Immunopeptidomics-guided identification of functional neoantigens in non-small cell lung cancer

*Ben Nicholas[1,2], *Alistair Bailey[1,2], Katy J McCann[3], Oliver Wood[3], Eve Currall[4], Peter Johnson[5], Tim Elliott[2,6], Christian Ottensmeier[3,4] and Paul Skipp[1]

[1]Centre for Proteomic Research, Biological Sciences and Institute for Life Sciences, Building 85, University of Southampton, Southampton, United Kingdom

[2]Centre for Cancer Immunology and Institute for Life Sciences, Faculty of Medicine, University of Southampton, Southampton, United Kingdom

[3]School of Cancer Sciences, Faculty of Medicine, University of Southampton, Southampton, United Kingdom

[4]Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, United Kingdom

[5]Cancer Research UK Clinical Centre, University of Southampton, Southampton, United Kingdom

[6]Oxford Cancer Centre for Immuno-Oncology and CAMS-Oxford Institute, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

*These authors contributed equally

ORCIDs:

Ben Nicholas: 0000-0003-1467-9643

Alistair Bailey: 0000-0003-0023-8679

21 Katy J McCann: 0000-0003-4388-7265

22 Eve Currall: 0009-0008-0400-2376

23 Peter Johnson: 0000-0003-2306-4974

24 Tim Elliott: 0000-0003-1097-0222

25 Christian Ottensmeier: 0000-0003-3619-1657

26 Paul Skipp: 0000-0002-2995-2959

27 Corresponding author: Dr Ben Nicholas email: bln1@soton.ac.uk

28 **Running title:**

29 Proteogenomics identification of neoantigens in lung cancer

30 **Keywords:**

31 HLA, immunopeptidome, non-small cell lung cancer, antigen presentation

32

Abstract

Non-small cell lung cancer (NSCLC) has poor survival even with modern checkpoint inhibitor therapies. Personalised vaccines based on short peptide neoantigens containing tumour mutations are an attractive precision medicine strategy, but identifying therapeutically relevant neoantigens remains challenging, with existing methods yielding positive responses in only 6% of candidates tested.

We developed an immunopeptidomics approach to improve neoantigen identification in 24 NSCLC patients (15 adenocarcinoma, 9 squamous cell carcinoma). We directly identified one neoantigen and using whole exome sequencing, transcriptomics and mass spectrometry-based immunopeptidomics, we filtered predicted neoantigens based on observed cohort HLA peptide presentation. This approach achieved positive functional responses in 5 of 6 patients tested (83% success rate) with 13% of putative neoantigens (9 out of 70) eliciting strong responses. Bayesian modelling of our initial rules-based neoantigen selection further revealed patient specific peptide presentation patterns and propensities.

Our findings demonstrate that incorporating donor specific HLA peptide presentation data substantially improves neoantigen identification success rates and immune response specificity, advancing personalised cancer vaccine development.

Introduction

Lung cancer is the second most common cancer in the UK and despite recent advances in checkpoint inhibitor therapies fewer than 20% of patients survive 5 years, with most surviving less than one year post-diagnosis[1,2].The majority are diagnosed at advanced stages (III-IV) with non-small cell lung cancer (NSCLC) accounting for 85-90% of cases. NSCLC comprises three histological subtypes: adenocarcinoma (LUAD, most common, peripheral), squamous cell

56 carcinoma (LUSC, central), and large cell undifferentiated carcinoma (least common, can form
57 throughout the lung).

58 Four immunotherapies targeting PD-1/PD-L1 are licensed for NSCLC but show reduced efficacy
59 in patients with EGFR or ALK mutations[3,4]. Consequently, immunotherapy typically follows
60 chemotherapy and/or targeted therapy. While LUAD generally shows more favourable survival
61 than LUSC, treatment resistance remains a major challenge.

62 The TRACERx study revealed evolutionary processes underlying this resistance. Whole
63 genome doubling, driven by tobacco smoke and cytidine deaminase activity, protects tumours
64 against high mutation burdens and chromosomal instability[5]. Smoking-induced truncal mutations
65 and deaminase-driven branch mutations create extensive intratumour heterogeneity that
66 predicts recurrence and death, yet confounds biomarker utility for immunotherapy response
67 prediction[6].

68 An attractive strategy for NSCLC treatment is vaccination targeting on HLA presented
69 neoantigens. This approach assumes neoantigens can be identified that expand tumour killing
70 T-cell populations and/or modulate the tumour microenvironment to make T-cell infiltration or
71 checkpoint inhibitors more effective. The personalised nature of neoantigens minimise the risk
72 of off-target effects and autoimmunity. However, direct identification of neoantigens is rare[7,8]
73 and most approaches to neoantigen discovery rely on predicting that a given mutation leads to
74 protein synthesis, antigen processing and HLA presentation. Direct observation is rare in part
75 due to limits in the sensitivity of the mass spectrometry proteomic detection of HLA ligands,
76 known as immunopeptidomics. Moreover, it is estimated that only a small fraction of mutations
77 are actually presented, possibly as low as 0.5% of non-silent mutations[9]. For example, a
78 NSCLC tumour with 600 missense variants might yield only 3 presented neoantigens amongst a
79 lung tissue immunopeptidome of around 60,000 unique class I and II HLA peptides[10]. A typical
80 experiment may identify 3,000 of these peptides. Assuming a hypergeometric distribution, the

81 probability of observing one class I or II HLA neoantigen is about 14%. Or to put it another way,
82 there is around an 86% chance of not seeing any neoantigens in any single mass spectrometry
83 proteomics experiment.

84 Given these odds, much effort has been put into the in silico prediction of mutations that will
85 give rise to neoantigens that would make effective vaccines. There are well established
86 algorithms that can predict the likelihood of a peptide of given amino acid sequence binding to
87 an HLA molecule, and immunopeptidomic evidence of the peptide length preference of peptide
88 for different HLA allotypes[11], and preferential regions of proteins favourable for presentation[12–14].
89 However, even with this knowledge prediction is stymied by the number of potential neoantigen
90 candidates each mutation might yield, creating large lists of candidate peptides. Moreover, the
91 key biochemical and structural parameters of immunogenic neoantigens remain unknown. The
92 best neoantigen prediction models have a success rate such that around 6% of their putative
93 neoantigens are T-cell reactive[15,16], although recent machine learning models report increased
94 predictive power[17–20].

95 Here we adopted a patient-specific approach where, rather than relying solely on various
96 general HLA peptide characteristics, we leveraged each patient's HLA-I and -II
97 immunopeptidomes alongside the cohort level immunopeptidomes to understand their individual
98 presentation propensity. Thus, immunopeptidomics was used as circumstantial evidence of the
99 biological availability of a mutated protein for presentation by HLA. First we mapped the
100 mutational and immunopeptidome landscapes of a cohort of LUAD and LUSC patients. We then
101 predicted HLA-restricted neoantigens using existing algorithms and used immunopeptidomic
102 data from their individual tumours to filter those predictions on the basis of evidence that they
103 could be presented. Neoantigen selection success was increased to 13%, with strong functional
104 responses confirmed in 5 of 6 patients (83.3%) for our neoantigen panels selected for each
105 donor. Finally, we used Bayesian modelling of the cohort's immunopeptidomes to quantify how

106 donor-specific presentation propensity and protein-level evidence combine to determine peptide
107 presentation, providing a statistical framework for understanding our rules-based selection
108 success.

109 These proof-of-concept results demonstrate how the information contained within the
110 immunopeptidome has the potential to enhance proteogenomics strategies for identifying
111 neoantigens for every patient according to their specific presentation propensity, enabling
112 precision neoantigen selection that accounts for each patient's unique presentation biology.

113

114 # Results

115 **A proteogenomics workflow for neoantigen identification**

116 Our NSCLC cohort consisted of 24 patients, 15 LUAD (8 female, 7 male) and 9 LUSC (5
117 female, 4 male). Median age at diagnosis was 69 (See Table 1 and Supplementary Data S1).
118 Tumour tissue and PBMCs were used for HLA typing, whole exome sequencing, RNA
119 sequencing and mass spectrometry proteomics of the HLA immunopeptidome (Figure 1). To
120 identify candidate neoantigens for each patient we developed a workflow that surveyed both the
121 genomic and immunopeptidomic landscapes. Somatic missense variants called from the whole
122 exome sequencing (WES, Supplementary Data S2 and S3) were used to generate a mutanome
123 for each individual against which the HLA immunopeptidome could be searched for direct
124 observation of neoantigens. Variants, gene expression and the patient HLA allotypes were also
125 used for the prediction of putative neoantigens using existing tools[21].

Table 1: Clinical summary of patients in this study with non-small cell lung cancer

Donor	Smoking status	Cancer subtype
A113	Current smoker	Squamous
A114	Ex smoker	Adenocarcinoma
A115	Ex smoker	Squamous
A116	Never smoker	Squamous
A117	Current smoker	Adenocarcinoma
A118	Ex smoker	Adenocarcinoma
A119	Ex smoker	Adenocarcinoma
A120	Ex smoker	Adenocarcinoma
A133	Ex smoker	Squamous
A134	Current smoker	Squamous
A136	Never smoker	Adenocarcinoma
A137	Ex smoker	Adenocarcinoma

A139	Ex smoker	Adenocarcinoma
A140	Ex smoker	Squamous
A141	Ex smoker	Adenocarcinoma
A142	Current smoker	Adenocarcinoma
A143	Ex smoker	Adenocarcinoma
A144	Ex smoker	Squamous
A145	Ex smoker	Squamous
A146	Current smoker	Adenocarcinoma
A147	Current smoker	Adenocarcinoma
A148	Ex smoker	Adenocarcinoma
A152	Ex smoker	Squamous
A153	Ex smoker	Adenocarcinoma

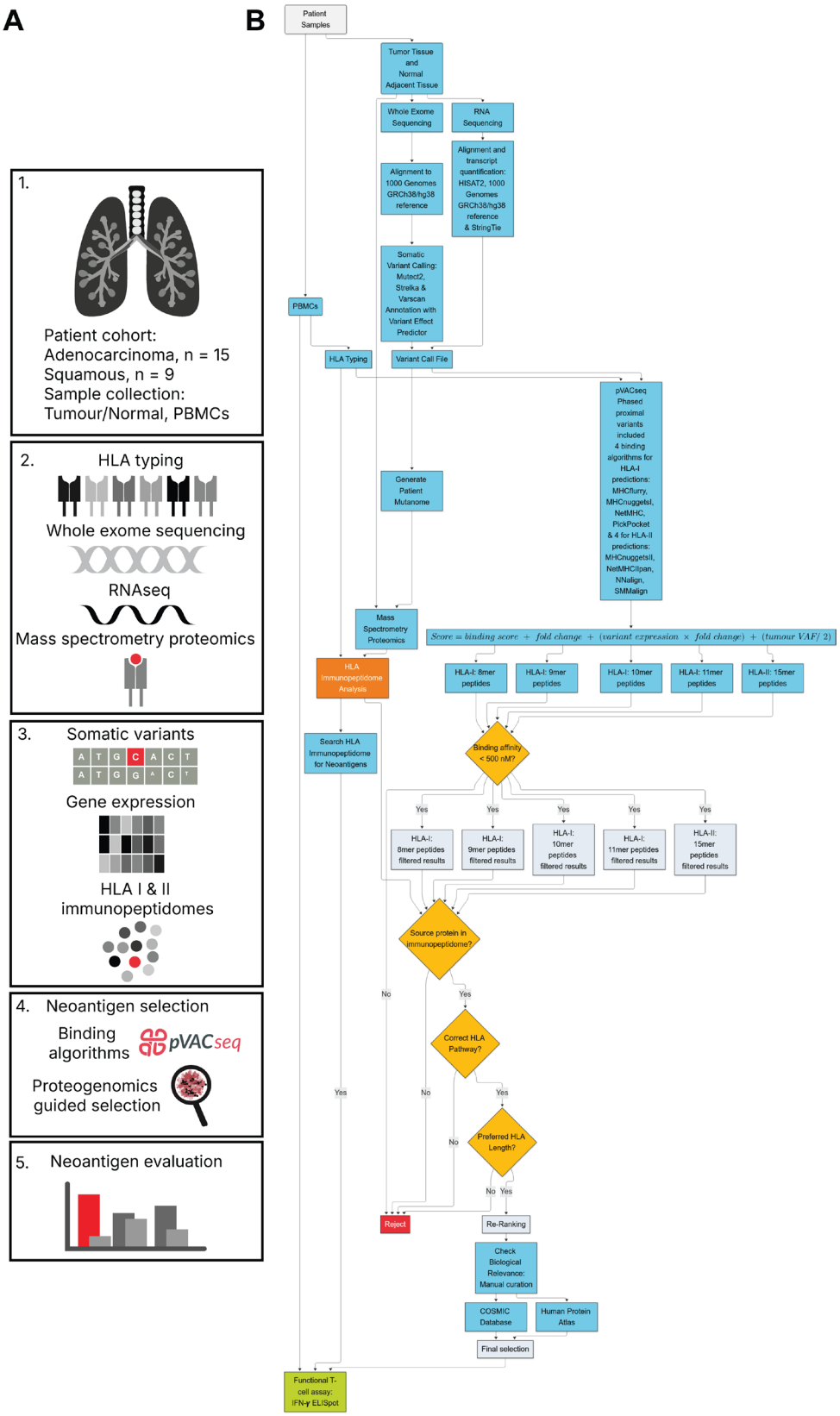
A
B
1.
Patient cohort:
Adenocarcinoma, n = 15
Squamous, n = 9
Sample collection:
Tumour/Normal, PBMCs
2.
HLA typing
Whole exome sequencing
RNAseq
Mass spectrometry proteomics
3.
Somatic variants
Gene expression
HLA I & II
immunopeptidomes
4.
Neoantigen selection
Binding
algorithms
pVACseq
Proteogenomics
guided selection
5.
Neoantigen evaluation
Patient Samples
Tumor Tissue and Normal Adjacent Tissue
Whole Exome Sequencing
RNA Sequencing
Alignment to 1000 Genomes GRCh38/hg38 reference
Alignment and transcript quantification: HISAT2, 1000 Genomes GRCh38/hg38 reference & StringTie
Somatic Variant Calling: Mutect2, Strelka & Varscan Annotation with Variant Effect Predictor
PBMCs
HLA Typing
Variant Call File
pVACseq Phased proximal variants included 4 binding algorithms for HLA-I predictions: MHCflurry, MHCnuggetsI, NetMHC, PickPocket & 4 for HLA-II predictions: MHCnuggetsII, NetMHCIIpan, NNalign, SMMalign
Generate Patient Mutanome
Mass Spectrometry Proteomics
Score = binding score + fold change + (variant expression × fold change) + (tumour VAF/ 2)
HLA Immunopeptidome Analysis
HLA-I: 8mer peptides
HLA-I: 9mer peptides
HLA-I: 10mer peptides
HLA-I: 11mer peptides
HLA-II: 15mer peptides
Search HLA Immunopeptidome for Neoantigens
Binding affinity < 500 nM?
Yes
HLA-I: 8mer peptides filtered results
HLA-I: 9mer peptides filtered results
HLA-I: 10mer peptides filtered results
HLA-I: 11mer peptides filtered results
HLA-II: 15mer peptides filtered results
Source protein in immunopeptidome?
No
Correct HLA Pathway?
Preferred HLA Length?
Reject
Re-Ranking
Check Biological Relevance: Manual curation
COSMIC Database
Human Protein Atlas
Final selection
Functional T-cell assay: IFN-γ ELISpot

