higher rate[34,35].
135

136 To identify potentially immunogenic epitopes derived from fs-mutations, we developed a fs-
137 neoantigen calling pipeline (see **Online Methods**). Using this pipeline, we analyzed the
138 distribution of fs-mutations, fs-peptides and corresponding fs-epitopes in MSI-H UCEC, COAD

Manuscript

139 and STAD cohorts of TCGA patients. As expected, we found that many genes were commonly
140 mutated via indels in MS regions in all the three tumor types. Up to 80% of MSI-H COAD and
141 STAD patients had several commonly mutated genes as a result of indels, e.g. ACVR2A,
142 MIKI67, RPL22. More than 50% of MS-H UCEC patients shared a different set of genes
143 affected by frameshifts e.g. CASP5, MUC6, KMT2C. Expectedly, the frequency of shared fs-
144 peptides, derived from exactly the same fs-mutations, dropped, with only a few fs-peptides
145 shared by 40% of MSI-H UCEC, or 50% of MSI-H COAD/STAD patients. Finally, the
146 frequency of shared HLA epitopes in MSI-H UCEC patients approached 25% having 28
147 immunogenic fs-epitopes, and to 50% of COAD patients who had 29 shared epitopes (**Figure**
148 **2A, B**). Interestingly, colon and stomach MSI-H tumors had twice as many shared immunogenic
149 fs-events than endometrial MSI-H tumors (**Figure 2B**), which likely stems from the different
150 pathways of tumorigenesis, and from the profiles of mutated oncogene drivers.
151
152 To identify immunogenic fs-peptides with confidence, we developed a mutation ranking system
153 based on maximization of four parameters (**Figure 2C**). Firstly, we introduced a quality metric
154 for each called fs-mutation in which higher quality implies higher confidence that this mutation
155 is truly somatic (**Supplement Figure 2A**). Also, we analyzed the distribution fs-peptide length:
156 on average, the length of MSI-H frameshift peptides is 20-30 aminoacid (aa) residues, suggesting
157 that these peptides may encode multiple immunogenic epitopes per fs-mutation (**Supplement**
158 **Figure 2B**). Secondly, we maximized the number of putative neoepitopes per each fs-peptide:
159 epitopes, predicted for each fs-peptide across all patients, were pooled and the total number of
160 unique epitopes was determined. Thirdly, we grouped all HLA alleles predicted to bind those
161 neoepitopes and considered this to be an HLA-ligandome of each fs-peptide. Maximization of

Manuscript

162 this parameter allowed us to pick poly-allelic fs-peptides, covering as diverse a set of HLA
163 alleles in a population as possible. Finally, to include the population HLA allele frequency
164 parameter, we quantified total amount of antigen-allele interactions per frameshift. Together, we
165 ensured the prediction of fs-peptides that are most likely immunogenic, encode poly-epitopes,
166 bind to a broad spectrum of HLA-alleles and widely shared. We identified 9, 37 and 23 fs-
167 peptides derived from frequently shared indel mutations that encode poly-epitopes in
168 endometrial, colorectal and stomach MSI-H patients, respectively (**Figure 2C**). Counting the
169 number of epitopes derived from each fs-peptide in each patient showed a high level of presence
170 of these epitopes in MSI-H cohort of patients. Alternatively, plotting the number of epitopes
171 derived from frameshift sequence per each predicted HLA allele showed that many epitopes are
172 predicted to bind to the most frequently expressed HLA alleles in the TCGA patient cohort
173 (**Figure 2D**). Finally, among the detected fs-peptides, five were found in all MSI-H patients,
174 pointing to the possibility to develop an off-the-shelf MSI-H vaccine for these three tumor types:
175 SLC35F5, SEC31A, TTK, SETD1B and RNF43.
176

177 **Nine poly-epitope fs-peptides shared in MSI-H endometrial carcinoma originate from**
178 **clonal tumor-specific alleles and are correctly translated**

179 We next focused our attention on frameshift-derived neoantigens from MSI-H endometrial
180 carcinoma, as these have not been characterized to date. At least two neoepitopes from nine
181 selected fs-peptides are detected in >95% of the MSI-H UCEC TCGA patient cohort. Moreover,
182 binding profiles of the predicted epitopes cover almost all HLA-alleles of the same patient
183 population as well as targeting the majority of frequent HLA types (**Figure 2D**). Importantly,
184 only the "mixture" of MHC-I epitopes derived from all 9 peptides has the potential to reach a

Manuscript

185 significantly high representation of all HLA allele-epitope diversity, than each fs-peptide
186 separately (**Figure 3A**).
187
188 Additionally, we analyzed the tumor allele frequencies from the MSI-H UCEC patient cohort to
189 estimate the abundance of the selected nine shared frameshifts in tumor genomes. To do that, we
190 compared corresponding fs-allele frequencies in normal and tumor samples. As expected, fs-
191 alleles are barely detectable in normal tissues, but in tumor biopsies their frequencies rise up to
192 40% on average in tumor biopsies, indicating these mutations are often clonal (**Figure 3B**). This
193 suggests the 9-peptide mix has a potential to prime T cell responses against almost all malignant
194 cells. However, the high mutation rates of MSI-H tumors might decrease the probability of the
195 shared fs-peptides being translated (**Supplementary Figure 2C**). Taking this into consideration,
196 we assessed the conditional probability of shared fs-peptide being correctly translated given this
197 shared mutation has happened. For this purpose, we estimated all disruptive upstream and
198 downstream mutation frequencies using the MSI-H TCGA patients and calculated posterior
199 probabilities (**Supplementary Figure 2D**). Though the majority of MSI-H UCEC shared fs-
200 peptides has a high probability of being correct (>0.9), TTK and RNF43 frameshifts have an
201 increased chance of being inactivated, with a ~0.8 probability of being correct.
202
203 **The predicted nine shared poly-epitope fs-peptides are expressed in MSI-H UCEC patients**
204 To estimate the expression level of genes encoding shared fs-peptides, we analyzed RNA
205 expression derived from TCGA RNAseq samples of matched MSI-H patients (**Figure 3C-F**).
206 We performed unsupervised clustering of MSI-H patients by genes with predicted shared fs-
207 events, obtaining the patient and gene rankings according to the shared fs-load (**Figure 3C**). We

Manuscript

208 then plotted FPKM expression values of genes encoding shared fs-peptides, and ranked them
209 according to the previously obtained patient and gene rankings (**Figure 3C**). We observed no
210 correlation between RNA expression and shared fs-load, suggesting that frameshifted genes are
211 not selectively epigenetically silenced in tumors (**Figure 3D**). We next analyzed the expression
212 of nine shared fs-mutations in MSI-H UCEC patients. As two of the shared fs-mutations are
213 present in one gene, formally we detected 8 uniquely mutated genes, and six of these were
214 expressed at the RNA level from matching tumor samples (**Figure 3E**). To assess the expression
215 level of fs-alleles, we compared the normalized read count containing the indel with total amount
216 of reads covering the targeted genomic loci from MSI-H RNAseq samples and MSS samples as a
217 control (**Figure 3E)**. We detected statistically significant and robust expression of fs-alleles in
218 MSI-H patients (**Figure 3E**). Also, we showed representative RNAseq expression histograms,
219 covering the predicted shared frameshift in RNF43 gene, and derived indel RNAseq read
220 frequencies in MSI-H and MSS patients (**Figure 3F**). Thus, we confirmed that fs-alleles are
221 indeed expressed and detected in bulk RNAseq.
222
223 We then tested if high load of shared poly-epitope fs-mutations had a survival benefit within the
224 MSI-H patient cohort and/or may be skewed towards later tumor stage. This questions is
225 important for ongoing immunotherapy trials, where MSI-H becomes widely accepted as a
226 predictive biomarker to anti-PD-1 therapy[36,37], although deeper genetic differences between
227 patients may underlie the incomplete responsiveness. For this purpose, we performed Cox
228 regression and survival analyses of TCGA MSI-H UCEC patients stratified by shared fs-load, as
229 high and low, and analyzed tumor stage and patient age in the same strata. Patient age and tumor
230 stage were evenly represented in both, fs-high and fs-low MSI-H cohorts (**Supplemental Figure**

Manuscript

231 **3**). Finally, we did not detect any significant benefit in patients' survival based on shared fs-load
232 in any of the MSI-H tumor types (**Supplementary Figure 4**).
233
234 **Tumor fs-epitopes are more likely to be presented and are less similar to viral antigens**
235 **than missense derived epitopes**
236 We next examined the intrinsic properties of fs-derived epitopes when compared to missense-
237 derived epitopes and viral antigens. In previously published studies, similarity of tumor-derived
238 neoantigens to pathogen derived (viral) antigens has been shown to be predictive in the
239 checkpoint blockade immunotherapy setting[1,2]. First, we calculated the total amount of
240 neoantigens derived from missense mutations of MSI-H patients and compared it to fs-epitope
241 load. Despite the fact that the total frameshift and the missense mutation loads were similar, the
242 number of MHC-I epitopes per mutation were different: 4 epitopes per one frameshift and 2 per
243 one missense mutation on average (**Supplement Figure 5A**). This observation suggests, that fs-
244 mutations may be more immunogenic then missense due to an increased probability of
245 presentation. While most of missense-derived epitopes are, by definition, one aa different from a
246 self-peptide, the majority of fs-derived epitopes are unique, "non-host" peptide sequences and
247 hence exhibit minimal similarity to the human proteome (**Supplement Figure 3B**). This implies,
248 that fs-derived epitopes might have never been tolerized by the host immune system, and that the
249 frameshift-specific T-cells may have little or minimal autoreactivity.
250
251 We also compared these two epitope datasets with virus-derived antigens. At multiple search
252 parameters, the overall number of missense epitopes, when matched with viral ones, was higher,
253 than matched fs-epitopes (**Supplement Figure 3C**). We speculate this observation is due to the

Manuscript

254 overall viral adaptation to the human proteome and host T-cell epitopes, as viruses need to hijack
255 particular host functionalities in order to interact with the host cellular machinery as well as
256 escape host immune recognition. As a consequence of this finding, we conclude fs-epitopes
257 appear "less self' than either missense or virus derived epitopes.
258

259 **Predicted shared poly-epitope fs-peptides are detected and expressed in diverse set of**
260 **cancer cell lines from CCLE**

261 To validate the predicted shared fs-mutations in an external dataset, we verified the presence of
262 these mutations in the Cancer Cell Line Encyclopedia (CCLE) (**Figure 4**). 34 from 45 of
263 predicted shared poly-epitope fs-peptides were detected in multiple cancer cell lines, derived
264 from different tumor types. Diverse cancer cell lines had different numbers of shared fs-peptides,
265 e.g. intestine, endometrium, stomach and prostate cancer cell lines having 5-10 shared fs-
266 mutations per cell line, while hematopoietic, ovarian and lung cancer cell lines had 1-5 shared fs-
267 mutations per cell line (**Figure 4A**). The presence of predicted shared fs-mutations in the last
268 three tumor types suggests a broader occurrence of shared poly-epitope fs-peptides across
269 tumors, than initially suggested. The frequency of shared fs-positive cell lines per tissue type was
270 around 20, 40 and 60% of intestine, stomach and endometrial cell lines (Figure 4B).
271 Interestingly, the majority of fs-mutations differentially shared by MSI-H UCEC, COAD and
272 STAD TCGA tumors were evenly detected in endometrial, intestine and stomach cancer cell
273 lines (**Figure 4C**), suggesting epigenetic remodeling of certain chromatin areas making them
274 accessible for genome destabilization. Initially predicted in TCGA, shared fs-peptides were the
275 only fs-mutations statistically significantly shared in cancer cell lines than any other fs-mutation
276 derived from the same genes (**Figure 4D**). Fs-allele coverage analysis suggested that predicted

Manuscript

277 shared fs-mutations are clonal and present in shared fs-positive cancer cell lines at 30-50% allele
278 frequency on average (**Figure 4E**). Also, RNAseq data indicates the gene expression patterns are
279 unchanged upon acquiring shared frameshift (**Supplement Figure 6**). Overall, the CCLE dataset
280 analysis confirms the widespread occurrence of predicted shared fs-peptides in cancer.
281
282 **Nine poly-epitope fs-peptides shared in MSI-H UCEC patients are highly immunogenic**
283 To assess the immunogenicity of the nine predicted fs-peptides from MSI-H UCEC patient
284 cohort, we evaluated the T cell responses against each neopeptide using a T cell immunogenicity
285 assay developed in our laboratory that is designed to rapidly prime naïve T cells (see **Online**
286 **Methods**). In brief, long overlapping peptide (OLP) libraries spanning each fs-peptide were
287 designed (**Supplementary Figure 7**). Using these OLP pools, T cells from 15 randomly picked
288 healthy donors (HD) were primed and expanded. After expansion, the cells were stimulated with
289 the OLP pools and fs-peptide-specific T cell responses were evaluated by measuring IFN-γ
290 production using ELISPOT (**Figure 5A**). Results showed that each fs-peptide could elicit T cells
291 responses in a subset of subjects tested. Furthermore, some subjects had reactive T cells against
292 multiple fs-peptides (**Figure 5B-D**). Importantly, when combined, the fs-peptide-specific T cells
293 were significantly enriched in the subject cohort in accordance with the predictions detailed in
294 Figure 2D. We also characterized the fs-peptide-specific T cell responses in the same HD cohort
295 by intracellular staining (ICS). Results from both assays showed similar stimulation profiles.
296 Moreover, responses to fs-peptides were observed primarily in CD8+ T cells, reaching up to 10%
297 of T cells indicating strong priming to these neoantigens. (**Figure 5E-G**). In total, a majority of
298 HD (13 out of 15) responded to at least one fs-peptide. Importantly, the reactive T cells produced
299 TNF-α, in addition to IFN-γ, suggesting that fs-peptide-specific T cells are polyfunctional