Figure 1: Integrated proteogenomics workflow. (A) HLA typing, whole exome sequencing, RNASeq and mass spectrometry-based proteomic of the HLA immunopeptidome were collected for 24 lung cancer patients providing mutational, gene expression and immunopeptidomic data from which to identify candidate neoantigens using binding algorithms and manual inspection of the combined proteogenomic data. (B) Flowchart of neoantigen selection workflow

127

128 **The mutational landscape of NSCLC in the studied cohort**

129 To assess the likelihood of identifying HLA presented neoantigens we first examined the
130 mutational landscape of the NSCLC cohort and found it to be consistent with previous
131 reports[22,23]. Somatic variants were identified by WES of tumour and matched normal adjacent
132 tissues. Tumour mutational burden (TMB) quantifies the number of mutations per million bases
133 (Mb). From WES it is calculated as the number of variants divided by the size of the exome
134 targets; here the target size was 35.7 Mb and the number of variants were either the total
135 number of all variants, or only the protein coding missense variants: $N_{vars}/(35.7) = N_{vars}/Mb$.

136 This revealed that both cancer types have relatively high mutational burdens calculated from all
137 variants, ranging from 27 to 280 mutations per Mb (Mt/Mb) with similar median mutational
138 burdens of 109 Mt/Mb for LUAD and 104 Mt/Mb LUSC, but a broader range for LUAD (Figure 2
139 A, Supplementary Data S1). In terms of missense variants alone, this scales as ranging from 5
140 to 43 Mt/Mb and medians of 20 Mt/Mb for LUAD and 16 Mt/Mb LUSC

141 Approximately one third of all nucleotide transitions and transversions were C>A transversions
142 in both LUSC and LUAD (Figure 2 B), a known mutational signature of smoking[22]. Of the 51,810
143 LUAD and 32,344 LUSC single nucleotide variants, approximately 20% were missense variants
144 (10,565 LUAD, 6,772). These missense SNVs along with approximately 15% insertion/deletion
145 variants (9,697 LUAD, 6,780 LUSC) predict amino acid changes at the protein level and are
146 therefore potential sources of HLA neoantigens (Figure 2 C).

147 For each cancer subtype, patterns of single base substitutions created by the somatic mutations
148 were extracted to identify mutational signatures that were fitted to those identified in COSMIC[24–
149 26] (Figure 2 D-F). LUAD signature A fit SBS36 indicating base excision repair deficiency
150 characterised by C>A transversions. LUAD signature C fit SBS2, which is common in lung
151 cancer and thought to indicate APOBEC cytidine deaminase activity as characterised by C>T

152 transitions. LUSC signature C fit SBS29, another signature characterised by C>A transversions
153 and linked to tobacco chewing.

154 In addition to examining the potential for neoantigen generation at the exon level, we sought to
155 examine the potential for neoantigen recognition within the tumours using bulk gene expression
156 data from RNAseq to assess the fractions of immune cells present in the tumours[27] (Figure 2
157 G). This estimation also provides an indication of the tumour sample purity. All expressed genes
158 not used as markers for immune cells are labelled as 'otherCells' and we would expect this
159 category to comprise the largest proportion of cells in a tumour sample. Therefore if a sample
160 has lower proportion of 'otherCells' it is indicative of a less pure tumour sample. For LUAD and
161 LUSC, the median proportions of 'otherCells' are one third. In cases with very low proportions of
162 'otherCells' such as A134 and A145, the corresponding histology reports indicate these were
163 fibrotic samples consistent with the very high proportions of cancer associated fibroblasts
164 identified by RNAseq. However at the cohort level, proportions of T-cells estimated capable of
165 responding to neoantigens presented by HLA were estimated with similar medians for CD4+ T-
166 cells of 18% and 15% for LUAD and LUSC respectively, and medians for CD8+ T-cells of 2%
167 and less than 1% for LUAD and LUSC respectively.

168 In summary, the mutational landscape of the NSLC cohort is characterised by a high tumour
169 mutational burden in both cancer subtypes, the largest proportion of variants with the potential
170 for generation of neoantigens arising from C>A transversions. Furthermore, gene expression
171 data estimates the presence of limited populations of T-cells with the potential to recognise HLA
172 presented neoantigens.

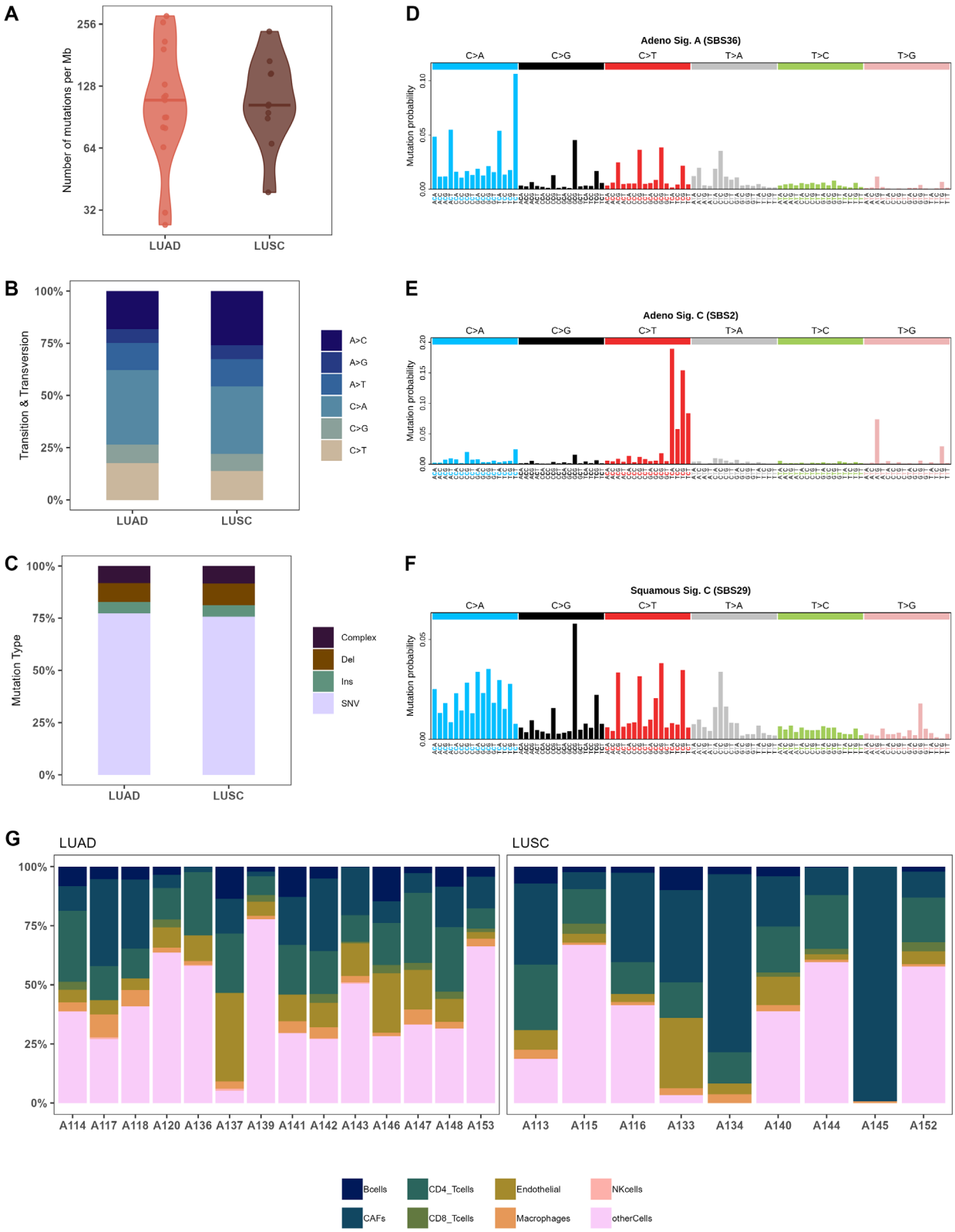
A
Number of mutations per Mb
256
128
64
32
LUAD
LUSC
B
Transition & Transversion
100%
75%
50%
25%
0%
LUAD
LUSC
A>C
A>G
A>T
C>A
C>G
C>T
C
Mutation Type
100%
75%
50%
25%
0%
LUAD
LUSC
Complex
Del
Ins
SNV
D
Adeno Sig. A (SBS36)
C>A
C>G
C>T
T>A
T>C
T>G
Mutation probability
E
Adeno Sig. C (SBS2)
C>A
C>G
C>T
T>A
T>C
T>G
Mutation probability
F
Squamous Sig. C (SBS29)
C>A
C>G
C>T
T>A
T>C
T>G
Mutation probability
G
LUAD
LUSC
100%
75%
50%
25%
0%
A114 A117 A118 A120 A136 A137 A139 A141 A142 A143 A146 A147 A148 A153
A113 A115 A116 A133 A134 A140 A144 A145 A152
Bcells
CD4_Tcells
Endothelial
NKcells
CAFs
CD8_Tcells
Macrophages
otherCells

Figure 2: The mutational landscape of lung cancer in the studied cohort. (A) The mutational burden of each cancer type: squamous (n=9) and adenocarcinoma (n=15). (B) Mutation frequency of six transition and transversion categories for each cancer type. (C) Mutation frequencies each cancer type. (D-F) Mutational signatures identified in each cancer subtype. (G) The proportions of immune cells estimated from bulk tumor RNASeq in each tumour sample.

173 # The peptidome landscape of NSCLC in the studied cohort

174 Mass spectrometry proteomics of the HLA immunopeptidomes identified large distributions of
175 peptides with their characteristic modes of 9 amino acids (AA) and 15 AA for class I and II HLA
176 peptides respectively (Figure 3 A). Median class I immunopeptidome sizes were 5422 and 2998
177 for Adenocarcinoma and Squamous NSCLC respectively. Median class II immunopeptidome
178 sizes were 2849 and 1125 for Adenocarcinoma and Squamous NSCLC respectively.

179 **The lung cancer peptidome resembles the healthy lung tissue peptidome**

180 We compared the distinct source protein populations yielding the class I and II HLA peptidomes
181 between our LUAD, LUSC samples and healthy lung tissues from the Human HLA Ligand
182 Atlas[10] to examine their similarities and differences (Figure 3 B-C, Supplementary Data S5) ,
183 considering only proteins present in at least two-thirds of our samples peptidomes. Our analysis
184 suggests that healthy lung and tumour tissues immunopeptidomes sample largely the same
185 protein populations. 92% of HLA-I proteins and 52% of HLA-II proteins were common to all
186 three tissue types. The remaining proteins most likely represent experimental variation.

187 **Direct identification of a neoantigen in a LUAD patient immunopeptidome**

188 Across the cohort of 24 patients we identified a single missense variant product by direct mass
189 spectrometric observation in the class I HLA immunopeptidome of one LUAD patient (A147)
190 (Figure S1-S3). This derived from a C>A variant in the ALYREF gene yielding an Asp10Tyr
191 mutation in its protein product THO complex subunit 4 (Uniprot: Q86V81). This mutation yielded
192 seven nested 15-18mer peptides with the mutation Y before the start of core sequence of
193 MSLDDIIKL. No wild type peptides were observed for this protein in either the HLA I or II
194 immunopeptidomes, suggesting this mutation altered either the binding affinity of these peptides
195 or the source protein processing in the antigen processing pathway. The rarity of this
196 observation is in keeping with estimates of frequencies in the order of 0.5% of missense
197 variants encoding presented neoantigens[7,9]. The length of the ALYREF peptides suggested
198 these may be class II HLA peptides that we had captured by chance in this assay. The motif
199 most closely matched the patients HLA-DRB1*03:01 allotype with peptide AYKMDMSLDDIIKLN
200 predicted as a weakly binding peptide[28]. We identified 1135 missense mutations for patient
201 A147 (Supplementary Data S1). An estimate of 0.5% missense yielding neoantigens would
202 represent 6 neoantigens. Here we observed one.

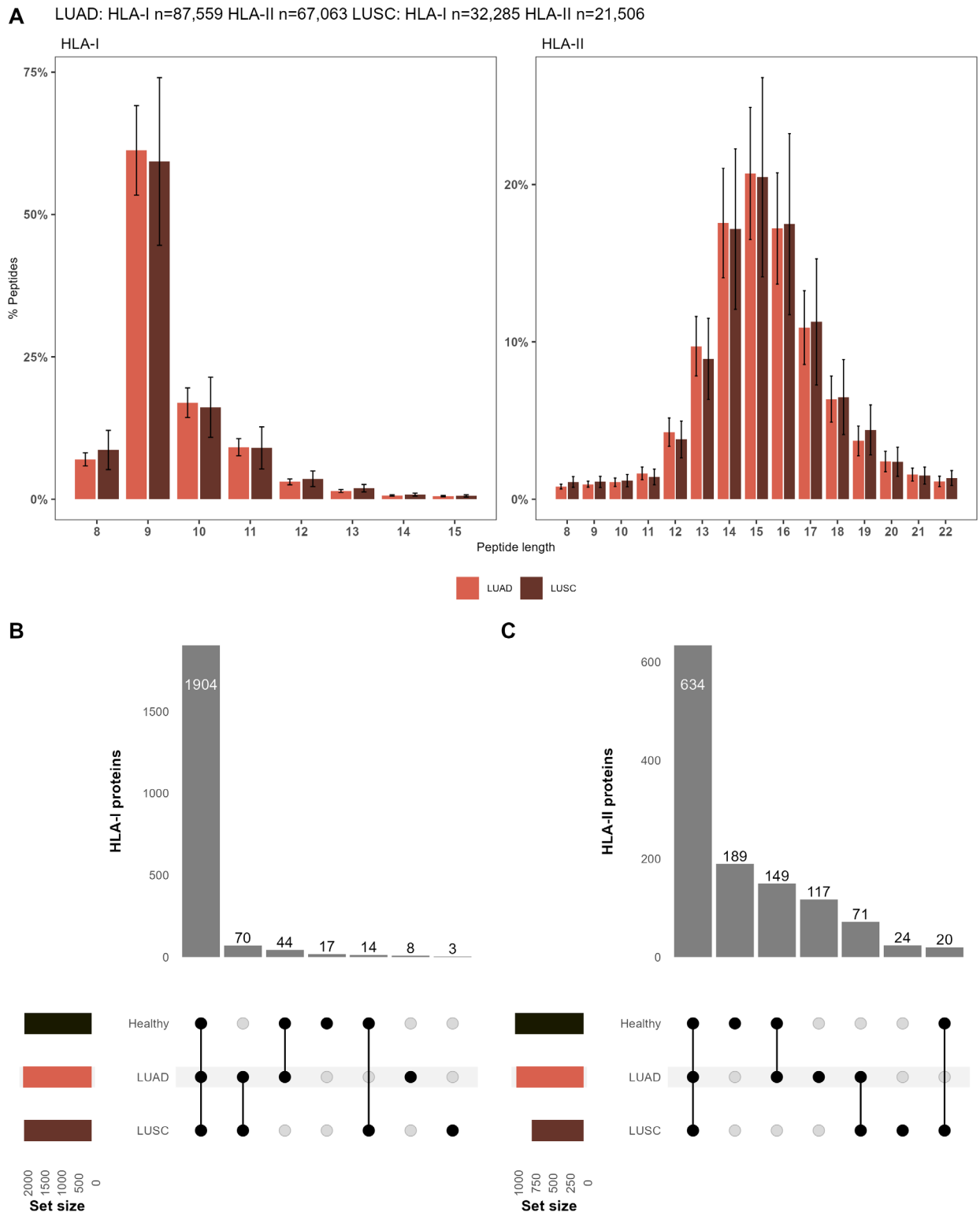

Figure 3: The peptidome landscape of lung cancer. (A) Length distributions of

immunopeptides from tumour tissues. (B-C) Upset plots of proteins presented by HLA molecules (class I left, class II right) comparing proteins between cancer subtypes and healthy lung tissues from the HLA ligand atlas. Proteins were included if observed in at least two-thirds of the peptidomes for each cancer subtype.

203

204 # A proteogenomics view of NSCLC in the studied cohort

205 Consistent with the view that non-silent mutations rarely encode HLA presented neoantigens[9],
206 we observed that LUAD and LUSC driver genes[29] are not mutated and presented by HLA
207 molecules with the same frequencies. Some drivers are frequently mutated, but rarely
208 presented e.g. APC, whereas others are rarely mutated, but frequently present in the HLA
209 immunopeptidomes e.g. KEAP1 (Figure 4 A). TP53 is both frequently mutated and presented in
210 the class I HLA peptidomes of both NSCLC subtypes (Figure 4 A).

211 We found that loss of class I HLA heterozygosity in the genome[30] is reflected in the peptidome.
212 In heterozygous patients, immunopeptides identified as presented by HLA molecules from the
213 retained allele were observed at higher proportions in the peptidome than from the lost allele for
214 HLA-A and B allotypes (Figure 4 B, Supplementary Data S4).

215 We also found that mutations are distributed across the cellular compartments at the same
216 frequencies as the genes are expressed (Figure 4 C left), but the HLA pathways sample the
217 compartments preferentially. Class I HLA immunopeptides are derived preferentially from
218 nuclear and cytosolic proteins, whilst class II HLA immunopeptides are derived preferentially
219 from membrane and extracellular proteins (Figure 4 C right).

220 Collectively, these observations imply that the likelihood of a putative neoantigen being
221 presented by either HLA class is influenced by the cellular compartment origin of the source
222 protein and secondly, putative neoantigens with motifs[31] for the retained class I HLA allotypes
223 are more likely to be presented than those from the lost allotype.

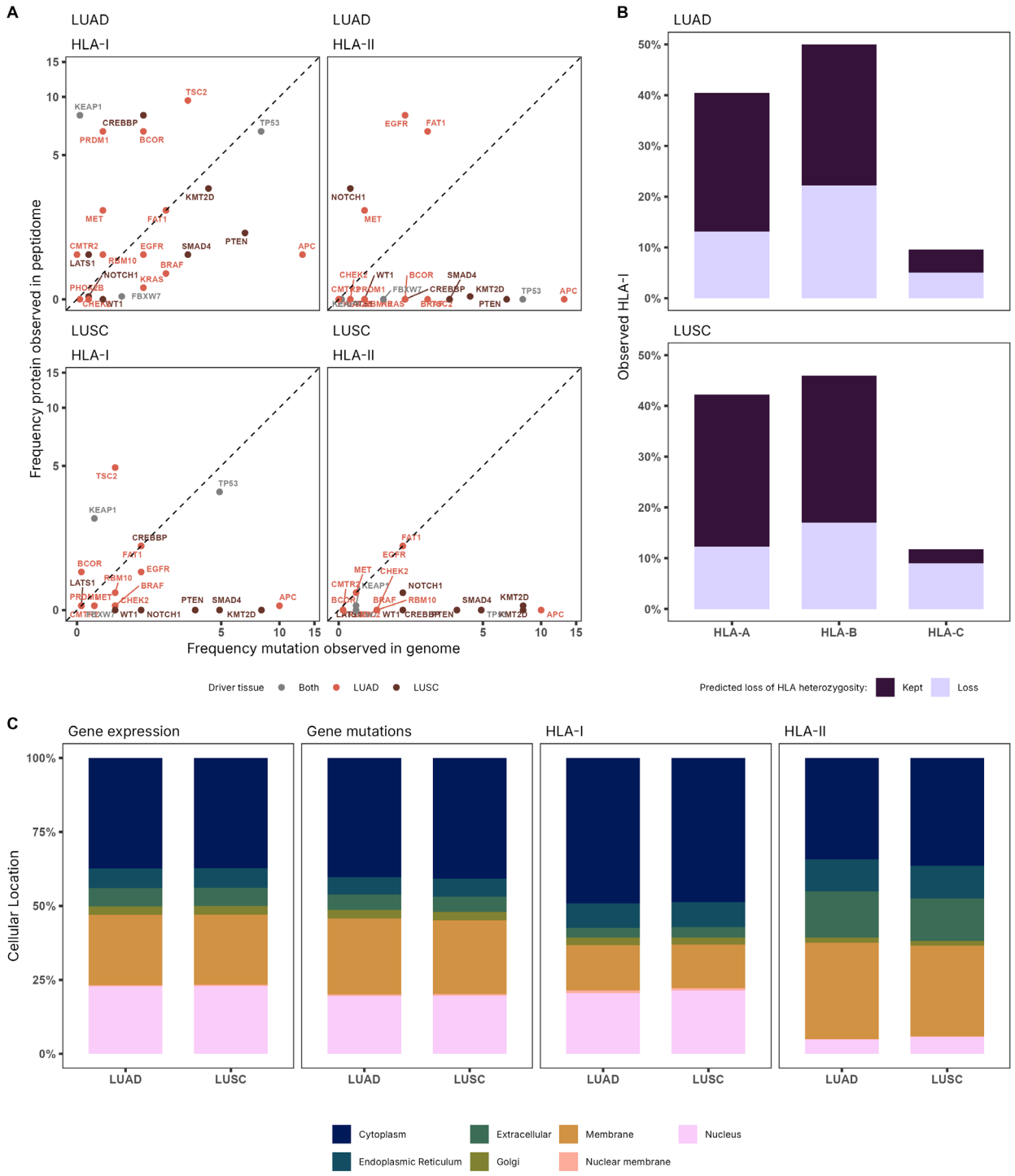

Figure 4: Integrating the mutational and immunopeptidome landscape reveals previously unclear relationships between mutations and peptide presentation. (A) Parity plots of the frequency of mutated initiating driver genes for each cancer type plotted against the

frequency of observed presentation from the corresponding protein in either the HLA-I (left) or HLA-II peptidome (right). Parity of frequency is indicated by the dotted line. The colour indicates which cancer type the gene was identified as an initiating driver in[5]. (B) The relative proportions of immunopeptides assigned to each of the two HLA-A, B and C allotypes for heterozygous patients.[7,31,32]. The colour represents whether an allele is predicted to have loss of heterozygosity in the genome[30]. (C) Comparison of the frequency between the cellular compartments in which gene expression and somatic mutations occur, and those from which HLA peptides are observed in each cancer type.