Manuscript

300 (**Supplementary Figure 8**). Additionally, we synthesized control peptides (15-aa) for each fs-
301 peptide using their wild type sequence surrounding the fs-mutation site. Responses by HD T
302 cells to stimulation with WT OLP pool were not higher than the background (**Figure 5H**),
303 suggesting that the observed T cell responses were specific to fs-peptides. Next, we investigated
304 whether the fs-peptide-specific T cells responses that were observed in the HD cohort correlated
305 with the predicted high affinity epitope load. To determine the predicted epitope load, we
306 identified the HLA-I alleles of each subject by sequence-based HLA-I genotyping and
307 investigated the predicted binding affinity of epitopes from fs-peptides to each subject's unique
308 HLA. We found no significant correlation between the total epitope load per patient and
309 experimentally observed response rate (**Figure 5I-K**). This observation is expected since many
310 studies have reported that the majority of predicted epitopes fail to elicit T cell responses[38] and
311 that high HLA-peptide binding affinity does not equate to immunogenicity. Altogether, our data
312 show that MSI high patients have an increased frequency of high-quality T cell epitopes derived
313 from shared fs-peptides, binding to a broad spectrum of HLA alleles, capable of inducing
314 immunogenicity for CD8+ T cell in particular.
315

316 # Discussion

317 In this study we evaluated the MSI-H patients from the TCGA database for the presence of
318 shared, immunogenic tumor-associated neoantigens, providing a basis for the design of a
319 common "off-the-shelf" cancer vaccine. We proposed fs-events to have a high probability of
320 being shared across multiple MSI-H patients. To this end, we characterized the MSI-H
321 population and the tumor genomic frequencies of corresponding indel mutations, as well as their
322 tumor expression profiles. Our approach to detect immunogenic, shared neoantigens relies on

Manuscript

323 two assumptions: (i) indel mutations occurring in frequently mutated MS regions will lead to
324 identical fs-peptide extensions, (ii) these frequent, identical fs-peptide extensions encode poly-
325 epitopes with broad HLA allele specificity. We confirmed the validity of our neoantigen
326 selection approach by testing the immunogenicity of the selected fs-peptides. We found that the
327 selected peptides were highly immunogenic and generated strong CD3+CD8+ T cell responses
328 in a broad range of subjects.
329
330 One important finding is the existence of shared mutations that generate immunogenic
331 neoantigens. Indeed, the majority of tumors develop a restricted profile of mutations, either
332 disrupting activity of oncosuppressors, like p53 and KREB, or promoting gain-of-function
333 activity of oncogenes, like BRAF V600E. However, these shared mutations often escape
334 immune recognition and are less likely to encode tumor-associated neoantigens[17]. By contrast, in
335 MSI-H tumors indel mutagenesis is generally restricted to MS regions, thus increasing the
336 probability of being shared among patients (**Supplement Figure 1A, B**). Another key
337 observation is that indel mutations in MS within protein coding regions are more likely to
338 produce common fs-peptide, which may encode immunogenic neoantigens. With an average
339 length of 15 – 50 aa residues, frameshift extensions can easily accommodate multiple 9-mer T
340 cell epitopes. By running conventional antigen-prediction pipelines, we annotated several shared
341 fs-peptides as immunogenic.
342
343 From the tumor's evolutionary perspective, there is no need to keep those immunogenic
344 mutations, so multiple tumor-intrinsic escape mechanisms might exist as a consequence. As
345 described elsewhere, MSI-H tumors upregulate checkpoint molecules to evade the development

Manuscript

346 of antitumor T cell responses[12,26,37,39]. Blockade of this mechanism has been proven to be
347 effective in clearing a range of MSI-H tumors in multiple clinical trials[37,40]. However, several
348 other immune resistance mechanisms might exist: downregulation of HLA-allele and/or β-
349 microglobulin expression; inactivation/loss of antigen processing genes; inactivation/loss of
350 interferon-γ-response pathway genes; disruption of immunogenic neoantigens by additional,
351 acquired mutations[41–43]. The latter have been examined by us with respect to mutational escape
352 of shared poly-epitope fs-peptides. Indeed, in certain cases we observed mutations that were
353 potentially disruptive to the predicted neoantigens (**Supplementary Figure 2**). Another
354 important finding is the high-occurrence of predicted shared fs-peptides in commonly utilized
355 cancer cell lines (CCLE, **Figure 4**). We therefore suggest the potential utility of these shared fs-
356 positive cell lines in immunological studies of predicted fs-epitopes, such as understanding the
357 efficacy of antigen-dependent tumor cell killing as well as studying prospective tumor-escape
358 mechanisms.
359
360 Finally, we investigated the sequence space produced by fs-mutations. Previous reports provided
361 evidence that high quality immunogenic neoantigens are related to viral epitope sequences and
362 are predictive for outcome of checkpoint blockade immunotherapy[1,2]. To determine whether the
363 same rules applied to fs-derived neoantigens, we investigated similarities between viral epitopes
364 and immunogenic neoantigens. We found missense-derived neoantigens to be 3 times more
365 similar to viral epitopes than fs-neoantigens. We attributed this to viruses evolving within the
366 cellular environment of their host, trying to mimic the host's functionalities and potentially avoid
367 immune responses through extensive similarity to the host proteome. The fs-mutations are
368 therefore even "further from self" than viral antigens (**Supplemental Figure 3**). Taken together,

Manuscript

369 we conclude that frameshifts represent unique, intrinsically different epitope space, with a great
370 potential for discovering immunogenic epitopes which can be targeted by immune therapies. In a
371 recently published study authors investigated the presence of fs-derived neoepitopes in TCGA
372 and arrived at similar conclusions[44]. Also, a few previous reports investigated the
373 immunogenicity of unique fs-mutations on a small scale[45–49]. In addition to this shared research
374 goal, our group performed in depth computational characterization and experimental validation
375 of shared fs-epitopes in MSI-H tumors specifically.
376
377 Preselected fs-mutations might also be used for developing targeted sequencing panels for
378 diagnostic purposes. Precise detection and mapping of predicted mutations in each tumor patient
379 may enhance the chances of achieving a positive response with an applied vaccine. The usage of
380 targeted sequencing panels for diagnostics have already proven essential for developing
381 actionable treatments, particularly in the selection of targeted regimens. We believe the same
382 paradigm will become useful for precise immunotherapies, with physicians being able to select
383 the ideal cancer vaccine formulation based on the results of targeted sequencing panels.
384
385 Our work revealed the possibility of designing common cancer vaccines in specific tumor
386 subtypes with broad HLA-allele specificity. By applying tailored vaccines for MSI-H
387 endometrial, colorectal and stomach carcinomas, one can potentially achieve immunological
388 responses against existing neoplasms or develop preventive memory T cell responses in high-risk
389 patient populations, like those with Lynch syndrome.
390

391 # Online Methods

Manuscript

392 **Computational analysis of mutational data from TCGA**

393 Tumor-associated antigens were predicted using somatic mutation datasets, called by internal
394 mutation pipelines of The Cancer Genome Atlas (TCGA, or Genomics Data Commons,
395 https://gdc.cancer.gov). Briefly, annotated somatic missense and frameshift mutations by Mutect,
396 Somatic Sniper, Varscan and Muse were combined together per each patient. In case of somatic
397 missense mutations, corresponding 17-aminoacid residue-length normal peptides, surrounding
398 mutation site, were converted to tumor-specific peptides and used for MHC-I epitope prediction.
399 In case of frameshift mutations, the tumor specific peptide was called as following: major
400 mRNA isoform was mutated according to the frameshift mutation, translated starting from "-8"
401 aminoacid residue position from the mutation site till the stop codon within the new open reading
402 frame, defined by the frameshift. Resulting frameshift peptides were used for MHC-I epitope
403 prediction. NetMHC v4.0 and NetMHCpan v3.0[50,51] were used to predict missense and
404 frameshift epitopes. HLA allele types for >5000 patients from TCGA were taken from
405 previously published paper[3]. Collected epitope data was analyzed using statistical packages,
406 available in Prism and R, using custom written scripts.

407

408 **Expression analysis of TCGA**

409 Hg19-aligned RNAseq bam files were downloaded from GDC (https://gdc.cancer.gov). Obtained
410 .bam files were processed with samtools to extract RNAseq reads, covering 250 nt genomic loci
411 around shared fs-mutation (samtools view -b -L chr19.region.bed $i/*.bam >
412 $filename.chr19.bam). To count indel events in extracted RNAseq bam fiels, we applied
413 samtools mpileup tool: samtools mpileup -uf /work/scratch/index/TCGA/GRCh38.d1.vd1.fa $i |

Manuscript

414 bcftools view -l region.bed - | grep "INDEL" >> $filename.vcf. Finally, obtained data was
415 processed with custom scripts and analysed in PRISM 8.
416
417 **Peptide comparison with virus epitope databases**
418 The collection of viral MHC-I epitopes was downloaded from IEDB database and preformatted
419 for BLAST usage (makeblastdb -in iedb.fasta -parse_seqids -dbtype prot). Predicted frameshift
420 and missense originated T-cell epitopes form MSI-H patients were BLASTed against IEDB. For
421 comparison with viral we used following command: blastp -db iedb.fasta -query
422 frameshift.neoantigen.netMHC.score.fasta -outfmt "6 qseqid sseqid pident ppos positive
423 mismatch gapopen length qlen slen qstart qend sstart send qseq sseq evalue bitscore" -word_size
424 3 -gapopen 32767 -gapextend 32767 -evalue 1 -max_hsps_per_subject 1 -matrix BLOSUM62 -
425 max_target_seqs 10000000 -out frameshift.neoantigen.iedb.blast.out. To compare predicted
426 epitopes with human proteome, we used last command. First, we preformatted human proteome
427 (ensemble archive from December 2016): lastdb -p human.proteome human.proteome.fasta.
428 Then we used following command to compare epitopes to this database: lastal -f MAF -r 2 -q 1 -
429 m 100000000 -a 100000 -d 15 -l 4 -k 1 -j1 -P 10 human.proteome
430 frameshift.neoantigen.netMHC.score.fasta > frameshift.neoantigen.human.last.out. Finally,
431 obtained results were processed with custom scripts (bash, python) and analyzed in PRISM 8.
432
433 **Computational analysis of CCLE database**
434 Somatic mutation data and normalized RNAseq expression values for genes with shared fs-
435 mutations were obtained from https://portals.broadinstitute.org/ccle. Total mutation counts per

Manuscript

436 all cell lines were taken from CCLE_DepMap_18q3_maf_20180718.txt. Obtained data was
437 processed with custom scripts and analyzed in PRISM 8.
438
439 **Code availability**
440 The scripts and code are freely available on git account: https://github.com/VladimirRoudko
441
442 **Peptide synthesis**
443 Custom peptide libraries for WT and mutated peptides were chemically synthesized by
444 GenScript (USA/China). Each peptide had >80% purity as determined by high performance
445 liquid chromatography. MOG and CEFT peptide pools were commercially available at JPT
446 Peptide Technologies (Germany). Each peptide was resuspended in DMSO and used at a final
447 concentration of 1 µg/mL. Sequences of mutated peptides are displayed in supplemental figure 7
448 and the WT sequences are as follows: for SLC35F5 GKLTATQVAKISFFF, for SEC31A
449 QAVQSQGFINYCQKK, for SLC22A9 LEILKSTMKKELEAA, for TTK
450 ESHNSSSSKTFEKKR and YSGGESHNSSSSKTF, for SETD1B MENSHPPHHHHQQPP, for
451 OR7E24 MSYFPILFFFFLKRC, for RNF43 KSSLSARHPQRKRRG and for ASTE1
452 AEIFLPKGRSNSKKK.
453
454 **Rapid T-cell activation protocol**
455 Healthy donor PBMCs were cultured in X-VIVO15 media (LONZA) with cytokines promoting
456 dendritic cell (DC) differentiation overnight and then stimulated with peptide pool as displayed
457 in supplemental figure 7 (each peptide at 1 µg/mL) in the presence of adjuvant promoting DC
458 maturation in X-VIVO15. Stimulation with DMSO (vehicle) and MOG pool (JPT, 1 µg/mL)

Manuscript

459 were used as negative controls and CEFT pool (JPT, 1 μg/mL) were used as positive controls.
460 Next day, cells were fed with IL-2 (R&D Systems, 10 IU/mL) and IL-7 (R&D Systems, 10
461 ng/mL) in RPMI media (Gibco) containing 10% human serum. Cells were fed every 2-3 days.
462 IL-2 and IL-7 were not added at the last feeding. After 10 days of culture, cells were harvested
463 and re-stimulated with peptides (1 μg/mL) in the presence of anti-CD28 (0.5 mg/mL) and anti-
464 CD49d (0.5 mg/mL) antibodies. Where indicated, cells were stimulated with PMA (Sigma-
465 Aldrich, 50 ng/mL) and ionomycin (Sigma-Aldrich, 1 μg/mL), as positive control. IFN-γ
466 formation was measured by flow cytometry or ELISPOT. For flow cytometry, 1 hour after re-
467 stimulation with peptides, cells were added BD GolgiStop™, containing monensin and
468 BD GolgiPlug™, containing brefeldin A according to manufacturer's suggestion. IFN-
469 γ production was measured 12-hours after the addition of protein transport inhibitors by
470 intracellular staining usingBD Cytofix/Cytoperm™ reagents according to manufacturer's
471 protocol. For ELISPOT analysis, cells were re-stimulated in plates with mixed cellular ester
472 membrane that were coated with anti-IFN-γ antibody (Mabtech, 4 μg/mL). Plates were processed
473 for IFN-γ detection after 48-hours of culture.
474

475 **Author Contributions**

476 N.B. and S.B. initiated the project; N.B, V.R., C.C.B and T.O. designed the study; C.C.B.
477 collected the samples; V.R and C.C.B. acquired and analyzed data. N.B, B.G., V.R., C.C.B and
478 T.O interpreted the data; V.R. and C.C.B. wrote the manuscript; V.R., C.C.B., S.B., B.G. and
479 N.B. revised the manuscript.
480

481 **Competing Interests**

Manuscript

482 N.B. receives research funds from Novocure, Celldex, Ludwig institute, Genentech, Oncovir,
483 Melanoma Research Alliance, Cancer Research Institute, Leukemia & Lymphoma Society,
484 NYSTEM, Regeneron, and is on the advisory boards of Neon, Tempest, Checkpoint Sciences,
485 Curevac, Primevax, Novartis, Array BioPharma, Roche, Avidea; N.B. receives grant support and
486 serves on advisory board at Parker Institute for Cancer Immunotherapy; N.B. has National
487 Institutes of Health grants R01CA201189, R01CA180913 and R01AI081848
488 B.G. has National Institutes of Health grants 7R01AI081848-04 and 1P30CA196521-01; B.G.
489 has Stand Up To Cancer–National Science Foundation–Lustgarten Foundation Convergence
490 Dream Team Grant sponsored by Stand Up to Cancer, the Lustgarten Foundation, the V
491 Foundation and the National Science Foundation grant NSF 1545935; B.G. is The Pershing
492 Square Sohn Prize—Mark Foundation Fellow supported by funding from The Mark Foundation
493 for Cancer Research.