224 # Proteogenomics guided NSCLC neoantigen selection and
225 # testing

226 In selecting neoantigens we initially used the pVACseq tool to create a list of putative
227 neoantigens for each patient and HLA allotype and different peptide lengths[33]. Briefly, we used
228 pVACseq with the whole exome and transcriptome outputs and patient HLA allotypes to predict
229 8-11mer peptides for class I HLA and 15-mer peptides for class II HLA-DRB allotypes across
230 eight binding algorithms. This combined genomic and binding score creates an overall score for
231 each peptide (Details in Section 1.8.0.10). For our 24 patients this comprises 524 HLA class I
232 tables and 74 HLA class II tables of ranked predictions. Discarding any prediction for a peptide
233 with >500 nM binding affinity, pVACseq yielded 27,466 class I HLA and 127,015 class II HLA
234 predicted neoantigen peptides (Supplementary Data S6-S8)

235 We were able to test predictions for six patients, but this still required selecting from thousands
236 of possible candidate peptides. We consequently filtered the candidate peptides according to
237 whether peptides arising from the gene product with a missense variant were already present in
238 the patients' respective class I or class II HLA peptidome (Supplementary Data S3) and

239 according to HLA peptide length preferences[34]. This reduced the number of candidates to a few
240 hundred peptides for each patient. We finally manually curated the ranked peptide candidates
241 for biological relevance using auxiliary information from the literature, the Human Protein Atlas
242 and COSMIC. (Figure 1, Section 1.8.0.10).

243 Our exploratory filtering process for candidate neoantigens can be summarised as: Does a
244 missense mutation exist? Is there evidence that the mutated gene product enters the antigen
245 processing pathway for presentation, and if so in which HLA pathway? Is the candidate
246 neoantigen of the preferred HLA allotype length? Is the candidate neoantigen predicted to bind
247 to the HLA allotype according to pVACseq? To preferentially select one neoantigen candidate
248 over another in the final ranking we looked for evidence of biological relevance for the candidate
249 in COSMIC[26] and the HPA[35].

250 As HLA peptidome observation took precedence over pVACseq rank, some candidates such as
251 peptide 08-FAT1 were low ranking (70th percentile) but still with a predicted binding affinity
252 lower than 500 nM (Table 2).

253 For six patients, 3 LUAD and 3 LUSC, we selected 9 to 14 putative neoantigens per patient (70
254 in total) and synthesised the specific putative HLA-I or HLA-II peptides in the mutant neoantigen
255 and wildtype forms (Supplementary Data S9). Using their donor matched PBMCs, we identified
256 nine strong neoantigen specific responses to putative neoantigens in five out of six patients,
257 including for the directly observed ALYREF peptide (Figure 5 A-F, Table 2). This represents an
258 83% success in achieving a donor functional response and a 13% response rate in antigenic
259 peptide selection, twice the genomics-based peptide prediction rate of 6% reported in the
260 literature[15]. We observed responses to both class I and class II HLA candidate neoantigens in
261 LUAD (Figure 5 A-C), but only class II HLA candidate neoantigens in LUSC (Figure 5 E-F).
262 LUSC patient A116 yielded no responses (Figure 5 D).

Table 2: Peptides yielding IFN-γ ELISPOT responses

Tissue	ID / HLA / Peptide Length[a]	Peptide[b]	Rank %[c]	Peptidome support[d]	Auxiliary support[e]	ELISpot response
LUAD	A119 / DRB1*04:04 / 15	01-CANT1	11	II	HPA	Strong
LUAD	A119 / HLA-A*31:01 / 10	12-PTPRT	10	I	COSMIC	Strong
LUAD	A147 / Observed / 15	01-ALYREF	-	I	-	Strong
LUAD	A147 / DRB1*04:01 / 15	08-FAT1	70	I+II	COSMIC	Strong
LUAD	A147 / HLA-A*01:01 / 9	14-TP53	7	I	COSMIC	Weak
LUAD	A148 / HLA-A*26:01 / 9	01-KMT2C	20	I	COSMIC	Strong
LUAD	A148 / DRB1*01:01 / 15	05-NT5E	6	I+II	HPA	Strong
LUSC	A134 / DRB1*01:03 / 15	06-KRT8	10	I+II	-	Strong
LUSC	A144 / DRB1*04:01 / 15	04-FAT1	24	I+II	COSMIC	Strong
LUSC	A144 / DRB1*04:01 / 15	08-NF1	3	I+II	COSMIC	Strong

[a]The patient ID, predicted HLA allotype for the peptide and peptide length. ALYREF was an observed peptide.

[b]The peptide identifier and gene name corresponding with those in Figure 5.

[c]Rank % is the rank of the peptide in the table for that Donor and HLA allotype, a lower rank corresponds with better pVACseq score as detailed in the materials and methods.

[d]Peptidome support indicates from which class HLA peptidome source protein peptides were observed.

[e]Auxiliary support indicates support for biological relevance either from the lung cancer associated proteins in the Human Protein Atlas (HPA) or the Top 20 mutated genes in COSMIC.

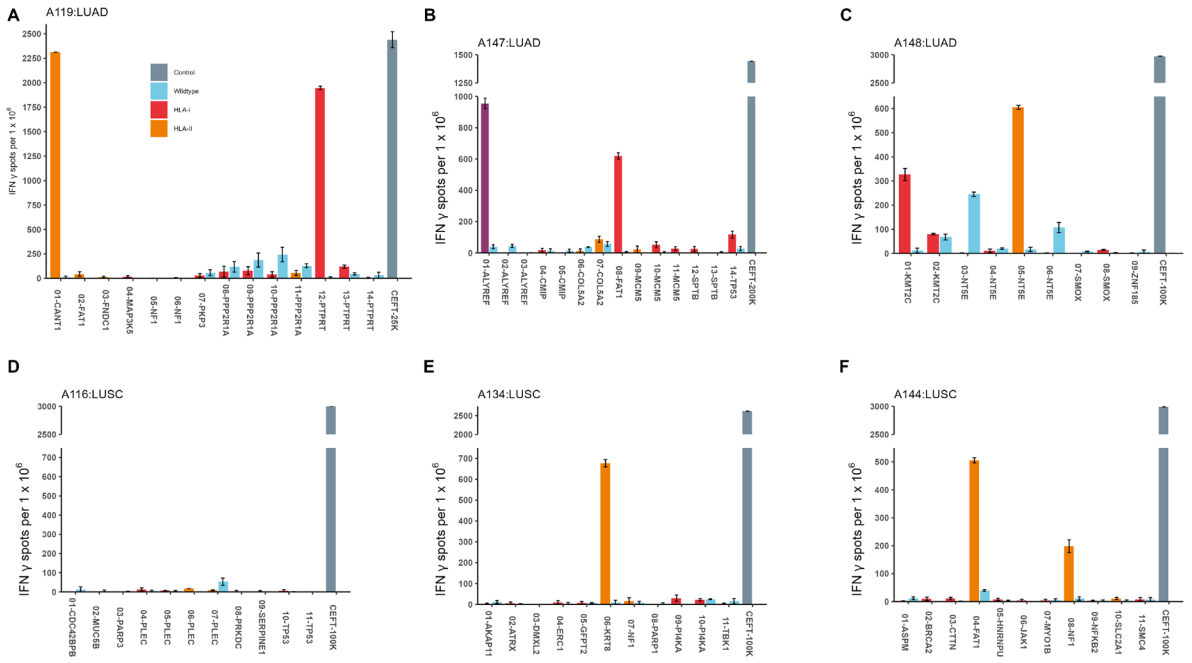

Figure 5: Proteogenomics guided NSCLC neoantigen selection identifies nine strong candidates (A-F) IFN-γ ELISPOT of putative neoantigens three LUAD and three LUSC patients donor matched PBMCs. Wildtype peptides are represented by blue bars, putative HLA-I neoantigens by red bars, putative HLA-II neoantigens by orange bars, the observed ALYREF neoantigen in purple and the CEFT control peptide mix in grey.

263 **Model of peptide presentation**

264 To retrospectively examine the mechanisms underlying our rules based selection approach and
265 quantify the relative importance of factors affecting peptide presentation, we created Bayesian
266 multilevel regression models of the NSCLC immunopeptidomes[36]. The goal was to understand
267 the utility of immunopeptidomes in the context of neoantigen selection by quantifying what
268 determines observed peptide presentation.

269 Several features have been identified as causal: protein abundance, protein hotspots, peptide
270 length preferences of HLA molecules, peptide binding motifs and other features of antigen
271 processing proteins such as tapasin, TAPBR and TAP on HLA-I peptide selection[37–41].

272 Our findings indicated that driver genes do not correlate with peptide presentation (Figure 4 A)
273 and that the presented proteome of LUAD, LUSC and healthy lung tissues are highly similar
274 (Figure 3 B-C). Additionally, as others have also found, direct observations of neoantigens are
275 rare[7,9]. Although not exclusively, we found that proteins were preferentially presented via HLA
276 pathways corresponding with the cellular compartment they occupy (Figure 4 C).

277 To construct the models we began with the source proteins, which we did not observe directly.
278 Proteins from the underlying proteome were latent variables from which peptides derived, bound
279 to and were presented by HLA molecules. Proteins were represented in the form of peptide
280 counts for each protein observed, for each HLA allotype and donor. We assumed that there
281 were relationships between donor or NSCLC subtype cohort and protein properties
282 e.g. abundance, with other peptides on binding to HLA and presentation, represented as a
283 causal model by a directed acyclic graph (DAG) (Figure 6 A). Arrows indicated the direction of
284 the causal effects the model assumed. The DAG indicated several sources of confounding,
285 specifically the need for the model to adjust during fitting for protein, donor, cohort, peptide and
286 HLA specific effects in estimating the probability of each peptide being presented. Within the
287 limits of detection, a peptide was either presented or not, hence we assumed this outcome
288 variable (yes or no) as being a Bernoulli distribution arising from a linear combination of our
289 predictor variables: protein, donor, cohort, peptide and HLA properties. Importantly, having
290 assumed the outcome distribution, Bayes theorem was used to pool observations across each
291 subtype cohort and share information to adjust the estimation for each peptide as indicated by
292 the DAG. Pooled adjustment helped to account for differing immunopeptidome sizes between
293 donors and also enabled estimation of the probability of presentation of a peptide even if it was

294 missing in any given immunopeptidome. These models quantified the relative importance of
295 donor specific versus protein specific factors in determining peptide presentation.

296 Model diagnostics confirmed robust estimation with high signal-to-noise ratios (SNR > 10),
297 adequate effective sample sizes, and good MCMC convergence across all models
298 (Supplementary Tables S1-S2, Figures S4-S11).

299 To examine the relative importance of donor differences, protein properties and donor-specific
300 protein preferences, we calculated the contribution to the total variance of each model
301 component (Figure 6 B). We found moderately greater contributions of donor effects to protein
302 effects for LUAD HLA-I and HLA-II presentation, ratios of 1.3 and 3 times greater respectively
303 (Figure 6 C). For the LUSC cohort, donor effects were much larger contributors to presentation
304 variance than protein effects for both HLA-I and II, ratios of 8 and 22 times respectively
305 (Figure 6 C). Donor-specific protein contributions varied and were greatest for HLA-II
306 immunopeptidomes, and was the largest variance component for LUAD HLA-II.

307 To further quantify these effects, we calculated the average difference in peptide presentation
308 probability for each donor relative to median donor presentation (Figure 6 D). For LUAD, most
309 donors showed HLA-I presentation close to the median, but differences in the order of 50% for
310 some donors with respect to median presentation for HLA-II, albeit with much greater
311 uncertainty. The model estimated donors such as A148 and A147 were more likely to present
312 peptides, despite A148 and A147 having only the 6th and 10th highest number of HLA-I and 8th
313 and 9th number of HLA-II peptide observations respectively (Supplementary Data S1). The
314 estimates for LUSC donors showed greater variation and uncertainty, with most donors below
315 the median. Donor A116 had the greatest number of LUSC HLA-I peptide observations, and 3rd
316 greatest LUSC HLA-II peptide observations (Supplementary Data S1), but the model estimated
317 A116 below the median for presentation for both HLA classes in LUSC, and we were unable to
318 select a reactive neoantigen peptide for this donor.

319 In our rules-based selection of neoantigens we ranked a prediction more highly if a donor
320 presented peptides from the same protein as the putative neoantigen (Figure 1 B). To quantify
321 this effect and examine our rules-based assumption, we calculated the average effect on
322 peptide presentation of observing multiple peptides from the same protein across each NSCLC
323 subtype cohort versus the effect of observing more than one peptide from the same protein
324 presented by a donor.

325 The model identified competing effects such that when a protein presented multiple peptides
326 across the cohort, this decreased the probability of observing an additional peptide by between
327 18-23% depending on NSCLC subtype and HLA class (Figure 6 E). In contrast, a donor
328 presenting one peptide from a protein increased the probability of observing another peptide
329 from the same protein by between 23-41% depending on NSCLC subtype and HLA class
330 (Figure 6 E), supporting our rules-based assumption. The model quantified that donor-specific
331 protein preferences substantially influenced presentation probability, but also revealed that
332 cohort-level presentation patterns created competing effects.

333 In summary, by pooling information across immunopeptidomes in the NSCLC cohorts, the
334 models identified NSCLC subtype variation and donor-specific effects on peptide presentation:
335 each donor's general propensity to present peptides and their protein-specific preferences.

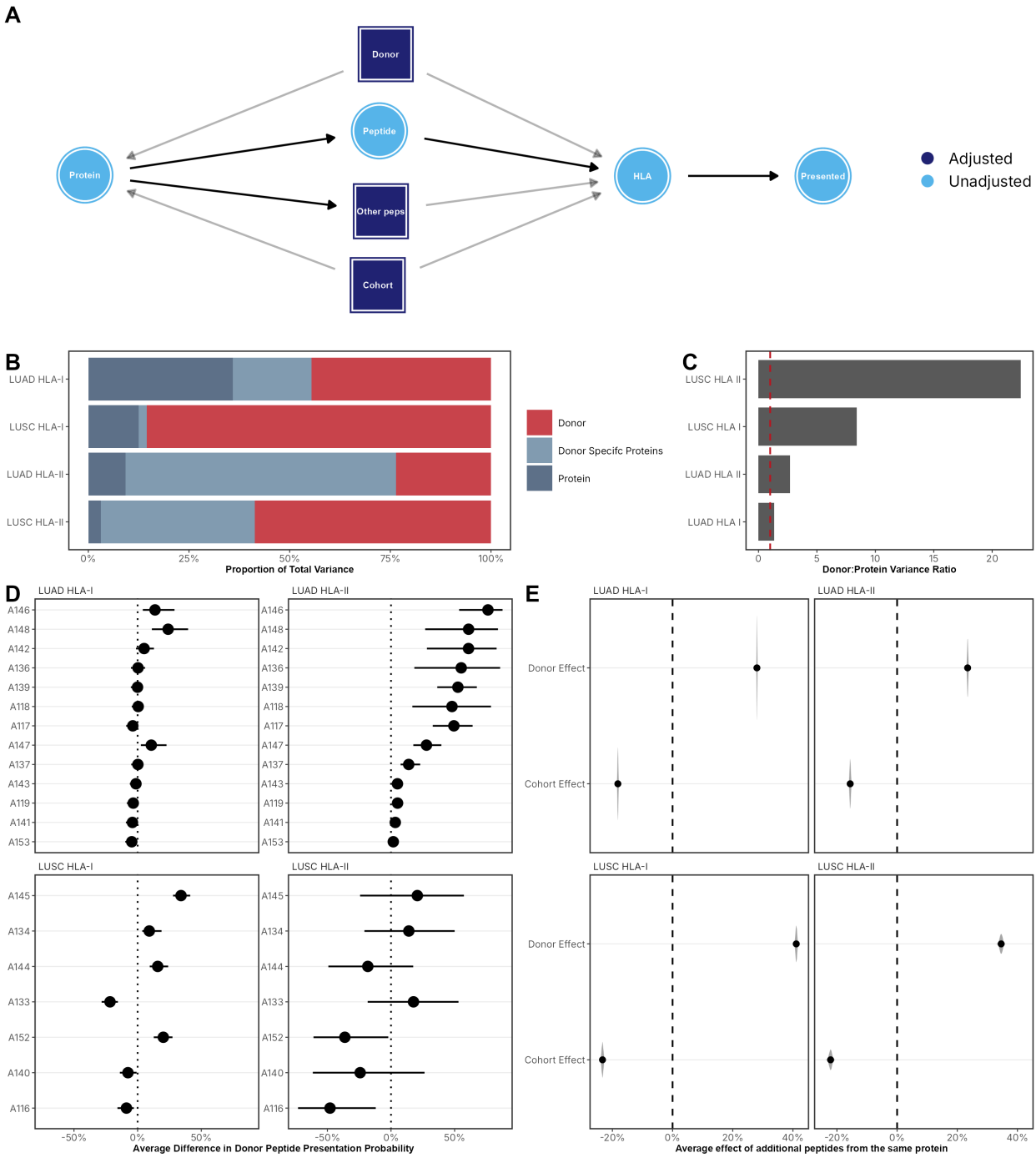

Figure 6: Bayesian multilevel regression model of HLA peptide presentation (A) Directed Acyclic Graph of HLA presentation for NSCLC cohorts. Variables adjusted for confounding in dark blue, unadjusted variables in light blue. (B) Contributions of donor differences, protein properties and donor specific proteins to the total variance in

peptide presentation in Bayesian multilevel regression models of NSCLC immunopeptidomes. (C) Barplot of the ratio between donor and protein variance components in the peptide presentation models. The red dashed line indicates where these components are of equal value. (D) Comparison of the average propensity of donor to present peptides with respect to median presentation for each NSCLC subtype and HLA class. (E) Comparison of the average effect on the probability of presentation of an additional peptide from the same protein if a donor has already presented a peptide versus multiple peptides from that protein being presented across the cohort for each NSCLC subtype and HLA class.

336 # Discussion

337 The strategy of using HLA presented peptides as a basis for immunotherapy is long
338 standing[42,43]. Researchers have sought to identify either peptides common to a cancer type, so-
339 called tumour associated antigens, or peptides unique to a patient's tumour, so called
340 neoantigens. Here we sought to improve our ability to identify HLA presented neoantigens in
341 two NSCLC sub-types using a proteogenomics approach that combines exome sequencing,
342 transcriptomics, mass spectrometry immunopeptidomics and mechanistic modelling. In tumour
343 samples from cohort of 24 NSCLC patients we found relatively high mutational burdens with
344 exonic mutations characterised by a predominance of C>A transversions and containing small
345 populations of T-cells. Consistent with previous reports[8,9] using mass spectrometry
346 immunopeptidomics, we only directly identified one neoantigen amongst tens of thousands of
347 peptide identifications. However, following a rules based approach, we utilised the remaining
348 observations to inform our selection of neoantigens from ranked lists generated by in silico
349 prediction algorithms[33] to the extent that we were able to identify positive functional assay
350 neoantigens for 5 out of the 6 patients (83%) we were able to test. This included a positive

351 response for the directly observed neoantigen. These findings also represent a two-fold
352 improvement over previous reports for neoantigen prediction and identification[15,16]. We then
353 applied Bayeisan inference to retrospectively examine our rules based neoantigen selection and
354 quantify patient specific differences in peptide presentation propensity.

355 Although identifying patient specific neoantigens and was the primary goal of our study, the data
356 included some interesting related observations: With the exception of TP53 and FAT1 in the
357 HLA-I and HLA-II immunopeptidomes respectively, there was no correlation between driver
358 gene mutation frequency and their peptide presentation frequency. This provides some
359 circumstantial support for these two genes as sources for neoantigens[44,45]. TP53 mutations can
360 be either truncal or late-stage[46], but a third of TP53 mutations occur in so-called hotspot
361 regions[47], making them of interest both for early detection and as targets for immunotherapy[48,49].
362 Overall we found that the tumour peptidomes contained peptides derived from the same source
363 proteins as healthy tissue. Furthermore, although somatic mutations are not random, as seen in
364 the mutational signatures and driver genes, their distribution amongst cell compartments
365 corresponds with gene expression frequency. There is no enrichment for mutations in genes
366 expressing proteins in specific cell compartments. This implies a connection between the cell
367 compartment from where the source protein derives and the subsequent HLA antigen
368 processing pathway it primarily feeds, consistent with previous reports[50]. TP53 is a
369 predominantly a nuclear protein, whilst FAT1 is predominantly extracellular, hence their higher
370 frequencies in HLA I and II immunopeptidomes respectively. Whilst this might seem tautological,
371 it does indicate that it would be unwise to preferentially select class I neoantigen predictions for
372 FAT1 and vice versa for TP53, and yet this is not explicitly considered in existing neoantigen
373 prediction algorithms. Hence, we selected source protein cell compartment as a relevant
374 neoantigen parameter in our rules based approach. This choice was supported by the
375 mechanistic model prediction of an increased probability of observing another peptide from the

376 same protein if a patient has presented one peptide already by between 23-41%. However, our
377 mechanistic model also identified competing effects between proteins that were widely
378 presented in NSCLC subtype and those proteins specific to each donor, highlighting the need
379 for patient specific predictions that aren't solely reliant on general HLA peptide characteristics.

380 Loss of class I HLA heterozygosity in the genome was reflected in the proportions of peptides
381 observed for each HLA-I allotype, and although we did not use this information to select
382 neoantigens, this might be another useful parameter when ranking candidates.

383 There are various limitations in our study which might be addressed by future studies. Perhaps
384 most significantly from a methodological perspective, mass spectrometry as a technique does
385 not have the amplification step found in many genomic sequencing methodologies. Therefore
386 the strength of the input signal arises almost entirely from sample quality and preparation and
387 sensitivity is determined by the mass spectrometer itself. The complexity of the input mixture
388 and the differential ability of peptides to ionise, along with their relative abundances all affect
389 what fraction of the immunopeptidome is observed. Various single molecule technologies are
390 being developed that may address this problem, of which pore-based technologies, possibly in
391 combination with fluorescence fingerprinting, seem well suited to identification of short
392 peptides[51–53]. Sequencing peptides using pore technologies offers the tantalising prospect of
393 providing much greater coverage of the immunopeptidome, and therefore direct observation of
394 neoantigens. There are many challenges to this approach, not least post-translational
395 modifications and the non-polar nature of protein peptides, but much progress has already been
396 made[54–56].

397 In our study we only considered canonical neoantigens arising from missense variants. This
398 was a limitation largely arising from choosing whole exome sequencing, but there is increasing
399 evidence for non-canonical antigens arising from non-coding regions of the genome[57–59]. There
400 are several sources of non-canonical peptides such as circRNA translation, unannotated open

401 reading frames (nuORFs), lncRNA and alternative reading frames. As yet, clinical efficacy is
402 unknown and it is also unclear how abundant non-canonical peptides are in the
403 immunopeptidome due their precursor instability and/or relative abundance[60]. Likewise,
404 detection of non-canonical peptides is further hampered by the increase in search space they
405 represent and subsequent impact on detection sensitivity, though this is being addressed[61].
406 Immunopeptidomic identification of circRNA and cancer specific nuORFs peptides have been
407 reported[63] and immune responses in mice to pancreatic cancer nuORF peptides[64]. Therefore
408 we'd expect inclusion of non-canonical targets to be part of future immunopeptidomic studies.

409 Here we have identified potential candidates for personalised vaccines that elicit strong positive
410 responses in functional T-cell assays, however further evaluation is required to determine their
411 clinical efficacy, and their effectiveness as vaccines cannot be guaranteed. The stage of the
412 cancer at which the patient receives the vaccine may be crucial for efficacy. Chronic neoantigen
413 exposure driving T-cells to dysfunctional states, late-differentiated T-cells dominating the tumour
414 microenvironment, and loss of HLA heterozygosity are all reasons NSCLC may become harder
415 to treat with neoantigen vaccines at later stages[6,65–67]. Heterogeneity in NSCLC tumours is likely
416 to influence the efficacy of neoantigen based vaccines[29].

417 Personalised neoantigen vaccines are already being trialled for the treatment of melanoma,
418 glioblastoma and pancreatic cancer[68–70]. These trials rely on the delivery of mRNA containing a
419 number of long sequences predicted to be processed into the final HLA presented neoantigens.
420 The vaccine response rate is in currently the order of 50% of patients when multivalent epitope
421 vaccines in combination with an checkpoint inhibitor adjuvant are used[71]. So whilst these results
422 are extremely promising, there is clearly room for improvement, including in the neoantigen
423 selection process. Immunogenic peptides are predicted by algorithms that incorporate machine
424 learnt parameters such as peptide binding affinity[72] or proteasomal cleavage[73], or more recently

425 using machine learning to identify features such as protein hotspots from large mass
426 spectrometry immunopeptidomics datasets[17,18].

427 The principal difference in our approach is one of tactics rather than strategy. Our tactical
428 difference being to look at which proteins yield peptides presented by HLA molecules and then
429 manually identifying supporting evidence for each neoantigen candidate protein in the literature.
430 To understand the mechanistic basis for this approach, we retrospectively applied Bayesian
431 inference to quantify how donor-specific and protein-specific factors contribute to peptide
432 presentation. This analysis validated our empirical observation that donor-presented source
433 proteins are informative for candidate ranking, while revealing competing cohort-level effects
434 that could inform parameter weighting in future automated selection algorithms. This tactic has
435 some similarity to the 'Tübingen approach' for identification of tumour associated neoantigens
436 which uses mass spectrometry proteomics identifications of HLA peptides to rank candidates[74],
437 as used in the glioblastoma vaccine[69].

438 Related immunopeptidomic-guided approaches have been reported. Tokita et al. used pooled
439 wildtype immunopeptidomes as "surrogate" immunopeptidomes to filter neoantigen candidates
440 against exonic mutations, successfully identifying two neoantigens in colorectal cancer[75]. A
441 large-scale immunopeptide-guided selection study used machine learning to analyse
442 approximately 0.3 million peptides[19]. In concordance with estimates that only 0.5% of mutations
443 yield presented neoantigen peptides[9], they predicted 0.7% of peptides as immunogenic (2,523
444 peptides total: 1,476 canonical and 1,047 noncanonical), with three canonical peptides
445 validated. Our approach differs in using patient-specific tumour immunopeptidomes rather than
446 surrogate immunopeptidomes, systematically testing multiple candidates per patient (70
447 peptides across 6 patients) and using mechanistic modelling to identify relevant parameters for
448 machine learning models.