494

495 # References

496 1. Luksza, M. *et al.* A neoantigen fitness model predicts tumour response to checkpoint
497 blockade immunotherapy. *Nature* **551**, 517–520 (2017).
498 2. Balachandran, V. P. *et al.* Identification of unique neoantigen qualities in long-term
499 survivors of pancreatic cancer. *Nature* **551**, S12–S16 (2017).
500 3. Charoentong, P. *et al.* Pan-cancer Immunogenomic Analyses Reveal Genotype-
501 Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade:
502 Cell Reports. *Cell Rep.* **18**, 248–262 (2017).
503 4. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy
504 in Metastatic Melanoma. *Cell* **165**, 35–44 (2016).

Manuscript

505 5. Kim, T. M., Laird, P. W. & Park, P. J. The landscape of microsatellite instability in
506 colorectal and endometrial cancer genomes. *Cell* **155**, 858–868 (2013).
507 6. Walther, A. *et al.* Genetic prognostic and predictive markers in colorectal cancer. *Nat.*
508 *Rev. Cancer* **9**, 489–499 (2009).
509 7. Ahmed, D. *et al.* Epigenetic and genetic features of 24 colon cancer cell lines.
510 *Oncogenesis* **2**, 1–9 (2013).
511 8. Diaz-Padilla, I. *et al.* Mismatch repair status and clinical outcome in endometrial cancer:
512 A systematic review and meta-analysis. *Crit. Rev. Oncol. Hematol.* **88**, 154–167 (2013).
513 9. Zighelboim, I. *et al.* Microsatellite instability and epigenetic inactivation of MLH1 and
514 outcome of patients with endometrial carcinomas of the endometrioid type. *J. Clin. Oncol.*
515 **25**, 2042–2048 (2007).
516 10. Veigl, M., Kastury, L., Olechnowicz, J., Ma, A. & Markowitz, S. Biallelic inactivation of
517 hMLH 1 by epigenetic gene silencing , a novel mechanism causing human MSI cancers.
518 *PNAS* **95**, 8698–8702 (1998).
519 11. Cunningham, J. M. *et al.* Hypermethylation of the hMLHl Promoter in Colon Cancer with
520 Microsatellite Instability. (1998).
521 12. Gatalica, Z., Vranic, S., Xiu, J. & Swensen, J. High microsatellite instability (MSI-H)
522 colorectal carcinoma : a brief review of predictive biomarkers in the era of personalized
523 medicine. *Fam. Cancer* **15**, 405–412 (2016).
524 13. Chung, D. C. & Rustgi, A. K. The Hereditary Nonpolyposis Colorectal Cancer Syndrome:
525 Genetics and Clinical Implications. *Ann. Intern. Med.* **138**, (2019).
526 14. Carethers, J. M. *et al.* Advances in Colorectal Cancer Lynch syndrome and Lynch
527 syndrome mimics : The growing complex landscape of hereditary colon cancer. *World J.*

Manuscript

528 *Gastroenterol.* **21**, 9253–9261 (2015).

529 15. Cohen, S. A., Pritchard, C. C. & Jarvik, G. P. Lynch Syndrome : From Screening to
530 Diagnosis to Treatment in the Era of Modern Molecular Oncology. *Annu. Revew*
531 *Genomics Hum. Genet.* **20**, 1–15 (2019).

532 16. Vasen, H. F. A. *et al.* Cancer Risk in Families With Hereditary Nonpolyposis Colorectal
533 Cancer Diagnosed by Mutation Analysis. *Gastroenterology* **110**, 1020–1027 (1996).

534 17. Marty, R. *et al.* MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell* **0**,
535 1–12 (2017).

536 18. Turajlic, S. *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the
537 immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.* **18**, 1009–1021 (2017).

538 19. Nangalia, J. *et al.* Somatic CALR mutations in myeloproliferative neoplasms with
539 nonmutated JAK2. *N. Engl. J. Med.* **369**, 2391–2405 (2013).

540 20. Iranzo, J., Martincorena, I. & Koonin, E. V. Cancer-mutation network and the number and
541 specificity of driver mutations. *PNAS* **115**, 6010–6019 (2018).

542 21. Buljan, M., Blattmann, P. & Aebersold, R. Systematic characterization of pan-cancer
543 mutation clusters. *Mol. Syst. Biol.* **14**, 1–19 (2018).

544 22. Gerstung, M. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*
545 **171**, 1029–1041 (2017).

546 23. Cortes-Ciriano, I., Lee, S., Park, W. Y., Kim, T. M. & Park, P. J. A molecular portrait of
547 microsatellite instability across multiple cancers. *Nat. Commun.* **8**, 1–12 (2017).

548 24. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science*
549 *(80-.).* **348**, 69–74 (2015).

550 25. Mcgranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to

Manuscript

551 immune checkpoint blockade. *Science (80-.).* **351**, 1463–1469 (2016).

552 26. Mlecnik, B. *et al.* Integrative Analyses of Colorectal Cancer Show Immunoscore Is a
553 Stronger Predictor of Patient Survival Than Microsatellite Instability. *Immunity* **44**, 698–
554 711 (2016).

555 27. Ghandi, M., Huang, F., Garraway, L. A. & Sellers, W. R. Next-generation characterization
556 of the Cancer Cell Line Encyclopedia. *Nature* (2019). doi:10.1038/s41586-019-1186-3

557 28. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of
558 anticancer drug sensitivity. *Nature* **483**, 4–11 (2012).

559 29. Frischmeyer, P. A. *et al.* An mRNA surveillance mechanism that eliminates transcripts
560 lacking termination codons. *Science* **295**, 2258–61 (2002).

561 30. Schweingruber, C., Rufener, S. C., Zünd, D., Yamashita, A. & Mühlemann, O. Nonsense-
562 mediated mRNA decay - mechanisms of substrate mRNA recognition and degradation in
563 mammalian cells. *Biochim. Biophys. Acta* **1829**, 612–23 (2013).

564 31. Isken, O. & Maquat, L. E. The multiple lives of NMD factors: balancing roles in gene and
565 genome regulation. *Nat. Rev. Genet.* **9**, 699–712 (2008).

566 32. Conti, E. & Izaurralde, E. Nonsense-mediated mRNA decay: Molecular insights and
567 mechanistic variations across species. *Curr. Opin. Cell Biol.* **17**, 316–325 (2005).

568 33. Kurosaki, T., Popp, M. W. & Maquat, L. E. Quality and quantity control of gene
569 expression by nonsense-mediated mRNA decay. *Nat. Rev. Mol. Cell Biol.* (2019).
570 doi:10.1038/s41580-019-0126-2

571 34. Apcher, S. *et al.* Major source of antigenic peptides for the MHC class I pathway is
572 produced during the pioneer round of mRNA translation. *Proc. Natl. Acad. Sci. U. S. A.*
573 **108**, 11572–7 (2011).

Manuscript

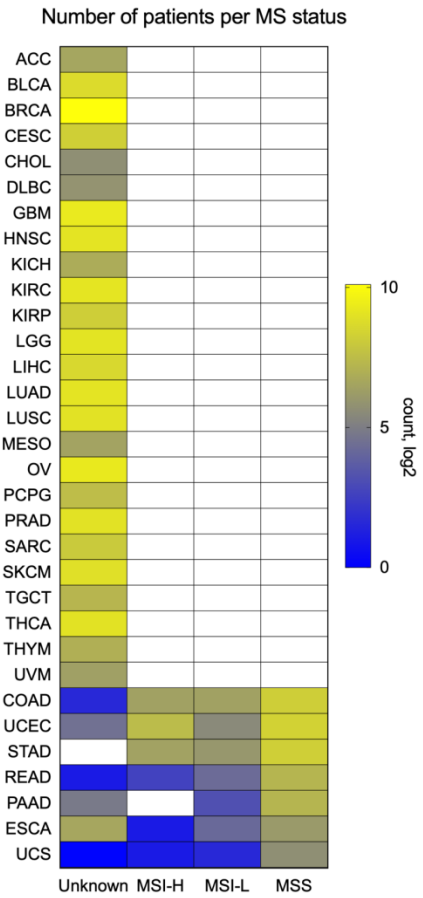
574 35. Buchwald, G. *et al.* Insights into the recruitment of the NMD machinery from the crystal
575 structure of a core EJC-UPF3b complex. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 10050–10055
576 (2010).
577 36. Dudley, J. C., Lin, M. T., Le, D. T. & Eshleman, J. R. Microsatellite instability as a
578 biomarker for PD-1 blockade. *Clin. Cancer Res.* **22**, 813–820 (2016).
579 37. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N. Engl. J.*
580 *Med.* **372**, 2509–2520 (2015).
581 38. Editorial. The problem with neoantigen prediction. *Nat. Biotechnol.* **35**, 97–97 (2017).
582 39. Mittica, G., Ghisoni, E., Giannone, G., Aglietta, M. & Valabrega, G. Checkpoint
583 inhibitors in endometrial cancer : preclinical rationale and clinical activity. *Oncotarget* **8**,
584 90532–90544 (2017).
585 40. Le, D. T. *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1
586 blockade. *Science (80-.).* **413**, 409–413 (2017).
587 41. Gao, J. *et al.* Loss of IFN-γ Pathway Genes in Tumor Cells as a Mechanism of Resistance
588 to Anti-CTLA-4 Therapy. *Cell* **167**, 397-404.e9 (2016).
589 42. Sharma, P., Hu-Lieskovan, S., Wargo, J. A. & Ribas, A. Primary, Adaptive, and Acquired
590 Resistance to Cancer Immunotherapy. *Cell* **168**, 707–723 (2017).
591 43. Roh, W. *et al.* Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and
592 PD-1 blockade reveals markers of response and resistance. *Sci. Transl. Med.* **9**, (2017).
593 44. Koster, J. & Plasterk, R. H. A. A library of Neo Open Reading Frame peptides (NOPs) as
594 a sustainable resource of common neoantigens in up to 50 % of cancer patients. *Sci. Rep.*
595 1–8 (2019). doi:10.1038/s41598-019-42729-2
596 45. Garbe, Y., Maletzki, C. & Linnebacher, M. An MSI Tumor Specific Frameshift Mutation

Manuscript

597 in a Coding Microsatellite of MSH3 Encodes for HLA-A0201-Restricted CD8+ Cytotoxic
598 T Cell Epitopes. *PLoS One* **6**, 2–9 (2011).
599 46. Maletzki, C., Schmidt, F., Dirks, W. G., Schmitt, M. & Linnebacher, M. Frameshift-
600 derived neoantigens constitute immunotherapeutic targets for patients with microsatellite-
601 instable haematological malignancies: Frameshift peptides for treating MSI+ blood
602 cancers. *Eur. J. Cancer* **49**, 2587–2595 (2013).
603 47. Wagner, S., Mullins S, C. & Linnebacher, M. Colorectal cancer vaccines: Tumor-
604 associated antigens vs neoantigens. *World J. Gastroenterol.* **24**, 5418–5432 (2018).
605 48. Schwitalle, Y. *et al.* Immune Response Against Frameshift-Induced Neopeptides in
606 HNPCC Patients and Healthy HNPCC Mutation Carriers. *Gastroenterology* **134**, 988–997
607 (2008).
608 49. Woerner, S. M. *et al.* Pathogenesis of DNA repair-deficient cancers: A statistical meta-
609 analysis of putative Real Common Target genes. *Oncogene* **22**, 2226–2235 (2003).
610 50. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks:
611 application to the MHC class I system. *Bioinformatics* **32**, 511–517 (2016).
612 51. Nielsen, M. & Andreatta, M. NetMHCpan-3.0; improved prediction of binding to MHC
613 class I molecules integrating information from multiple receptor and peptide length
614 datasets. *Genome Med.* **8**, 1–9 (2016).
615
616
617 # Figures and Tables

Manuscript

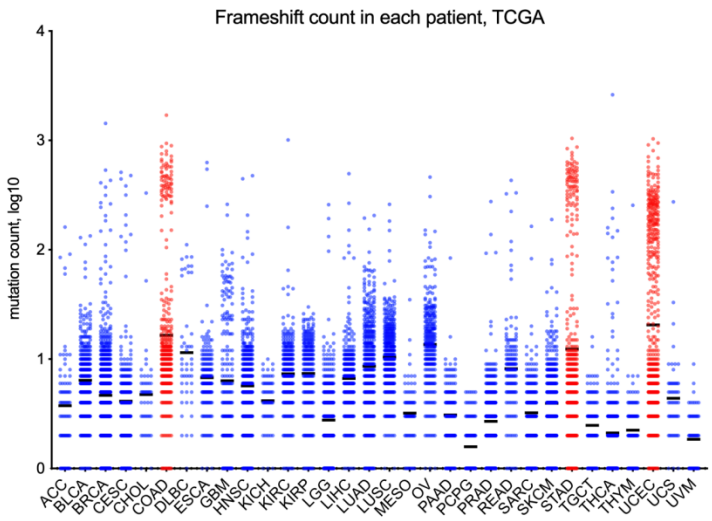
A.



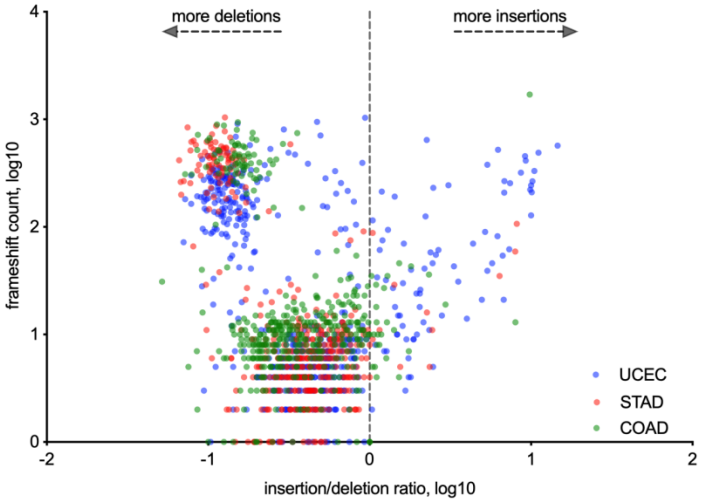
B.