449 Although still far from perfect with 87% of our predictions failing, our approach outperformed
450 current machine learning models and, in the tested patients, identified peptides that elicited a
451 strong functional response in ~83% of cases tested. Perhaps most importantly, our findings
452 indicate that patient specific knowledge beyond their mutational differences, incorporating
453 knowledge of each individual's HLA-presented peptides are critical parameters in improving
454 neoantigen selection workflows and precision therapies.

455

456 # Materials and Methods

457 **Ethics statement**

458 This study was performed in accordance with the Declaration of Helsinki. Ethical approval was
459 obtained from the local research ethics committee (LREC reference 14-SC-0186 150975) and
460 written informed consent was provided by the patients.

461 **Tissue preparation**

462 Tumours were excised from resected lung tissue post-operatively by pathologists and
463 processed either for histological evaluation of tumour type and stage, or snap frozen at −80°C.
464 Whole blood samples were obtained, and PBMCs were isolated by density gradient
465 centrifugation over Lymphoprep prior to storage at −80°C.

466 **HLA typing**

467 HLA typing was performed by Next Generation Sequencing by the NHS Blood and Transplant
468 Histocompatibility and Immunogenetics Laboratory, Colindale, UK.

469 **DNA and RNA extraction**

470 DNA and RNA were extracted from tumor tissue that had been obtained fresh and immediately
471 snap frozen in liquid nitrogen. Ten to twenty 10 μm cryosections were used for nucleic acid
472 extraction using the automated Maxwell® RSC instrument (Promega) with the appropriate
473 sample kit and according to the manufacturer's instructions: Maxwell RSC Tissue DNA tissue kit
474 and Maxwell RSC simplyRNA tissue kit, respectively. Similarly, DNA was extracted from snap
475 frozen normal adjacent tissue as described above. DNA and RNA were quantified using Qubit
476 fluorometric quantitation assay (ThermoFisher Scientific) according to the manufacturer's

477 instructions. RNA quality was assessed using the Agilent 2100 Bioanalyzer generating an RNA
478 integrity number (RIN; Agilent Technologies UK Ltd.).

479 **Whole exome sequencing**

480 The tumor and normal adjacent samples were prepared using SureSelect Human All Exon V7
481 library (Agilent, Santa Clara USA). 100 bp paired end reads sequencing was performed using
482 the Illumina NovaSeq 6000 system by Edinburgh Genomics (Edinburgh, UK) providing ~100X
483 depth. Reads were aligned to the 1000 genomes project version of the human genome
484 reference sequence (GRCh38/hg38) using the Burrows-Wheeler Aligner (BWA; version 0.7.17)
485 using the default parameters with the addition of using soft clipping for supplementary
486 alignments. Following GATK Best Practices, aligned reads were merged[76], queryname sorted,
487 de-duplicated and position sorted[77] prior to base quality score recalibration[78].

488 **Tumour purity and ploidy**

489 Tumour purity and ploidy were estimated from WES using the ASCAT R package v3.3.0
490 following the recommended steps on the Github repository: https://github.com/VanLoo-
491 lab/ascat[79].

492 **Somatic variant calling**

493 Somatic variant calling was performed using three variant callers: Mutect2 (version 4.1.2.0)[80],
494 Varscan (version 2.4.3)[81], and Strelka (version 2.9.2)[82]. For Mutect2, a panel of normals was
495 created using 40 samples (20 male and 20 female) from the GBR dataset. Variants were
496 combined using gatk GenomeAnalysisTK (version 3.8-1) with a priority order of Mutect2,
497 Varscan, Strelka. Variants were then left aligned and trimmed, and multi-allelic variants split[83].
498 Hard filtering of variants was performed such that only variants that had a variant allele fraction
499 > 5%, a total coverage > 20 and variant allele coverage > 5 were kept. Filtered variants were

500 annotated using VEP (version 97)[84] and with their read counts (https://github.com/genome/bam-
501 readcount) to generate the final filtered and annotated variant call files (VCF).

502 **RNA sequencing**

503 Samples were prepared as TruSeq stranded mRNA libraries (Illumina, San Diego, USA) and
504 100bp paired end sequencing was performed using the Illumina NovaSeq 6000 system by
505 Edinburgh Genomics (Edinburgh, UK). Raw reads were pre-processed to using fastp (version
506 0.20.0)[85]. Filtered reads were aligned to the 1000 genomes project version of the human
507 genome reference sequence (GRCh38/hg38) using hisat2 (version 2.1.0)[86], merged and then
508 transcripts assembled and gene expression estimated with stringtie2 (version 1.3.5)[87] using
509 reference guided assembly.

510 **Mutanome generation**

511 The annotated and filtered VCFs were processed using Variant Effect Predictor (version 97)[84]
512 plugin ProteinSeqs to derive the amino acid sequences arising from missense mutations for
513 each sample for use in immunopeptide analyses.

514 **Neoantigen prediction**

515 Variant call files were prepared for the pvacseq neoantigen prediction pipeline (version
516 1.5.10)[21,33] by adding tumor and normal DNA coverage, and tumor transcript and gene
517 expression estimates using vatools (version 4.1.0) (http://www.vatools.org/). Variant call files of
518 phased proximal variants were also created for use with the pipeline[88]. Prediction of
519 neoantigens arising from somatic variants was then performed using pvacseq with the patient
520 HLA allotypes to predict 8-11mer peptides for class I HLA and 15-mer peptides for class II HLA-
521 DRB allotypes. Four binding algorithms were used for class I predictions (MHCflurry,
522 MHCnuggetsI, NetMHC, PickPocket) and four for class II predictions (MHCnuggetsII,

523 NetMHCIIpan, NNalign, SMMalign). Unfiltered outputs were post-processed in R[89] and split into
524 individual tables for each peptide length and HLA allotype for each patient, and each table was
525 then ranked according to the pvacseq score, where:

526 $$score = binding\ score\ +\ fold\ change\ +\ (variant\ expression\ \times\ fold\ change)$$
527 $$+\ (tumour\ VAF/\ 2)$$

528 Here $binding\ score$ is 1/median neoantigen binding affinity, $fold\ change$ is the difference in
529 median binding affinity between neoantigen and wildtype peptide (agretopicity).

530 Each table was then filtered according to whether wildtype peptide(s) from the same protein as
531 predicted neoantigen was present in the individual's peptidome, and further filtered manually
532 according to biological relevance e.g. the ontology of the protein and its likely presence in the
533 relevant HLA pathway, for example a cytoplasmic resident protein would be considered more
534 likely to yield a HLA-I neoantigen than a HLA-II one. The Human Protein Atlas list of 354 genes
535 identified for unfavourable prognosis in lung cancer, the COSMIC top 20 mutated genes and
536 literature searches were also used as a screen for genes/proteins/peptides of biological
537 relevance.

538 **Immunopeptidomics**

539 Snap frozen tissue samples were briefly thawed and weighed prior to 30s of mechanical
540 homogenization (Fisher, using disposable probes) in 4 mL lysis buffer (0.02M Tris, 0.5% (w/v)
541 IGEPAL, 0.25% (w/v) sodium deoxycholate, 0.15mM NaCl, 1mM EDTA, 0.2mM iodoacetamide
542 supplemented with EDTA-free protease inhibitor mix). Homogenates were clarified for 10 min at
543 2,000g, 4°C and then for a further 60 min at 13,500g, 4°C. 2 mg of anti-MHC-I mouse
544 monoclonal antibodies (W6/32) covalently conjugated to Protein A sepharose (Repligen) using
545 DMP as previously described[90,91] were added to the clarified supernatants and incubated with
546 constant agitation for 2 h at 4°C. The captured MHC-I/β_2m/immunopeptide complex on the

547 beads was washed sequentially with 10 column volumes of low (isotonic, 0.15M NaCl) and high
548 (hypertonic, 0.4M NaCl) TBS washes prior to elution in 10% acetic acid and dried under
549 vacuum. The MHC-I-depleted lysate was then incubated with anti-MHC-II mouse monoclonal
550 antibodies (IVA12) and MHC-II bound peptides were captured and eluted in the same
551 conditions.
552 Immunopeptides were separated from MHC-I/β_2m or MHC-II heavy chain using offline HPLC on
553 a C18 reverse phase column, as previously described[90]. Briefly, dried immunoprecipitates were
554 reconstituted in buffer (1% acetonitrile,0.1% TFA) and applied to a 10cm RP-18e 100-4.6
555 chromolith column (Merck) using an Ultimate 3000 HPLC equipped with UV monitor.
556 Immunopeptides were then eluted using a 15 min 0-40% linear acetonitrile gradient at a flow
557 rate of 1 mL/min. Peptide fractions were eluted and pooled at between 0 and 30% acetonitrile,
558 and the β_2m and MHC heavy chains eluted at >40% acetonitrile.

559 HLA peptides were separated by an Ultimate 3000 RSLC nano system (Thermo Scientific)
560 using a PepMap C18 EASY-Spray LC column, 2 µm particle size, 75 µm x 75 cm column
561 (Thermo Scientific) in buffer A (0.1% Formic acid) and coupled on-line to an Orbitrap Fusion
562 Tribrid Mass Spectrometer (Thermo Fisher Scientific,UK) with a nano-electrospray ion source.
563 Peptides were eluted with a linear gradient of 3%-30% buffer B (Acetonitrile and 0.1% Formic
564 acid) at a flow rate of 300 nL/min over 110 minutes. Full scans were acquired in the Orbitrap
565 analyser using the Top Speed data dependent mode, performing a MS scan every 3 second
566 cycle, followed by higher energy collision-induced dissociation (HCD) MS/MS scans. MS
567 spectra were acquired at resolution of 120,000 at 300 m/z, RF lens 60% and an automatic gain
568 control (AGC) ion target value of 4.0e5 for a maximum of 100 ms. MS/MS resolution was 30,000
569 at 100 m/z. Higher energy collisional dissociation (HCD) fragmentation was induced at an
570 energy setting of 28 for peptides with a charge state of 2–4, while singly charged peptides were
571 fragmented at an energy setting of 32 at lower priority. Fragments were analysed in the Orbitrap

572 at 30,000 resolution. Fragmented m/z values were dynamically excluded for 30 seconds. Data
573 desposited in PRIDE[92], see Data Availability for details.

574 **Proteomic data analysis**

575 Raw spectrum files were analyzed using Peaks Studio 10.0 build 20190129[93,94] and the data
576 processed to generate reduced charge state and deisotoped precursor and associated product
577 ion peak lists which were searched against the UniProt database (20,350 entries, 2020-04-07)
578 plus the corresponding mutanome for each sample (~1,000-5,000 sequences) and
579 contaminants list in unspecific digest mode. Parent mass error tolerance was set a 5ppm and
580 fragment mass error tolerance at 0.03 Da. Variable modifications were set for N-term acetylation
581 (42.01 Da), methionine oxidation (15.99 Da), carboxyamidomethylation (57.02 Da) of cysteine.
582 As previously described, carbamidomethylated cysteines were treated as variable modifications
583 due to the low concentration of 0.2 mM of iodoacetamide used in the lysis buffer to inhibit
584 cysteine proteases[95]. A maximum of three variable modifications per peptide was set. The false
585 discovery rate (FDR) was estimated with decoy-fusion database searches[93] and were filtered to
586 1% FDR. Downstream analysis and data visualizations of the Peaks Studio identifications was
587 performed in R using associated packages[89,96].

588 **Immunopeptide HLA assignment**

589 Identified immunopeptides were assigned to their HLA allotype for each patient using motif
590 deconvolution tools and manual inspection. For class I HLA peptides initial assignment used
591 MixMHCp (version 2.1)[7,31] and for class II HLA peptides initial assignment used MoDec (version
592 1.1)[32]. Downstream analysis and data visualizations was performed in R using associated
593 packages[89,96,97].

Synthetic peptides

Peptides for functional T-cell assays and spectra validation were synthesised using standard solid phase Fmoc chemistry (Peptide Protein Research Ltd, Fareham, UK).

Functional T-cell assay

PBMC ($2x10^6$ per well) were stimulated in 24-well plates with peptide (individual/pool) plus recombinant IL-2 (R&D Systems Europe Ltd.) at a final concentration of 5µg/mL and 20IU/mL, respectively, and incubated at 37°C with 5% CO2; final volume was 2mL. Media containing additional IL-2 (20IU/mL) was refreshed on days 4, 6, 8 and 11 and on day 13 cells were harvested. Expanded cells ($1x10^5$ cell/well) were incubated in triplicate with peptide (individual) at 5µg/mL final concentration for 22 hours at 37°C in 5% CO2; phytohemagglutinin (PHA; Sigma-Aldrich Company Ltd.) and CEFT peptide mix (JPT Peptide Technologies GmbH, Berlin, Germany), a pool of 27 peptides selected from defined HLA Class I- and II-restricted T-cell epitopes, were used as positive controls. Spot forming cells (SFC) were counted using the AID ELISpot plate reader system ELR04 and software (AID Autoimmun Diagnostika GmbH) and positivity calling for ELISpot data used the runDFR(x2) online tool (http://www.scharp.org/zoe/runDFR/). Downstream analysis and data visualizations was performed in R using associated packages[89,96].

Peptide presentation model

Peptide presentation was modelled in R using Bayesian Multilevel Regression Models[36].

For all simulations the Monte Carlo Markov Chain settings were:

```
 # MCMC settings
chains = 4,
cores = 4,
```

```
iter = 4000,
warmup = 1000,
control = list(
    adapt_delta = 0.95,
    max_treedepth = 15)
```

Model quality control was assessed using trace plots to assess convergence and effective parameter sampling (Figure S4-S7), and caterpillar plots to assess the uncertainty of model parameters (Figure S8-S11).

HLA Class I model

A peptide is either presented or not, so peptide presentation is 1 or 0. The model is therefore defined using a Binomial distribution with a logistic link function:

$$\text{peptide_presentation} \sim \text{Binomial}(1, p) \quad \text{logit}(p) = \eta$$

The linear predictor η is defined as:

$$\eta = \beta_0 + \beta_1 \cdot \text{binding_pre_threshold} + \beta_2 \cdot \text{binding_post_threshold} + \beta_3 \cdot \text{length_diff_from_9} + \Sigma_j \beta_{1j} \cdot (\text{binding_pre_threshold} \times I(\text{hla} = j)) + \Sigma_j \beta_{2j} \cdot (\text{binding_post_threshold} \times I(\text{hla} = j)) + \Sigma_j \beta_{3j} \cdot (\text{length_diff_from_9} \times I(\text{hla} = j)) + \beta_7 \cdot \text{z_sqrt_peptides_donor} + \beta_8 \cdot \text{z_sqrt_peptides_cohort} + \text{u_donor} + \text{u_donor:uniprot_id} + \text{u_uniprot_id}$$

Where:

- j is the index for each HLA allotypes ($j = 1, 2, ..., J$)
- $I(\text{hla} = j)$ equals 1 when the observed peptide has HLA allotype j, 0 otherwise
- β_{1j}, β_{2j}, β_{3j} are HLA-specific interaction coefficients

This allows binding and length preferences to vary by HLA allotype.

There are three levels of random effects:

1. Donor-level effects: `u_donor ~ Normal(0, σ²_donor)`
2. Donor-protein interaction effects: `u_donor:protein ~ Normal(0, σ²_donor:protein)`
3. Protein-level effects: `u_protein ~ Normal(0, σ²_protein)`

Partial pooling via these random effects helps to account for donor heterogeneity and any protein-specific effects.

The variables are:

HLA-I binding scores were estimated using MixMHCpred3.0 and the resulting `score_best_allele` scores for each peptide were centred and split around 0 to capture strong and weak binding effects.

- `binding_pre_threshold`: Weak binding peptide score
- `binding_post_threshold`: Strong binding peptide score
- `length_diff_from_9`: Difference in peptide length from 9 amino acids
- `hla`: HLA allotype
- `z_sqrt_peptides_donor`: Standardised square root of number of peptides per protein observed per donor
- `z_sqrt_peptides_cohort`: Standardised square root of number peptides peptides per protein observed per cohort
- `donor`: Donor identifier
- `uniprot_id`: UNIPROT protein identifier

Prior Distributions:

- β_0 ~ `Normal(-1, 1)` Intercept
- β_1 ~ `Normal(1, 0.5)` binding_pre_threshold (main effect)
- β_2 ~ `Normal(1, 0.5)` binding_post_threshold (main effect)
- β_3 ~ `Normal(0, 0.5)` length_diff_from_9 (main effect)
- β_{1j} ~ `Normal(0, 0.5)` binding_pre_threshold:hla interactions
- β_{2j} ~ `Normal(0, 0.5)` binding_post_threshold:hla interactions
- β_{3j} ~ `Normal(0, 0.5)` length_diff_from_9:hla interactions
- β_7 ~ `Normal(0.5, 0.8)` z_sqrt_peptides_donor
- β_8 ~ `Normal(-0.6, 0.8)` z_sqrt_peptides_cohort
- σ_donor ~ `Exponential(2)` Donor random effect SD
- σ_donor:uniprot_id ~ `Exponential(2)` Donor-protein random effect SD
- σ_uniprot_id ~ `Exponential(2)` Protein random effect SD

HLA Class II model

As for HLA-I, peptide is either presented or not, so peptide presentation is 1 or 0. The model is therefore defined using a Binomial distribution with a logistic link function:

`peptide_presentation ~ Binomial(1, p)` $\text{logit}(p) = \eta$

The linear predictor η is defined as:

$$\eta = \beta_0 + \Sigma_j\, s_j(\text{binding_score_centered}) \times I(\text{hla} = j) + \beta_1 \cdot \text{length_diff_from_15} + \Sigma_j\, \beta_{1j} \cdot (\text{length_diff_from_15} \times I(\text{hla} = j)) + \beta_2 \cdot \text{z_sqrt_peptides_donor} + \beta_3 \cdot \text{z_sqrt_peptides_cohort} + \text{u_donor} + \text{u_donor:uniprot_id} + \text{u_uniprot_id}$$

For HLA-II peptides, MixMHC2pred was used and produces scores as a binding rank and these were modelled as smooth splines:

$$\Sigma_j \; s_j(\texttt{binding_score_centered}) \times I(\texttt{hla} = j)$$

Where:

- $s_j(\cdot)$ are HLA allotype specific thin plate regression splines
- j indexes the HLA Class II allotypes ($j = 1, 2, \ldots, J$)
- $I(\texttt{hla} = j)$ equals 1 when the observed peptide has HLA allotype j, 0 otherwise

There are three levels of random effects:

1. Donor-level effects: $\texttt{u_donor} \sim \texttt{Normal}(0, \sigma^2\texttt{_donor})$
2. Donor-protein interaction effects: $\texttt{u_donor:protein} \sim \texttt{Normal}(0, \sigma^2\texttt{_donor:protein})$
3. Protein-level effects: $\texttt{u_protein} \sim \texttt{Normal}(0, \sigma^2\texttt{_protein})$

Partial pooling via these random effects helps to account for donor heterogeneity and any protein-specific effects.

The variables are:

HLA-II binding scores were estimated using MixMHC2pred and the resulting `percent_rank_best` scores for each peptide were centred:

- `binding_score_centered`: Centred binding peptide rank score
- `length_diff_from_15`: Difference in peptide length from 15 amino acids
- `hla`: HLA allotype

- `z_sqrt_peptides_donor`: Standardised square root of number of peptides per protein observed per donor
- `z_sqrt_peptides_cohort`: Standardised square root of number peptides peptides per protein observed per cohort
- `donor`: Donor identifier
- `uniprot_id`: UNIPROT protein identifier

Prior Distributions:

- β_0 ~ `Normal(-2, 1)` Intercept
- β_1 ~ `Normal(-0.25, 0.5)` length_diff_from_15
- β_{1j} ~ `Normal(0, 0.5)` length_diff_from_15:hla interactions
- β_2 ~ `Normal(0.5, 0.7)` z_sqrt_peptides_donor
- β_3 ~ `Normal(-0.5, 0.7)` z_sqrt_peptides_cohort
- s_j(`binding_score_centered`) ~ `Normal(0, 1)` Smooth spline prior (sds)
- σ_donor ~ `Exponential(2)` Donor random effect SD
- σ_donor:`uniprot_id` ~ `Exponential(2)` Donor-protein random effect SD
- σ_`uniprot_id` ~ `Exponential(2)` Protein random effect SD

720 # Author Contributions

721 Conceptualisation by PS, TE, CO, PJ, BN and AB. Project administration by PS, TE, CO, PJ
722 and KM. Supervision by PS, TE and CO. Methodology, investigation, visualisation and formal
723 analysis performed by BN, AB, KM, OW and EC. AB, BN, PJ, TE and CO wrote the manuscript.
724 All authors reviewed the manuscript.

725 # Funding

726 This study was funded by a Cancer Research UK Centres Network Accelerator Award Grant
727 (A21998). Instrumentation in the Centre for Proteomic Research is funded by the Biotechnology
728 and Biological Sciences Research Council, Grant/Award Number: BM/M012387/1. The funders
729 played no role in study design, data collection, analysis and interpretation of data, or the writing
730 of this manuscript.

731 # Competing Interests

732 All authors declare no financial or non-financial competing interests.

733 # Data availability

734 Whole Exome Sequencing and RNAseq data have been deposited at the European Genome-
735 phenome Archive (EGA) under EGA Study ID: EGAS00001005499

736 The mass spectrometry proteomics data have been deposited to the ProteomeXchange
737 Consortium via the PRIDE partner repository with the dataset identifier PXD028990 and
738 10.6019/PXD028990.

739 Supplementary Data are available on Github at https://github.com/ab604/lung-neoantigen-
740 supplement and Zenodo https://zenodo.org/doi/10.5281/zenodo.12820423

741 # References

742 1. Cancer survival in england - office for national statistics.

743 2. Cancer Survival in England, cancers diagnosed 2016 to 2020, followed up to 2021.

744 3. Gainor, J. F. *et al.* EGFR Mutations and ALK Rearrangements Are Associated with Low
745 Response Rates to PD-1 Pathway Blockade in NonSmall Cell Lung Cancer: A Retrospective
746 Analysis. *Clinical Cancer Research* **22**, 4585–4593 (2016).

747 4. Mazieres, J. *et al.* Efficacy of immune-checkpoint inhibitors (ICI) in non-small cell lung
748 cancer (NSCLC) patients harboring activating molecular alterations (ImmunoTarget). *Journal of*
749 *Clinical Oncology* **36**, 9010–9010 (2018).

750 5. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *New*
751 *England Journal of Medicine* **376**, 2109–2121 (2017).

752 6. Neoantigen-directed immune escape in lung cancer evolution | nature.

753 7. Bassani-Sternberg, M. *et al.* Direct identification of clinically relevant neoepitopes
754 presented on native human melanoma tissue by mass spectrometry. *Nature Communications* **7**,
755 13404 (2016).

756 8. Nicholas, B. *et al.* Identification of neoantigens in oesophageal adenocarcinoma.
757 *Immunology* **168**, 420–431 (2023).

758 9. Newey, A. *et al.* Immunopeptidomics of colorectal cancer organoids reveals a sparse
759 HLA class I neoantigen landscape and no increase in neoantigens with interferon or MEK-
760 inhibitor treatment. *Journal for ImmunoTherapy of Cancer* **7**, 309 (2019).

761 10. Marcu, A. *et al.* The HLA Ligand Atlas - A resource of natural HLA ligands presented on
762 benign tissues. *bioRxiv* 778944 (2020) doi:10.1101/778944.

763 11. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks:
764 Application to the MHC class i system. *Bioinformatics* **32**, 511–7 (2016).

765 12. Juncker, A. S. *et al.* Systematic Characterisation of Cellular Localisation and Expression
766 Profiles of Proteins Containing MHC Ligands. *PLoS ONE* **4**, e7448 (2009).

767 13. Müller, M., Gfeller, D., Coukos, G. & Bassani-Sternberg, M. 'Hotspots' of antigen
768 presentation revealed by human leukocyte antigen ligandomics for neoantigen prioritization.
769 *Frontiers in Immunology* **8**, (2017).

770 14. Gfeller, D., Liu, Y. & Racle, J. Contemplating immunopeptidomes to better predict them.
771 *Seminars in Immunology* **66**, 101708 (2023).

772 15. Wells, D. K. *et al.* Key Parameters of Tumor Epitope Immunogenicity Revealed Through
773 a Consortium Approach Improve Neoantigen Prediction. *Cell* **183**, 818–834.e13 (2020).

774 16. Buckley, P. R. *et al.* Evaluating performance of existing computational models in
775 predicting CD8+ t cell pathogenic epitopes and cancer neoantigens. *Briefings in Bioinformatics*
776 **23**, bbac141 (2022).

777 17. Gartner, J. J. *et al.* A machine learning model for ranking candidate HLA class I
778 neoantigens based on known neoepitopes from multiple human tumor types. *Nature Cancer* **2**,
779 563–574 (2021).

780 18. Müller, M. *et al.* Machine learning methods and harmonized datasets improve
781 immunogenic neoantigen prediction. *Immunity* **56**, 2650–2663.e6 (2023).