	Unknown	MSI-H	MSI-L	MSS
COAD	0.00 (3)	0.18 (83)	0.18 (83)	0.63 (292)
UCEC	0.02 (22)	0.30 (170)	0.08 (44)	0.58 (324)
STAD	0.00 (0)	0.19 (85)	0.14 (63)	0.67 (295)

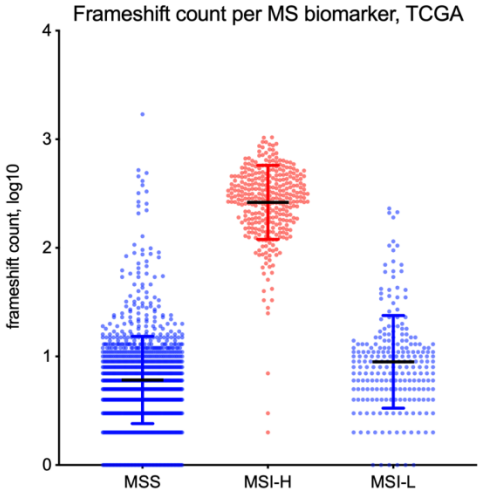
C.



D.



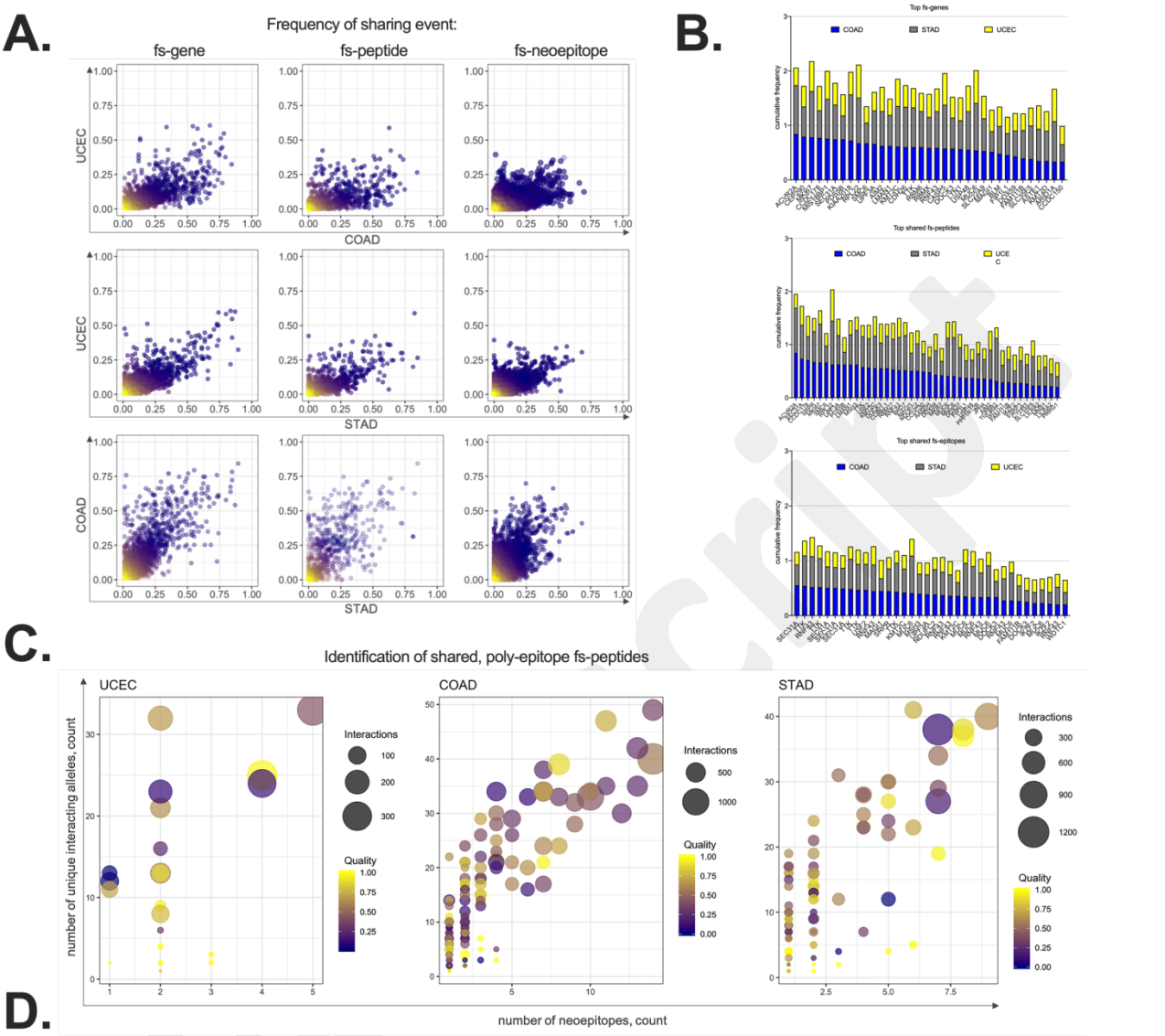
E.



618

619 **Figure 1**. Microsatellite instability is detected in COAD, STAD and UCEC tumors in TCGA.
620 Majority of MSI-H frameshifts are deletions. **A.** Quantification of patients with microsatellite
621 instable (MSI) tumors according to biomarker, applied by TCGA. MSI-H - MSI-high, MSI-L - MSI-
622 low, MSS - MS-stable, and Unknown – undetermined MS status. **B.** Table, showing the fraction
623 (absolute number) of patients with UCEC, COAD and STAD tumors identified as MSI-H, MSI-L,
624 MSS or Unknown. **C**. Frameshift (fs-) load (Y-axis, log10) in different tumor types across TCGA.
625 **D.** Comparison of fs-load (Y-axis, log10) with insertion-deletion ratio (X-axis, log10) in COAD,
626 STAD and UCEC tumors. **E**. Segregation of fs-load by MS biomarker.
627

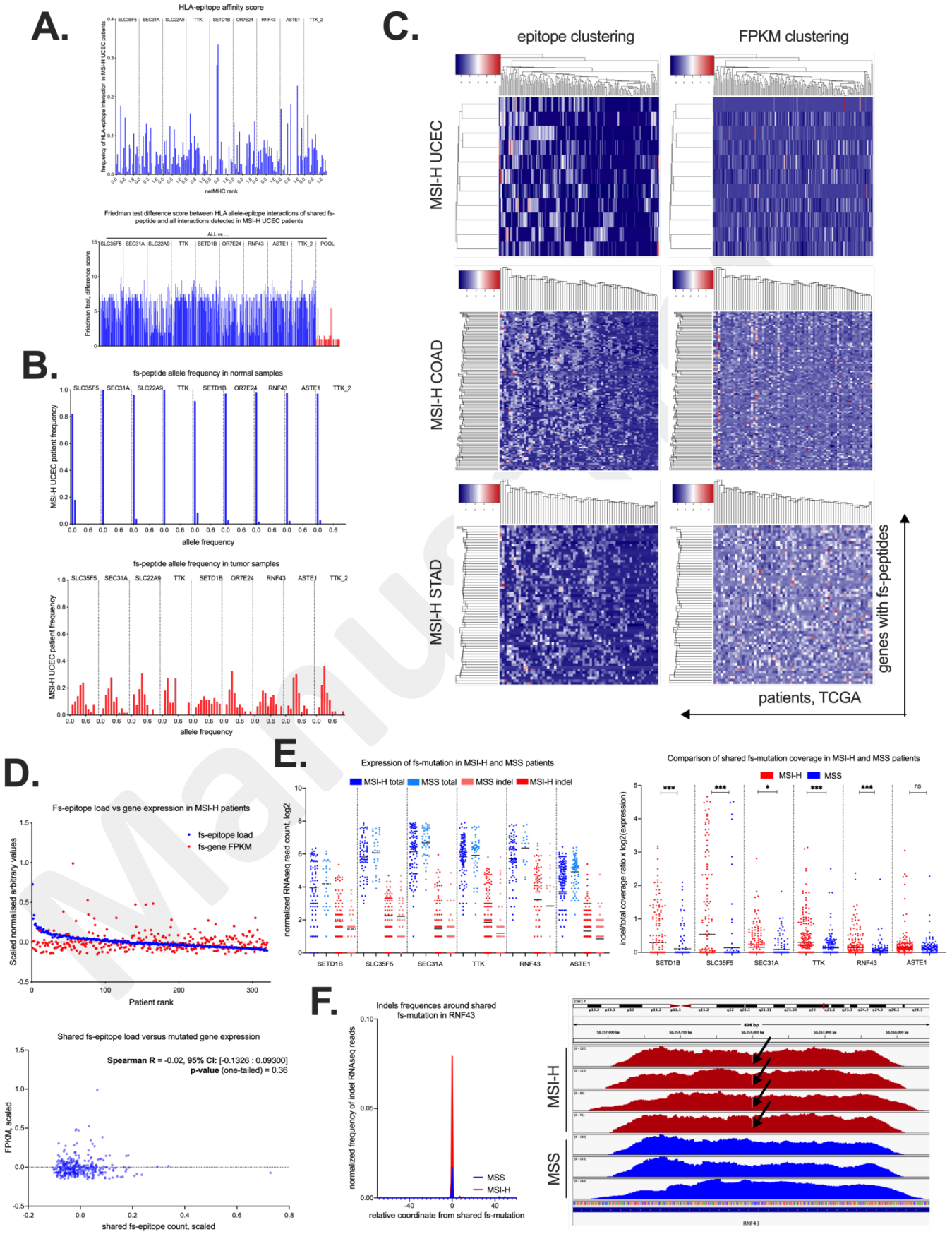
Manuscript


A.
Frequency of sharing event:
fs-gene
fs-peptide
fs-neoepitope
UCEC
COAD
STAD
B.
Top fs-genes
Top shared fs-peptides
Top shared fs-epitopes
COAD
STAD
UCEC
cumulative frequency
C.
Identification of shared, poly-epitope fs-peptides
UCEC
COAD
STAD
number of unique interacting alleles, count
number of neoepitopes, count
Interactions
Quality
D.
epitope count, log2
MSI-H UCEC patients
MSI-H UCEC HLA
MSI-H COAD patients
MSI-H COAD HLA
MSI-H STAD patients
MSI-H STAD HLA

628

629 **Figure 2**. Frequencies of shared, fs-events in STAD, COAD and UCEC MSI-H tumors. Greedy
630 pipeline for identification of shared, fs-peptides in MSI-H tumors. **A.** Scatterplots of patient
631 frequencies of frameshifted genes (LEFT), fs-peptides (CENTER) and fs-epitopes (RIGHT) in
632 UCEC, STAD and COAD MSI-H tumors. **B.** Cumulative frequency histograms, showing most
633 frequently mutated genes via indels (TOP); genes, mutated by shared fs-peptide (CENTER); genes,
634 encoding shared fs-neoantigens (BOTTOM). **C.** Three scatterplots showing selection procedure for
635 identification of shared fs-peptides in UCEC (LEFT), COAD (CENTER) and STAD (RIGHT)
636 patients. Each dot represents fs-peptide with the frequency of sharing above 20% in the patient
637 cohort. Amount of predicted 9-mer epitopes per peptide (X-axis) is plotted against a number of
638 predicted interacting HLA alleles (Y-axis). Size of the circle represents total number of predicted
639 interactions between HLA-alleles and encoded epitopes. Color of the dot reflects the ratio of
640 "PASS"ed fs-mutations to total number of called fs-mutations, according to somatic callers used by
641 TCGA consortium. **D.** Quantification of T cell epitopes (log2 scale) derived from shared fs-peptides
642 (columns) per each patient (rows, odd heatmaps) or each HLA allele (rows, even heatmaps) found in
643 UCEC, COAD and STAD MSI-H cohorts.
644

Manuscript



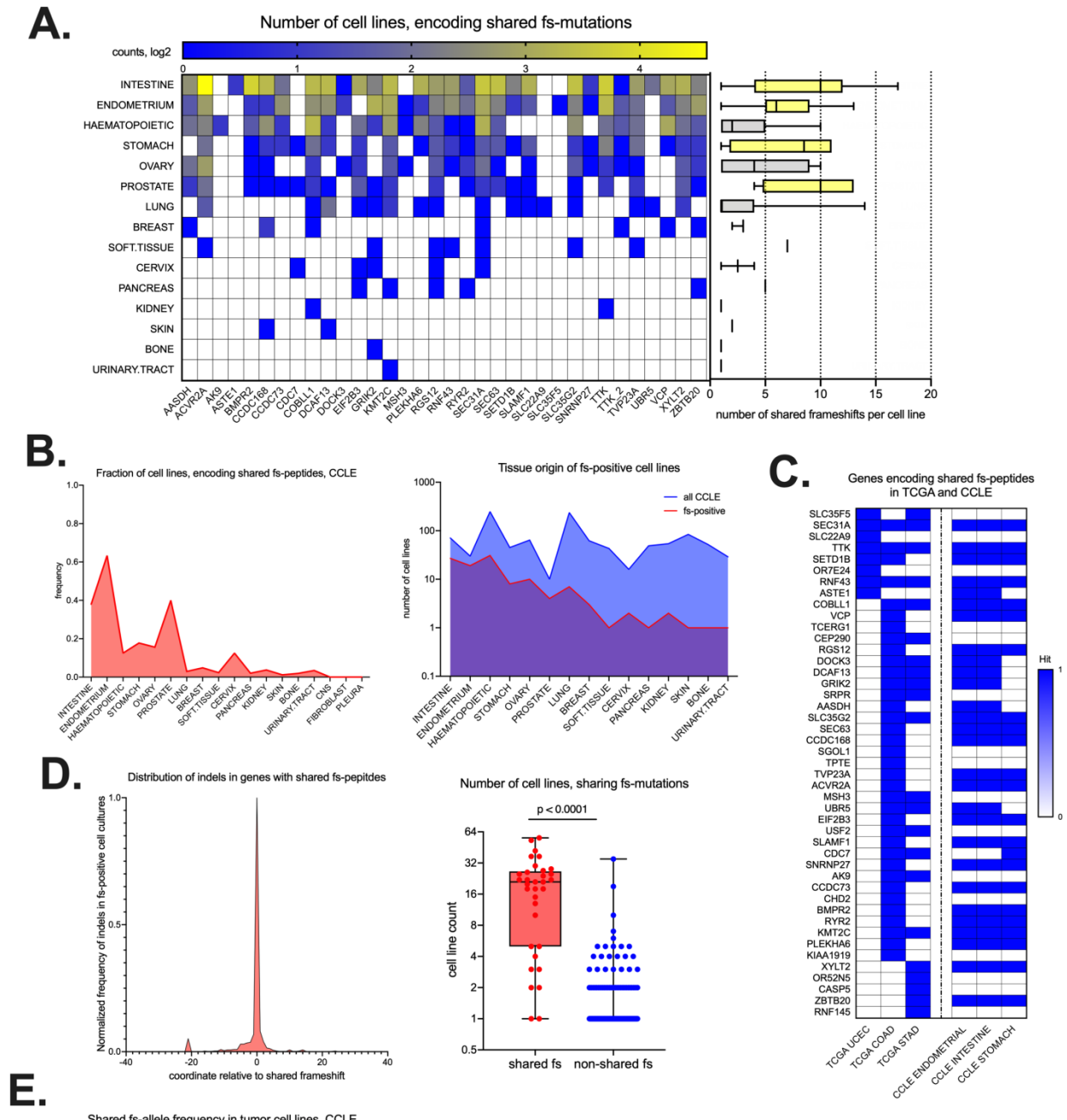
Manuscript

645

646 **Figure 3**. Genomic and expression properties of shared, immunogenic fs-mutations in MSI-H UCEC
647 tumor. Neoantigen quality of predicted epitopes. **A.** TOP – Distribution of predicted HLA allele-
648 epitope interaction ranking scores (NetMHC v3.4/v4.0), derived from nine shared fs-peptides in
649 MSI-H UCEC. BOTTOM – Friedman difference score, measuring the difference between total
650 amount of HLA-allele epitope interactions found in MSI-H UCEC patients (ALL) and either sum of
651 epitope-allele interactions per each fs-peptide separately, or pooled together (POOL). Each bar
652 represents the difference score between number of predicted epitopes per each HLA allele. **B.** tumor
653 allele frequency of nine shared frameshift mutations in normal (TOP) and tumor (BOTTOM) tissues
654 of MSI-H UCEC TCGA patients. **C.** Unsupervised hierarchical clustering of shared fs-mutations
655 (LEFT) and corresponding FPKM values of frameshifted genes (RIGHT) in MSI-H UCEC (TOP),
656 COAD (CENTER) and STAD (BOTTOM) tumors. Patients plotted in columns, genes plotted in
657 rows. **D.** TOP – scatterplot of shared fs-load and corresponding gene expression per each MSI-H
658 patient. BOTTOM – correlation plot between scaled fs-load and expression. **E**. Normalized
659 expression of nine shared fs-mutations in MSI-H UCEC patients. LEFT – normalized expression of
660 total and indel-containing reads spanning the genomic loci of six shared fs-mutations. RIGHT – ratio
661 of indel to total read count spanning the microsatellite region in MSI-H and MSS patient cohort of
662 UCEC, STAD and COAD tumors. Statistical significance is derived from non-parametric Mann-
663 Whitney two-tailed test. **F.** Representative example of shared frameshift detected in RNAseq
664 datasets. LEFT – normalized frequency of fs-mutation in RNF43 within 100 nucleotide genomic
665 loci: 50 nt upstream and 50 nt downstream the shared frameshift in MSI-H and MSS RNAseq
666 samples. RIGHT – RNAseq read histograms around the shared frameshift in RNF43 from MSI-H
667 (red) and MSS (blue) patients. The 1-nucleotide long drop in RNAseq read coverage at the site of
668 shared frameshift is indicated with an arrow.

Manuscript

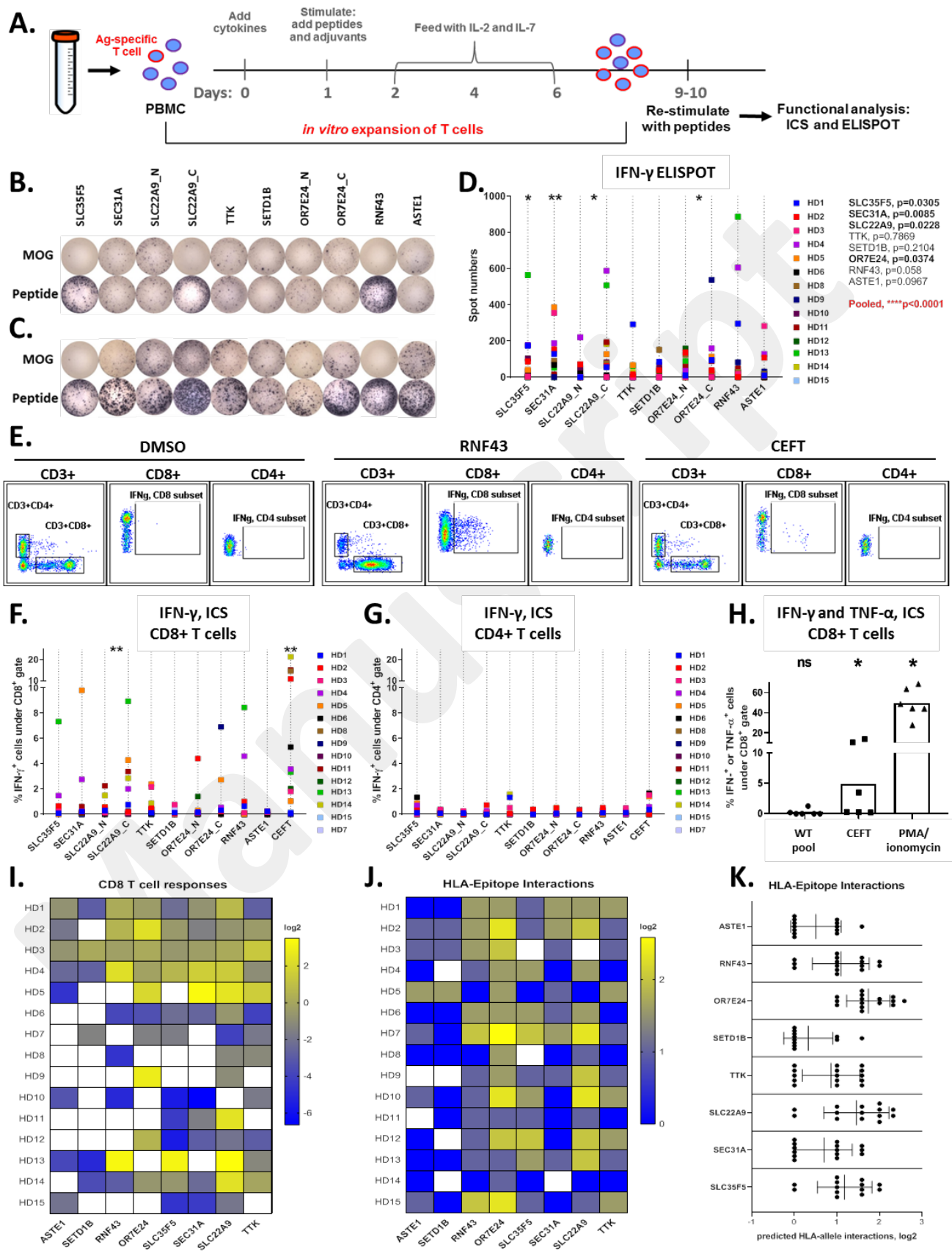
669


A.
Number of cell lines, encoding shared fs-mutations
counts, log2
INTESTINE
ENDOMETRIUM
HAEMATOPOIETIC
STOMACH
OVARY
PROSTATE
LUNG
BREAST
SOFT.TISSUE
CERVIX
PANCREAS
KIDNEY
SKIN
BONE
URINARY.TRACT
number of shared frameshifts per cell line
B.
Fraction of cell lines, encoding shared fs-peptides, CCLE
frequency
Tissue origin of fs-positive cell lines
all CCLE
fs-positive
number of cell lines
C.
Genes encoding shared fs-peptides in TCGA and CCLE
Hit
TCGA UCEC
TCGA COAD
TCGA STAD
CCLE ENDOMETRIAL
CCLE INTESTINE
CCLE STOMACH
D.
Distribution of indels in genes with shared fs-pepitdes
Normalized frequency of indels in fs-positive cell cultures
coordinate relative to shared frameshift
Number of cell lines, sharing fs-mutations
p < 0.0001
cell line count
shared fs
non-shared fs
E.
Shared fs-allele frequency in tumor cell lines, CCLE
cumulative frequency of fs-positive cell lines
shared fs-allele frequency
Allele coverage of frequent shared fs-mutations
frequency of shared frameshift
fs-allele frequency

670

671 **Figure 4.** Detection of shared fs-mutations in cancer cell line encyclopedia (CCLE). **A.**
672 Quantification of shared fs-mutations in cell lines per tissue of tumor origin and per each
673 frameshifted gene. Histogram plot on the left shows number of shared fs-mutations per cell line.
674 **B.** General statistics of cell lines, encoding shared fs-peptides. LEFT – fraction of cell lines,
675 positive for shared fs-mutation, per each tissue type. RIGHT – Absolute number of cell lines
676 with detected shared fs-mutation compared to total number of cell lines in CCLE. Cell lines are
677 sorted according to tissue origin. **C.** Gene and tumor specificity of shared fs-mutations in TCGA
678 and CCLE. **D.** Distribution of all detected indels in genes with shared fs-mutations (34 genes in
679 CCLE). LEFT – metagene, showing normalized frequency of all detected indels in 34 genes,
680 around shared fs-mutation. RIGHT – t-test of number of cell lines encoding shared fs-mutation
681 versus all other fs-mutations, detected in selected 34 genes. **E.** fs-allele frequency in WES
682 samples of shared fs-mutation positive cell lines. LEFT – cumulative histogram plot, showing fs-
683 allele frequency of each shared fs-mutation per gene of origin. RIGHT – box-plot of fs-allele
684 frequency plotted against frequency of fs-mutation in the pool of fs-positive cell lines.
685

Manuscript

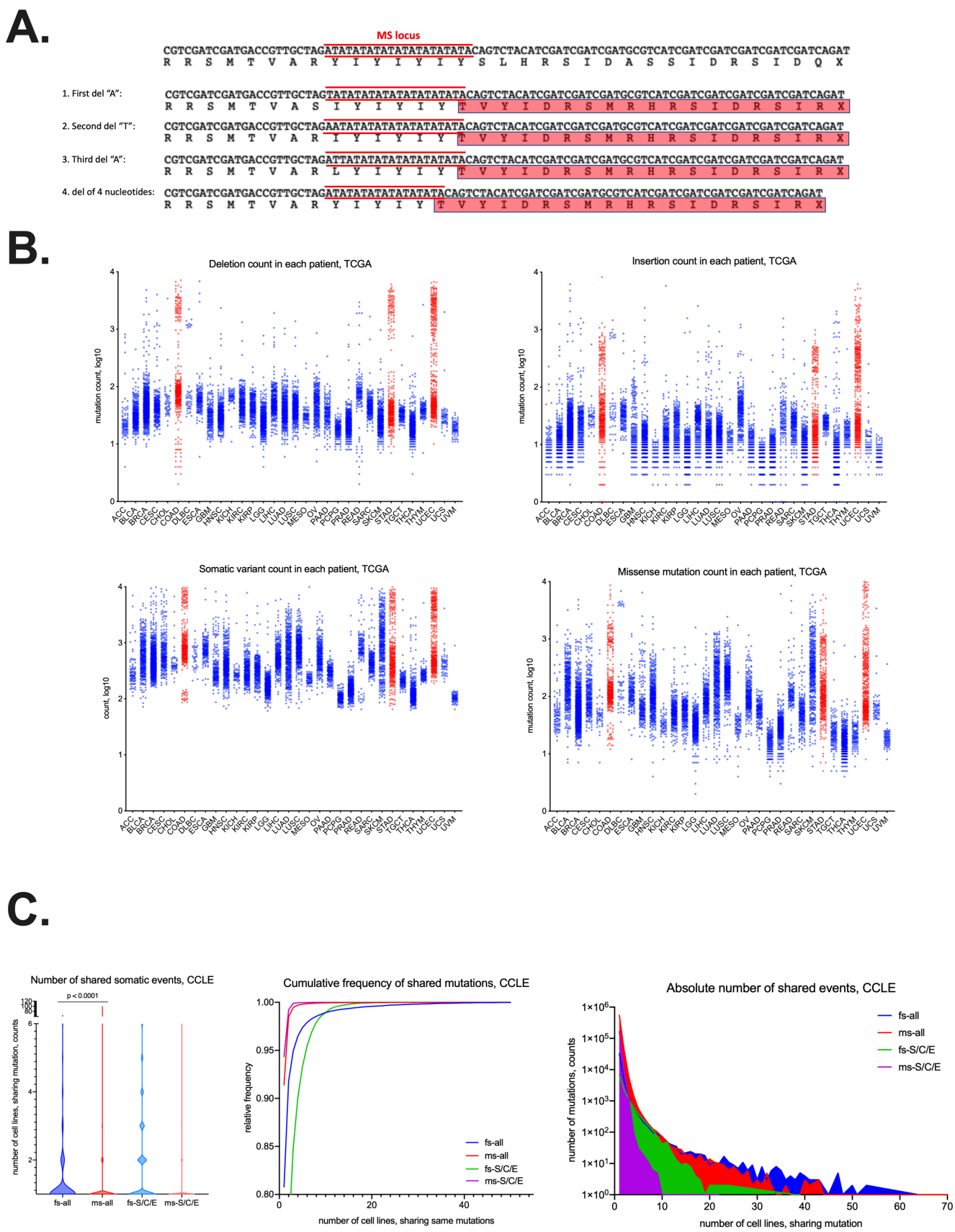

A.
Ag-specific T cell
PBMC
Add cytokines
Stimulate: add peptides and adjuvants
Feed with IL-2 and IL-7
Days: 0 1 2 4 6 9-10
Re-stimulate with peptides
Functional analysis: ICS and ELISPOT
in vitro expansion of T cells
B.
C.
MOG
Peptide
SLC35F5
SEC31A
SLC22A9_N
SLC22A9_C
TTK
SETD1B
OR7E24_N
OR7E24_C
RNF43
ASTE1
D.
IFN-γ ELISPOT
Spot numbers
SLC35F5, p=0.0305
SEC31A, p=0.0085
SLC22A9, p=0.0228
TTK, p=0.7869
SETD1B, p=0.2104
OR7E24, p=0.0374
RNF43, p=0.058
ASTE1, p=0.0967
Pooled, ****p<0.0001
E.
DMSO
RNF43
CEFT
CD3+
CD8+
CD4+
CD3+CD4+
CD3+CD8+
IFNg, CD8 subset
IFNg, CD4 subset
F.
IFN-γ, ICS CD8+ T cells
% IFN-γ+ cells under CD8+ gate
G.
IFN-γ, ICS CD4+ T cells
% IFN-γ+ cells under CD4+ gate
H.
IFN-γ and TNF-α, ICS CD8+ T cells
% IFN-+ or TNF-α+ cells under CD8+ gate
ns
WT pool
CEFT
PMA/ ionomycin
I.
CD8 T cell responses
log2
J.
HLA-Epitope Interactions
K.
HLA-Epitope Interactions
predicted HLA-allele interactions, log2