782 19. Cai, Y. *et al.* Immunopeptidomics-guided discovery and characterization of neoantigens
783 for personalized cancer immunotherapy. *Science Advances* **11**, (2025).

784 20. Huber, F. *et al.* A comprehensive proteogenomic pipeline for neoantigen discovery to
785 advance personalized cancer immunotherapy. *Nature Biotechnology* **43**, 1360–1372 (2024).

786 21. Hundal, J. *et al.* pVACtools: A Computational Toolkit to Identify and Visualize Cancer
787 Neoantigens. *Cancer Immunology Research* **8**, 409–420 (2020).

788 22. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature*
789 **500**, 415–421 (2013).

790 23. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types.
791 *Nature* **502**, 333–339 (2013).

792 24. Gori, K. & Baez-Ortega, A. *Sigfit: Flexible Bayesian Inference of Mutational Signatures*.
793 372896 https://www.biorxiv.org/content/10.1101/372896v2 (2020) doi:10.1101/372896.

794 25. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature*
795 **578**, 94–101 (2020).

796 26. Tate, J. G. *et al.* COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids*
797 *Research* **47**, D941–D947 (2019).

798 27. Racle, J., Jonge, K. de, Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous
799 enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6**,
800 e26476 (2017).

801 28. Nilsson, J. B. *et al.* Accurate prediction of HLA class II antigen presentation across all
802 loci using tailored data acquisition and refined machine learning. *Science Advances* **9**, (2023).

803 29. Jamal-Hanjani, M. *et al.* Tracking the evolution of nonsmall-cell lung cancer. *New*
804 *England Journal of Medicine* **376**, 2109–2121 (2017).

805 30. McGranahan, N. *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer
806 Evolution. *Cell* **171**, 1259–1271.e11 (2017).

807 31. Gfeller, D. *et al.* The Length Distribution and Multiple Specificity of Naturally Presented
808 HLA-I Ligands. *The Journal of Immunology* (2018) doi:10.4049/jimmunol.1800914.

809 32. Racle, J. *et al.* Robust prediction of HLA class II epitopes by deep motif deconvolution of
810 immunopeptidomes. *Nature Biotechnology* **37**, 1283–1286 (2019).

811 33. Hundal, J. *et al.* pVAC-seq: A genome-guided in silico approach to identifying tumor
812 neoantigens. *Genome Medicine* **8**, 11 (2016).

813 34. Gfeller, D. *et al.* The Length Distribution and Multiple Specificity of Naturally Presented
814 HLA-I Ligands. *The Journal of Immunology* (2018) doi:10.4049/jimmunol.1800914.

815 35. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, (2015).

816 36. Bürkner, P.-C. **brms**: An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal*
817 *of Statistical Software* **80**, (2017).

818 37. Howarth, M., Williams, A., Tolstrup, A. B. & Elliott, T. Tapasin enhances MHC class I
819 peptide presentation according to peptide half-life. *Proceedings of the National Academy of*
820 *Sciences* **101**, 11737–11742 (2004).

821 38. Boyle, L. H. *et al.* Tapasin-related protein TAPBPR is an additional component of the
822 MHC class I presentation pathway. *Proceedings of the National Academy of Sciences* **110**,
823 3465–3470 (2013).

824 39. Müller, M., Gfeller, D., Coukos, G. & Bassani-Sternberg, M. 'Hotspots' of antigen
825 presentation revealed by human leukocyte antigen ligandomics for neoantigen prioritization.
826 *Frontiers in Immunology* **8**, (2017).

827 40. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J. & Mann, M. Mass
828 Spectrometry of Human Leukocyte Antigen Class I Peptidomes Reveals Strong Effects of
829 Protein Abundance and Turnover on Antigen Presentation. *Molecular & Cellular Proteomics* **14**,
830 658–673 (2015).

831 41. Gfeller, D. *et al.* The Length Distribution and Multiple Specificity of Naturally Presented
832 HLA-I Ligands. *The Journal of Immunology* (2018) doi:10.4049/jimmunol.1800914.

833 42. Rosenberg, S. A. *et al.* Immunologic and therapeutic evaluation of a synthetic peptide
834 vaccine for the treatment of patients with metastatic melanoma. *Nature Medicine* **4**, 321–327
835 (1998).

836 43. Weinschenk, T. *et al.* Integrated functional genomics approach for the design of patient-
837 individual antitumor Vaccines1. *Cancer Research* **62**, 5818–5827 (2002).

838 44. Hsiue, E. H.-C. *et al.* Targeting a neoantigen derived from a common *TP53* mutation.
839 *Science* **371**, (2021).

840 45. Castle, J. C. *et al.* Exploiting the mutanome for tumor vaccination. *Cancer Research* **72**,
841 1081–1091 (2012).

842 46. Levine, A. J., Jenkins, N. A. & Copeland, N. G. The Roles of Initiating Truncal Mutations
843 in Human Cancers: The Order of Mutations and Tumor Cell Type Matters. *Cancer Cell* **35**, 10–
844 15 (2019).

845 47. Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. C. p53 Mutations in Human
846 Cancers. *Science* **253**, 49–53 (1991).

847 48. Lin, M. J. *et al.* Cancer vaccines: the next immunotherapy frontier. *Nature Cancer* **3**,
848 911–926 (2022).

849 49. Vadakekolathu, J. *et al.* Multi-Omic Analysis of Two Common P53 Mutations: Proteins
850 Regulated by Mutated P53 as Potential Targets for Immunotherapy. *Cancers* **14**, 3975 (2022).

851 50. Pearson, H. *et al.* MHC class Iassociated peptides derive from selective regions of the
852 human genome. *Journal of Clinical Investigation* **126**, 4690–4701 (2016).

853 51. Restrepo-Pérez, L., Joo, C. & Dekker, C. Paving the way to single-molecule protein
854 sequencing. *Nature Nanotechnology* **13**, 786–796 (2018).

855 52. Alfaro, J. A. *et al.* The emerging landscape of single-molecule protein sequencing
856 technologies. *Nature Methods* **18**, 604–617 (2021).

857 53. Lucas, F. L. R., Versloot, R. C. A., Yakovlieva, L., Walvoort, M. T. C. & Maglia, G.
858 Protein identification by nanopore peptide profiling. *Nature Communications* **12**, 5795 (2021).

859 54. Motone, K. & Nivala, J. Not if but when nanopore protein sequencing meets single-cell
860 proteomics. *Nature Methods* **20**, 336–338 (2023).

861 55. Martin-Baniandres, P. *et al.* Enzyme-less nanopore detection of post-translational
862 modifications within long polypeptides. *Nature Nanotechnology* **18**, 1335–1340 (2023).

863 56. Motone, K. *et al.* Multi-pass, single-molecule nanopore reading of long protein strands
864 with single-amino acid sensitivity. (2023).

865 57. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-
866 specific antigens. *Science Translational Medicine* **10**, (2018).

867 58. Chong, C. *et al.* Integrated proteogenomic deep sequencing and analytics accurately
868 identify non-canonical peptides in tumor immunopeptidomes. *Nature Communications* **11**,
869 (2020).

870 59. Ruiz Cuevas, M. V. *et al.* Most non-canonical proteins uniquely populate the proteome or
871 immunopeptidome. *Cell Reports* **34**, 108815 (2021).

872 60. Jeck, W. R. *et al.* Circular RNAs are abundant, conserved, and associated with ALU
873 repeats. *RNA* **19**, 141–157 (2012).

874 61. Horokhovskyi, Y. *et al.* An Automated Workflow to Address Proteome Complexity and
875 the Large Search Space Problem in Proteomics and HLA-I Immunopeptidomics. *Molecular &*
876 *Cellular Proteomics* **24**, 101039 (2025).

877 62. Ferreira, H. J. *et al.* Immunopeptidomics-based identification of naturally presented non-
878 canonical circRNA-derived peptides. *Nature Communications* **15**, (2024).

879 63. Ouspenskaia, T. *et al.* Unannotated proteins expand the MHC-I-restricted
880 immunopeptidome in cancer. *Nature Biotechnology* **40**, 209–217 (2021).

881 64. Ely, Z. A. *et al.* Pancreatic cancerrestricted cryptic antigens are targets for T cell
882 recognition. *Science* **388**, (2025).

883 65. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic
884 properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015).

885 66. Joshi, K. *et al.* Spatial heterogeneity of the T cell receptor repertoire reflects the
886 mutational landscape in lung cancer. *Nature Medicine* **25**, 1549–1559 (2019).

887 67. Ghorani, E. *et al.* The T cell differentiation landscape is shaped by tumour mutations in
888 lung cancer. *Nature Cancer* **1**, 546–561 (2020).

889 68. Sahin, U. *et al.* Personalized RNA mutanome vaccines mobilize poly-specific therapeutic
890 immunity against cancer. *Nature* **547**, 222–226 (2017).

891 69. Hilf, N. *et al.* Actively personalized vaccination trial for newly diagnosed glioblastoma.
892 *Nature* **565**, 240–245 (2019).

893 70. Rojas, L. A. *et al.* Personalized RNA neoantigen vaccines stimulate T cells in pancreatic
894 cancer. *Nature* **618**, 144–150 (2023).

895 71. Sethna, Z. *et al.* RNA neoantigen vaccines prime long-lived CD8+ T cells in pancreatic
896 cancer. *Nature* **639**, 1042–1051 (2025).

897 72. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks:
898 Application to the MHC class i system. *Bioinformatics* **32**, 511–7 (2016).

899 73. Nielsen, M., Lundegaard, C., Lund, O. & Keşmir, C. The role of the proteasome in
900 generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal
901 cleavage. *Immunogenetics* **57**, 33–41 (2005).

902 74. Singh-Jasuja, H., Emmerich, N. P. N. & Rammensee, H.-G. The Tübingen approach:
903 identification, selection, and validation of tumor-associated HLA peptides for cancer therapy.
904 *Cancer Immunology, Immunotherapy* **53**, 187–195 (2004).

905 75. Tokita, S. *et al.* Identification of immunogenic HLA class I and II neoantigens using
906 surrogate immunopeptidomes. *Science Advances* **10**, (2024).

907 76. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, (2021).

908 77. *Picard Toolkit*. (Broad Institute, 2019).

909 78. Genomics in the Cloud [Book].

910 79. Ross, E. M., Haase, K., Van Loo, P. & Markowetz, F. Allele-specific multi-sample copy
911 number segmentation in ASCAT. *Bioinformatics* **37**, 1909–1911 (2021).

912 80. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* 861054
913 (2019) doi:10.1101/861054.

914 81. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery
915 in cancer by exome sequencing. *Genome Research* **22**, 568–576 (2012).

916 82. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nature*
917 *Methods* **15**, 591–594 (2018).

918 83. Bonfield, J. K. *et al.* HTSlib: C library for reading/writing high-throughput sequencing
919 data. *GigaScience* **10**, (2021).

920 84. McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biology* **17**, 122
921 (2016).

922 85. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ
923 preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

924 86. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome
925 alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**, 907–
926 915 (2019).

927 87. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with
928 StringTie2. *Genome Biology* **20**, 278 (2019).

929 88. Hundal, J. *et al.* Accounting for proximal variants improves neoantigen prediction. *Nature*
930 *Genetics* **51**, 175–179 (2019).

931 89. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation
932 for Statistical Computing, Vienna, Austria, 2018).

933 90. Purcell, A. W., Ramarathinam, S. H. & Ternette, N. Mass spectrometrybased
934 identification of MHC-bound peptides for immunopeptidomics. *Nature Protocols* **14**, 1687–1707
935 (2019).

936 91. Bailey, A. *et al.* Characterization of the class i MHC peptidome resulting from DNCB
937 exposure of HaCaT cells. *Toxicological Sciences* (2020) doi:10.1093/toxsci/kfaa184.

938 92. Perez-Riverol, Y. *et al.* The PRIDE database resources in 2022: a hub for mass
939 spectrometry-based proteomics evidences. *Nucleic Acids Research* **50**, D543–D552 (2021).

940 93. Zhang, J. *et al.* PEAKS DB: De novo sequencing assisted database search for sensitive
941 and accurate peptide identification. *Molecular & Cellular Proteomics* **11**, M111010587 (2012).

942 94. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep
943 learning. *Proceedings of the National Academy of Sciences of the United States of America*
944 (2017) doi:10.1073/pnas.1705691114.

945 95. Chong, C. *et al.* High-throughput and Sensitive Immunopeptidomics Platform Reveals
946 Profound Interferonγ-Mediated Remodeling of the Human Leukocyte Antigen (HLA)
947 Ligandome*. *Molecular & Cellular Proteomics* **17**, 533–548 (2018).

96. Wickham, H. *et al.* Welcome to the tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

97. Jessen, L. E. *PepTools - an r-Package for Making Immunoinformatics Accessible*. (2018).