686

687 **Figure 5.** Shared fs-peptides predicted from UCEC MSI-H patients elicit T cell responses. **A.** An
688 overview of T cell immunogenicity assay used to evaluate antigen-specific T cell responses.
689 PBMCs from healthy donors (HD) were expanded *in vitro* following stimulation with fs-peptide
690 OLPs as shown in supplementary figure 7. Expanded T cells ($5x10^4$ cells/well) were re-
691 stimulated with either the peptide pool they were expanded with or the control peptide pool
692 MOG. Representative IFN-γ ELISPOT images for **B.** HD13 or **C.** for selected responsive HD. **D.**
693 Summary of ELISPOT data (n=14). Statistical significance for MOG vs OLPs was evaluated by
694 Wilcoxon signed-rank test. **E.** Representative flow cytometry plots and summary of data (n=15)
695 for IFN-γ in **F.** CD8 and **G.** CD4 T cell subsets. Stimulation with CEFT was used as a control.
696 Statistical significance for DMSO vs OLPs was evaluated by Wilcoxon signed-rank test.
697 **p=0.0032 for SLC22A9 and **0.0031 for CEFT. **H.** Frequency of IFN-γ or TNF-α producing
698 CD8+ T cells upon stimulation with WT OLP pool. CEFT and PMA/Ionomycin stimulation was
699 used as a control. The spot numbers and % IFN-γ values were calculated by subtracting the
700 values obtained after MOG or DMSO stimulation from the values after OLP pool stimulation
701 and negative values were set to zero. **I.** Summary of log2 transformed data for IFN-γ /TNF-α
702 response by CD8+ T cells against fs-peptides. **J.** and **K.** Quantification (log2 transformed) of HLA
703 allele-epitope interactions for each subject per fs-peptide. Interactions were counted when IC50<500
704 or percentile rank<2.
705

706 # Supplementary Information

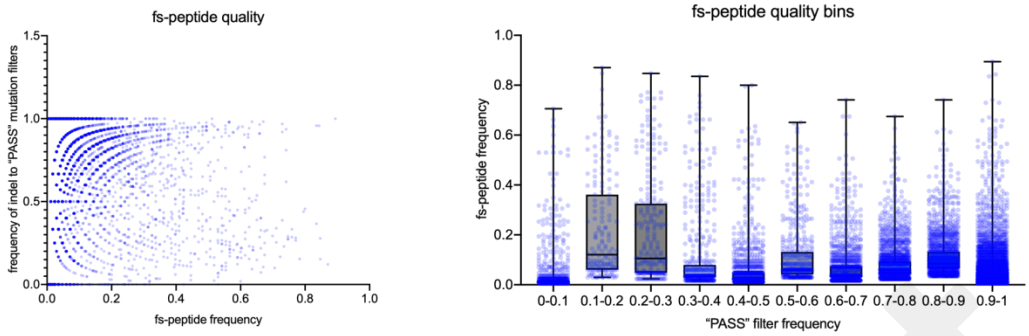
Manuscript


A.
MS locus
CGTCGATCGATGACCGTTGCTAGATATATATATATATATATATATACAGTCTACATCGATCGATCGATGCGTCATCGATCGATCGATCGATCGATCAGAT
R R S M T V A R Y I Y I Y I Y S L H R S I D A S S I D R S I D Q X
1. First del "A":
CGTCGATCGATGACCGTTGCTAGTATATATATATATATATATATACAGTCTACATCGATCGATCGATGCGTCATCGATCGATCGATCGATCGATCAGAT
R R S M T V A S I Y I Y I Y T V Y I D R S M R H R S I D R S I R X
2. Second del "T":
CGTCGATCGATGACCGTTGCTAGAATATATATATATATATATATACAGTCTACATCGATCGATCGATGCGTCATCGATCGATCGATCGATCGATCAGAT
R R S M T V A R I Y I Y I Y T V Y I D R S M R H R S I D R S I R X
3. Third del "A":
CGTCGATCGATGACCGTTGCTAGATTATATATATATATATATATACAGTCTACATCGATCGATCGATGCGTCATCGATCGATCGATCGATCGATCAGAT
R R S M T V A R L Y I Y I Y T V Y I D R S M R H R S I D R S I R X
4. del of 4 nucleotides:
CGTCGATCGATGACCGTTGCTAGATATATATATATATATATACAGTCTACATCGATCGATCGATGCGTCATCGATCGATCGATCGATCGATCAGAT
R R S M T V A R Y I Y I Y T V Y I D R S M R H R S I D R S I R X
B.
Deletion count in each patient, TCGA
mutation count, log10
Insertion count in each patient, TCGA
mutation count, log10
Somatic variant count in each patient, TCGA
count, log10
Missense mutation count in each patient, TCGA
mutation count, log10
ACC BLCA BRCA CESC CHOL COAD DLBC ESCA GBM HNSC KICH KIRC KIRP LGG LIHC LUAD LUSC MESO OV PAAD PCPG PRAD READ SARC SKCM STAD TGCT THCA THYM UCEC UCS UVM
C.
Number of shared somatic events, CCLE
p < 0.0001
number of cell lines, sharing mutation, counts
fs-all
ms-all
fs-S/C/E
ms-S/C/E
Cumulative frequency of shared mutations, CCLE
relative frequency
number of cell lines, sharing same mutations
Absolute number of shared events, CCLE
number of mutations, counts
number of cell lines, sharing mutation

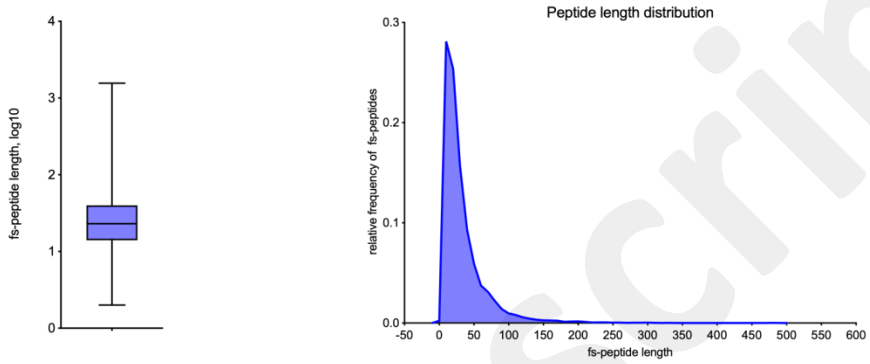
708 **Supplementary Figure 1**. Mutational analysis of TCGA. **A.** Hypothetical example of multiple
709 deletions happening in microsatellite region and leading to identical frameshift (fs-)peptide. **B.**
710 Comparison of frequencies of shared somatic indels and missense point mutations in human cancer
711 cell lines from CCLE. LEFT – number of cancer cell lines, sharing same fs-mutation or same
712 missense point mutation from all CCLE (fs-all and ms-all respectively) or from stomach, colon and
713 endometrium cell lines (fs-S/C/E and ms-S/C/E respectively). Statistical significance derived from
714 non-parametric Mann-Whitney, one-tailed test. CENTER – cumulative frequency plot of shared
715 indel and missense mutation from all CCLE (fs-all and ms-all respectively) or from stomach, colon
716 and endometrium cell lines (fs-S/C/E and ms-S/C/E respectively). RIGHT – absolute number of
717 shared mutations in cancer cell lines from all CCLE (fs-all and ms-all respectively) or from stomach,
718 colon and endometrium cell lines (fs-S/C/E and ms-S/C/E respectively). **C.** Mutational load across
719 different tumor types in TCGA. From left to right, top to bottom: deletion, insertion, somatic variant
720 and missense mutation load.
721

Manuscript

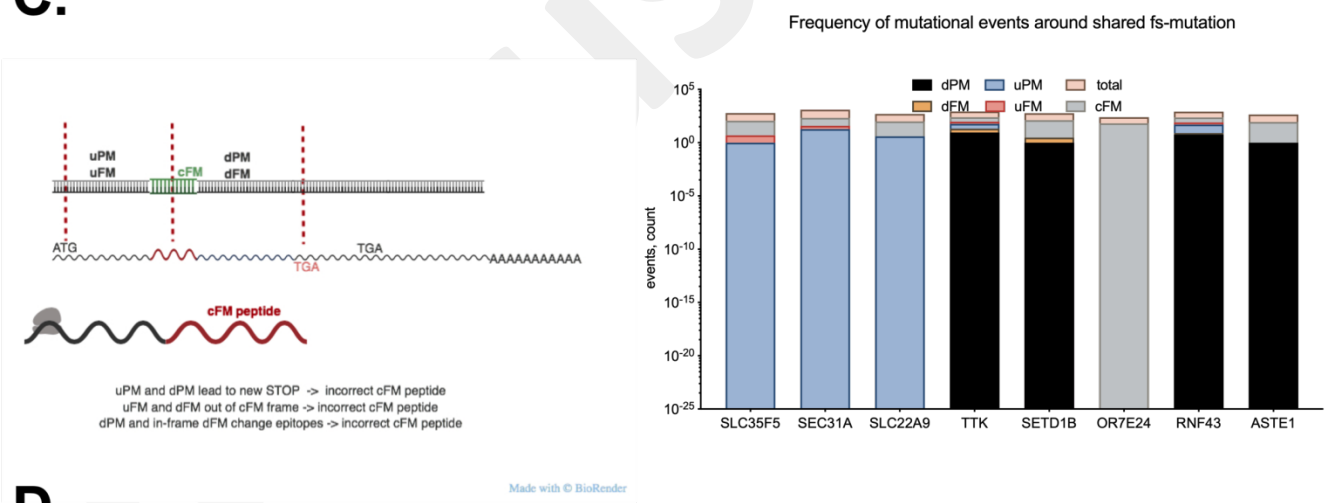
A.



B.



C.



D.

$$P_{ccFM\,|\,cFM} = 1 - [\,(P_{uPM\,|\,cFM} + P_{dPM\,|\,cFM}) \times P_{A,T,G} \times P_{STOP\,|\,A,T,G} + (P_{dPM\,|\,cFM} \times P_{missense} + P_{dFM\,|\,cFM} \times P_{inframe}) \times P_{epitope} + (P_{uFM\,|\,cFM} + P_{dFM\,|\,cFM}) \times P_{outframe}\,] \times P_{linkage}$$

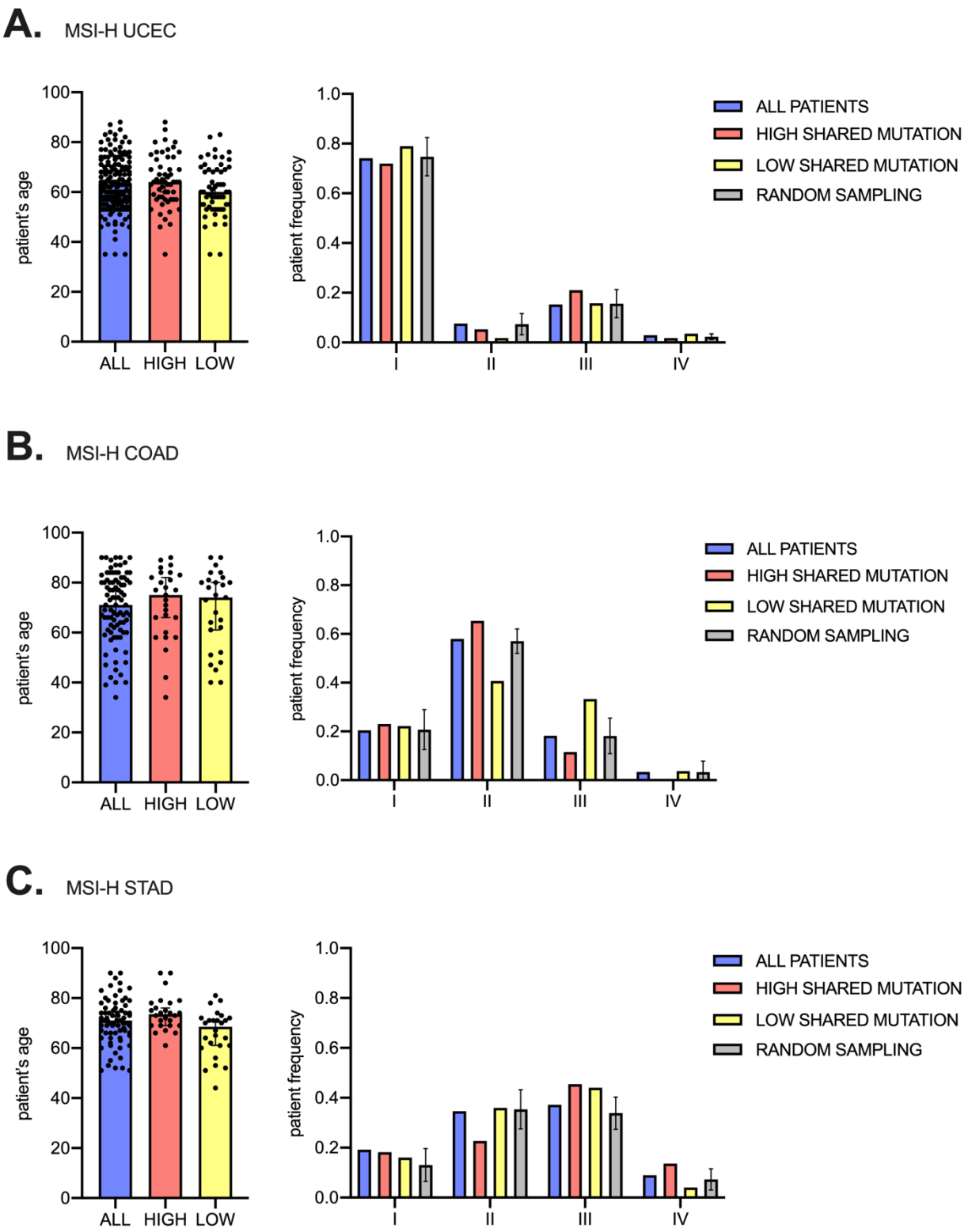
$$P_{ccFM\,|\,cFM} = 1 - [\,(P_{uPM\,|\,cFM} + P_{dPM\,|\,cFM}) \times 3/4 \times 3/27 + (P_{dPM\,|\,cFM} \times 2/3 + P_{dFM\,|\,cFM} \times 1/3) \times \Sigma_{k=1}^{n}\, C^{n}_{k} \times (7/len_{FM})^{k} + (P_{uFM\,|\,cFM} + P_{dFM\,|\,cFM}) \times 2/3\,] \times 1/2$$

GENE	PccFM | cFM
SLC35F5	0.98
SEC31A	0.92
SLC22A9	0.99
TTK	0.79
SETD1B	0.99
OR7E24	1.00
RNF43	0.82
ASTE1	1.00

722

723 **Supplementary Figure 2**. Properties of predicted frameshift peptides. **A.** Distribution of quality
724 metrics for each frameshift mutation detected in MSI-H patients. LEFT – ratio of "PASS" mutation
725 calls for each frameshift, plotted against fs-mutation frequency. RIGHT – box-plot of frameshift
726 frequency in MSI-H patient cohort per 10%-quality bin. The least shared fs-events are the most
727 confident, while majority of shared fs-mutations are within 10-50% "PASS" quality range. **B.** Box
728 plot (LEFT) and distribution (RIGHT) of fs-peptide frequency as function of its length. Most
729 frequent fs-peptides are in 15-50 aminoacid residue range. **C, D.** Estimation of correctness of nine
730 shared frameshift peptides in MSI-H UCEC population. In the reduced approximation, any
731 upstream mutation (upstream fs-mutation uFM or upstream point mutation uPM) may alter the
732 fs-peptide if uFM appears somewhere between start codon and predicted shared fs-mutation or if
733 uPM leads to an abortive stop codon; alternatively, any downstream mutation (downstream fs-
734 mutation dFM or downstream point mutation dPM) may alter shared fs-peptide if dFM happens
735 between shared mutation and novel stop codon, defined by the frame of fs-peptide or if dPM
736 happens within predicted T cell epitope or leads to abortive stop codon. **C.** LEFT: representation
737 of upstream and downstream mutagenesis, which can be detrimental to the predicted fs-peptide;
738 RIGHT – quantification of detrimental events around shared fs-peptides. **D.** LEFT: probabilistic
739 function to estimate conditional probability of fs-peptide being correct given that this fs-mutation is
740 happened; RIGHT: estimated posterior probabilities for each nine shared fs-peptides.
741

Manuscript

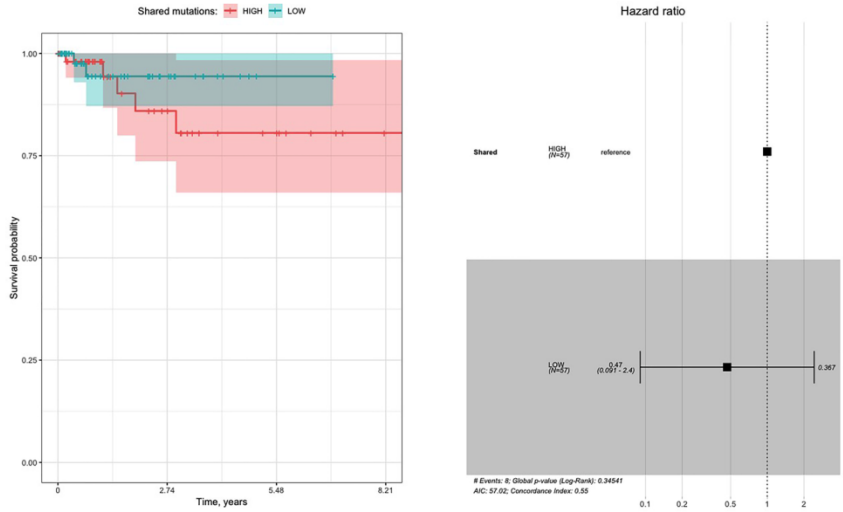

A.
MSI-H UCEC
patient's age
100
80
60
40
20
0
ALL
HIGH
LOW
patient frequency
1.0
0.8
0.6
0.4
0.2
0.0
I
II
III
IV
ALL PATIENTS
HIGH SHARED MUTATION
LOW SHARED MUTATION
RANDOM SAMPLING
B.
MSI-H COAD
C.
MSI-H STAD

742

743 **Supplementary Figure 3.** Distribution of patient's age and tumor stages based on shared fs-epitope
744 load in MSI-H tumors, TCGA. **A.** LEFT – distribution of patient's age in all, shared fs-epitope high
745 and shared fs-epitope low patients' cohort, MSI-H UCEC tumor type. RIGHT – patients' frequency
746 across tumor stages in the same cohorts, as above. **B.** LEFT – distribution of patient's age in all,
747 shared fs-epitope high and shared fs-epitope low patients' cohort, MSI-H COAD tumor type. RIGHT
748 – patients' frequency across tumor stages in the same cohorts, as above. **C.** LEFT – distribution of
749 patient's age in all, shared fs-epitope high and shared fs-epitope low patients' cohort, MSI-H STAD
750 tumor type. RIGHT – patients' frequency across tumor stages in the same cohorts, as above.
751

Manuscript

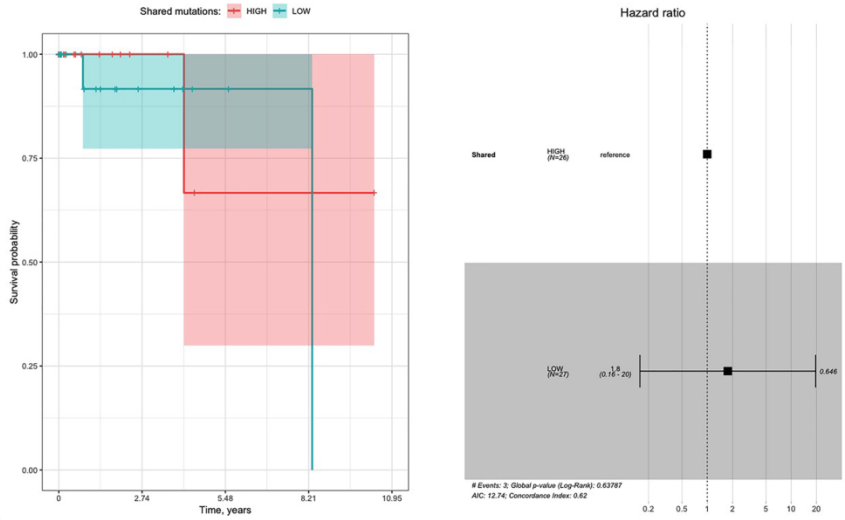
A. MSI-H UCEC



Log-Rank p-value: 0.34541

Hazard ratio: 0.47
(95% CI: 0.091 - 2.4)

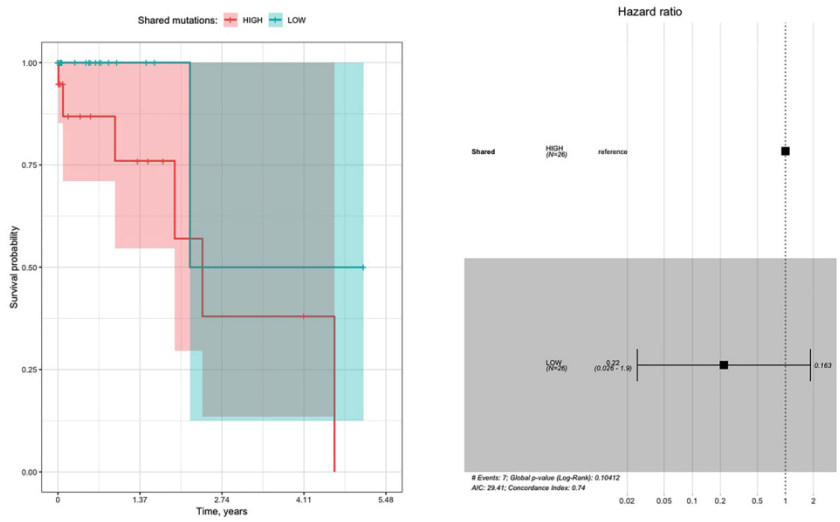
B. MSI-H COAD



Log-Rank p-value: 0.63787

Hazard ratio: 1.8
(95% CI: 0.16 - 20)

C. MSI-H STAD



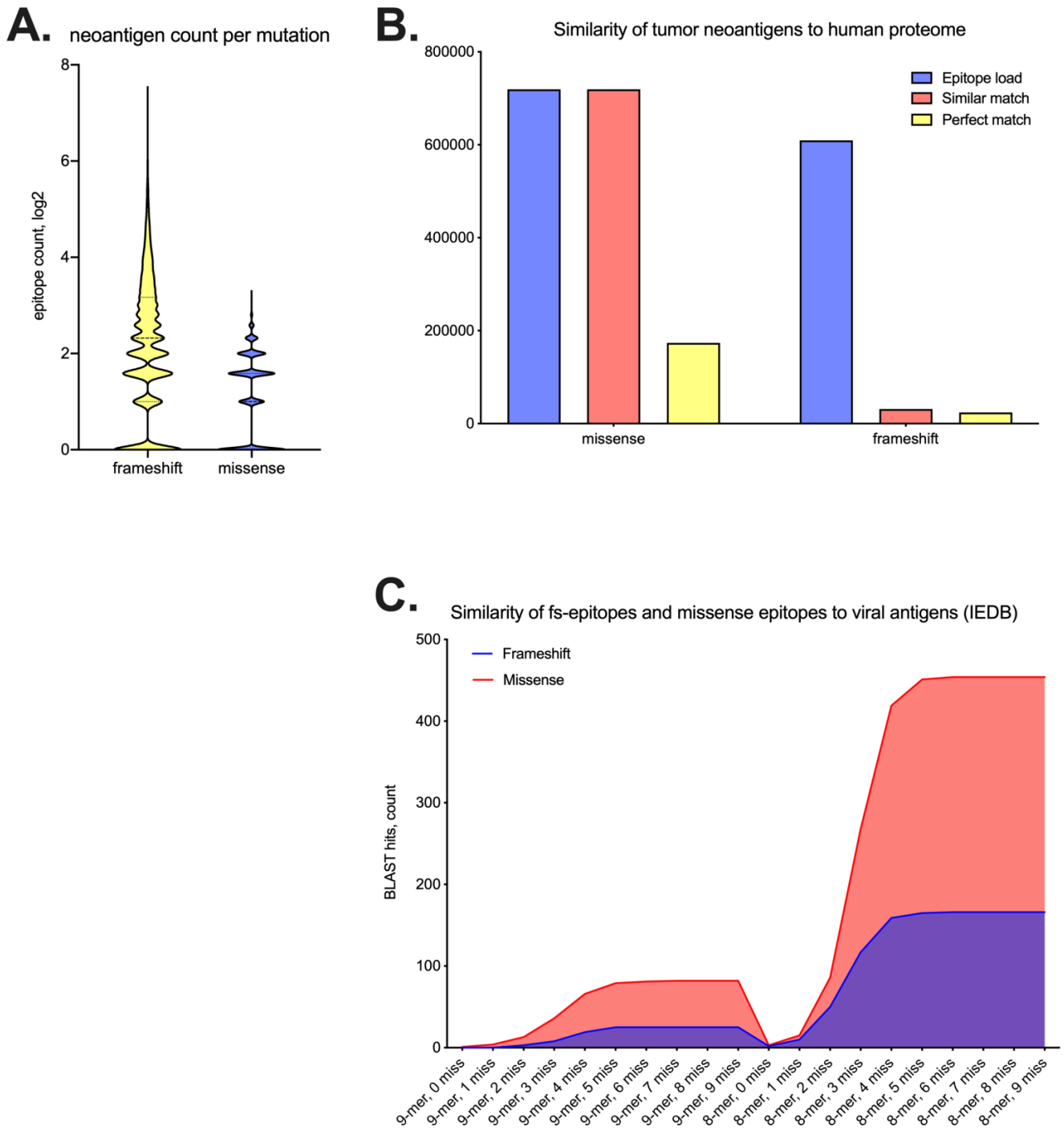
Log-Rank p-value: 0.10412

Hazard ratio: 0.22
(95% CI: 0.026 - 1.9)

752

753 **Supplementary Figure 4.** Kaplan-Meier plot of survival analysis (LEFT) and Cox's proportional
754 hazards model (RIGHT) of MSI-H UCEC (**A**), COAD (**B**) and STAD (**C**) patients, segregated by
755 shared fs-epitope load (HIGH vs LOW).
756

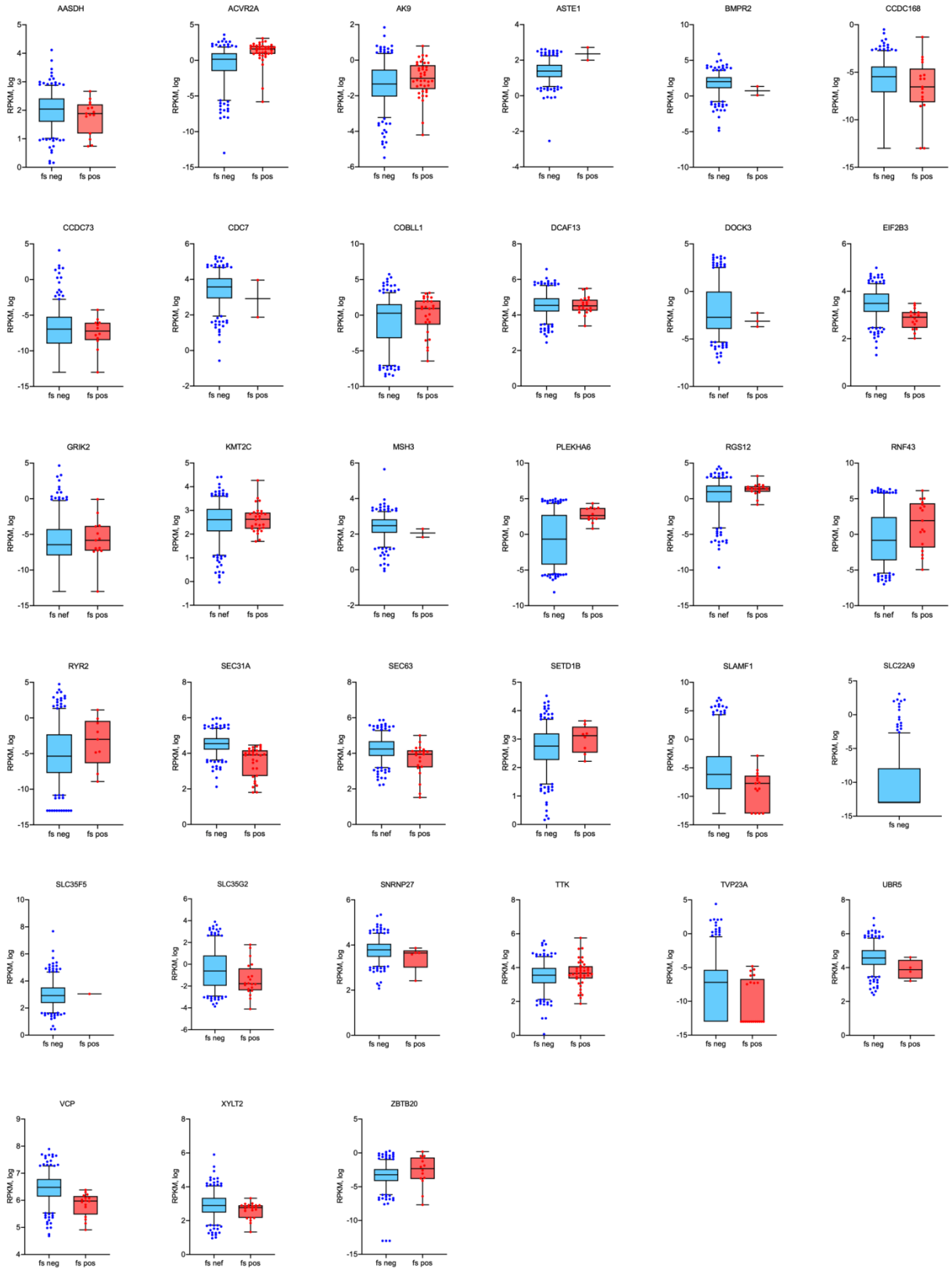
Manuscript



Supplementary Figure 5. Comparison of missense epitopes and fs-epitopes in MSI-H patients with human proteome and annotated viral epitopes (IEDB). **A.** Quantification of epitopes per missense or frameshift mutation. On average, frameshift mutation generates 4 epitopes, while missense - 2. **B.**

761 Mapping missense and fs-epitopes back to human proteome (ensemble, 2016). "Epitope load"
762 indicates total amount of predicted neoantigens from missense and frameshift mutations; "Similar
763 match" shows number of epitopes, successfully aligned to human proteome by sequence of 8 non-
764 gapped aminoacid residues, allowing 1 mismatch; "Perfect match" shows number of epitopes,
765 perfectly matched to human proteome by sequence of 9 non-gapped aa with no mismatches. **C.**
766 BLAST comparison of missense epitopes and fs-epitopes with viral epitopes from IEDB, allowing
767 different number of mismatches. Missense-derived epitopes are more similar to viral epitopes then
768 frameshift-derived on average.
769

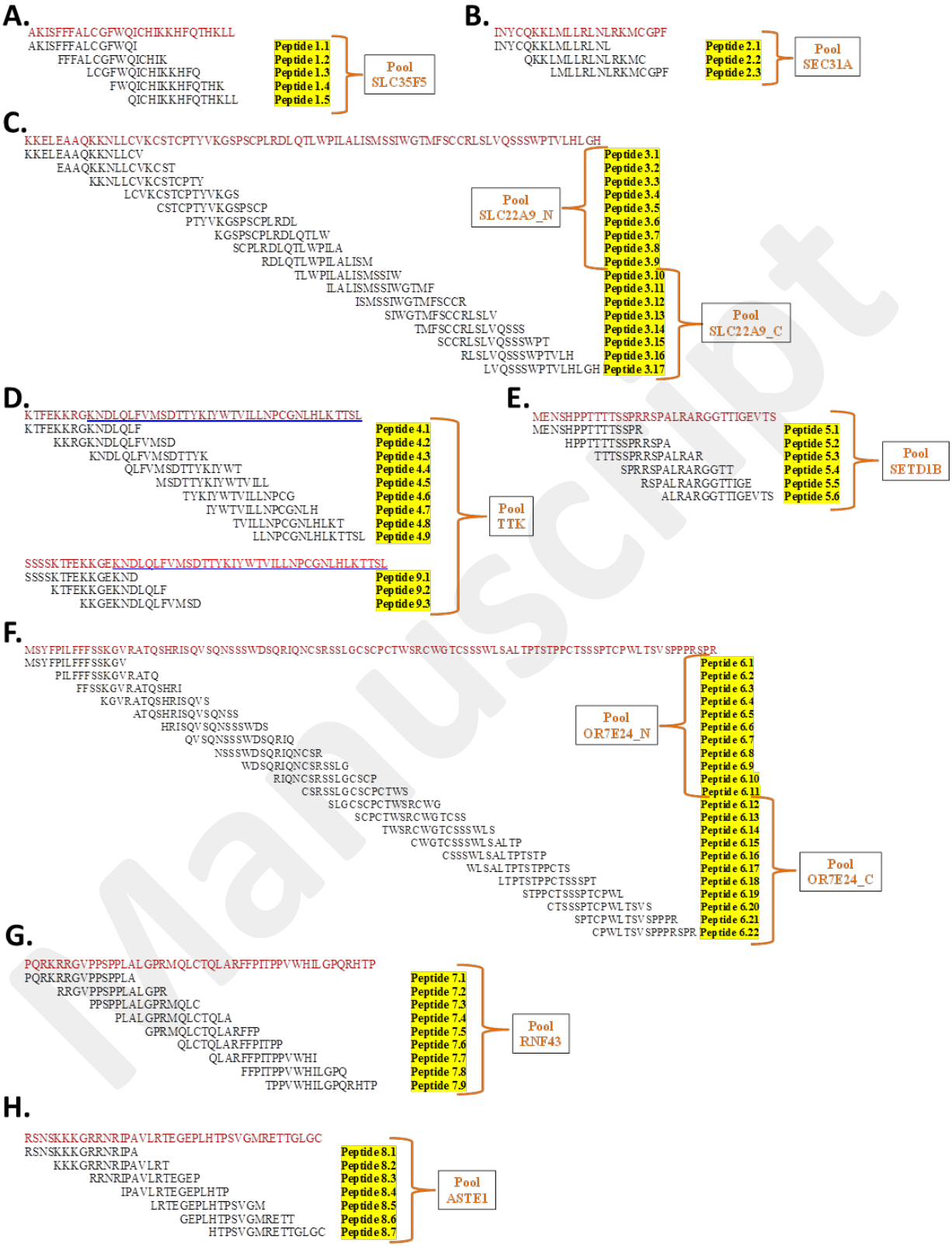
Manuscript


AASDH
ACVR2A
AK9
ASTE1
BMPR2
CCDC168
CCDC73
CDC7
COBLL1
DCAF13
DOCK3
EIF2B3
GRIK2
KMT2C
MSH3
PLEKHA6
RGS12
RNF43
RYR2
SEC31A
SEC63
SETD1B
SLAMF1
SLC22A9
SLC35F5
SLC35G2
SNRNP27
TTK
TVP23A
UBR5
VCP
XYLT2
ZBTB20
RPKM, log
fs neg
fs pos
fs nef

770

771 **Supplementary Figure 6.** RNA expression (RPKM, log2) of genes, carrying shared fs-mutation in
772 cell lines derived from CCLE. Each box-plot represents RPKM expression of gene of interest in
773 shared fs-negative and shared fs-positive cell lines.
774

Manuscript

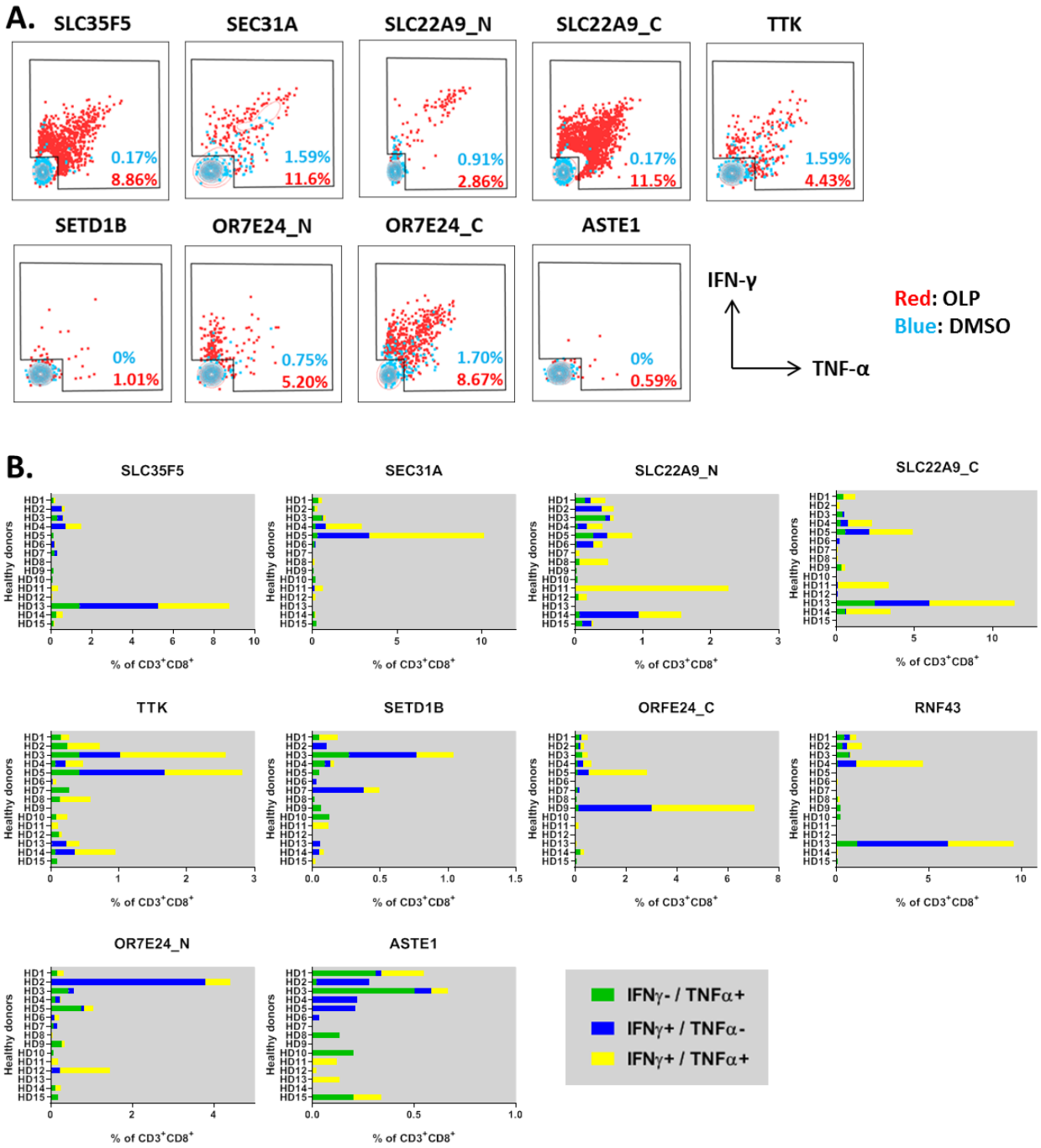


Manuscript

775

776 **Supplementary Figure 7**. Peptide design. 15-mer overlapping peptides spanning the upstream
777 8aa WT sequence and the entire mutated fs-peptide were synthesized for the predicted nine
778 shared fs-peptides from MSI-H UCEC. Peptide pools as they were utilized in the
779 immunogenicity assays were denoted. **A.** SLC35F5, **B.** SEC31A, **C.** SLC22A9**, D.** TTK**, E.**
780 SETD1B, **F.** OR7E24, **G.** RNF43, **H.** ASTE1.
781

Manuscript



Supplementary Figure 8. Fs-peptide-specific T cells are polyfunctional. PBMCs from healthy donors (HD) were expanded *in vitro* following stimulation with fs-peptide OLPs. DMSO was used vehicle control. **A.** Representative flow cytometry overlay plots for IFN-γ (y axis) and TNF-α (x axis) production by CD8+ T cells. Cytokine production upon OLP or DMSO

787 stimulation was shown in red or blue, respectively. **B.** Frequencies of CD8+ T cells producing
788 IFN-γ or TNF-α in response to each fs-peptide OLP pool were plotted for each subject. Green
789 denotes HD producing only TNF-α, blue denotes HD producing only IFN-γ and yellow denotes
790 HD producing both IFN-γ and TNF-α. % IFN-γ values were calculated by subtracting the values
791 obtained after DMSO stimulation from the values after OLP pool stimulation and negative
792 values were set to zero.

Manuscript