bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Proteogenomics 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, UK

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

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

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

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

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

ORCIDs:

Ben Nicholas: 0000-0003-1467-9643

Alistair Bailey: 0000-0003-0023-8679

Katy J McCann: 0000-0003-4388-7265

Eve Currall: 0009-0008-0400-2376

Peter Johnson: 0000-0003-2306-4974

Tim Elliott: 0000-0003-1097-0222

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Christian Ottensmeier: 0000-0003-3619-1657

Paul Skipp: 0000-0002-2995-2959

Correspondence to:

Dr Ben Nicholas

Centre for Proteomic Research
B85, Life Sciences Building
University of Southampton
University Road
Highfield
Southampton, Hants.
SO17 1BJ

Tel No: +44(0)2380 59 5503

email: bln1@soton.ac.uk

Running title:

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

Keywords:

HLA, peptidome, non-small cell lung cancer, antigen presentation

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Non-small cell lung cancer (NSCLC) has poor survival in both the short and long term even for those receiving modern checkpoint inhibitor therapies.

One attractive strategy for NSCLC therapy is personalised vaccines based upon short peptide neoantigens containing tumour mutations, presented to cytotoxic T-cells by human leukocyte antigen (HLA) molecules. However, identification of therapeutically relevant neoantigens is challenging. Existing methodologies yield positive functional assay responses in around 6% of candidate neoantigens tested, and neoantigen based vaccines in melanoma, glioblastoma and pancreatic cancer yield an immune response in around 50% of patients.

Here we report a proteogenomics approach to identify neoantigens in tumours from a cohort of 24 NSCLC patients: 15 adenocarcinoma, 9 squamous cell carcinoma. We characterised the mutational and HLA immunopeptide landscapes of NSCLC using whole exome sequencing, transcriptomics and mass spectrometry immunopeptidomics. We directly identified one neoantigen, and additional predicted neoantigens were generated using an existing in silico neoantigen prediction workflow. Using the immunopeptidomes to filter for candidate predicted neoantigens we identified positive functional assay responses for 5 out of the 6 patients we tested, with an overall success rate of 13%, inclusive of the directly observed neoantigen. Finally, for one patient using scRNAseq we identified a CD8+ effector T-cell clonotype expanded only in response to the putative class I HLA neoantigen.

These results represent an improvement in both the quantity of neoantigens identified and the specificity of immune responses to neoantigens, utilising knowledge of the HLA peptides presented on a tumour. Thus immunopeptidomics has the potential to improve the efficacy of neoantigen based personalised cancer vaccine workflows.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Lung cancer is the second most common cancer in the UK and is frequently diagnosed at an advanced stage, either locally advanced (stage III) or metastatic (stage IV). Non-small cell lung cancer (NSCLC) accounts for 85-90% of these cases and can be further classified into three histological subtypes: adenocarcinoma (LUAD), squamous cell carcinoma (LUSC), and large cell undifferentiated carcinoma. Of these types LUAD is the most common, often forming in the alveoli in the outer peripheral lung, whereas LUSC tends to form in squamous cells located more centrally and is the next most common type, whereas large cell undifferentiated carcinoma is least common, but can form anywhere in the lung [1]. In the UK, less than 20% of all lung cancer patients survive for 5 years, with the majority of patients surviving less than one year post-diagnosis [3]. Current treatments aim to prolong survival and improve quality of life, with options including surgery, chemotherapy, radiotherapy, and immunotherapy, subject to favourable biomarker profiles.

Currently, four immunotherapies targeting PD-1 or PD-L1 are licensed for use in NSCLC. However, these treatments are less effective for patients with well-defined mutations in Epidermal growth factor receptor (EGFR) and Anaplastic lymphoma kinase (ALK) [4,5]. As a result, immunotherapy is typically offered as a second-line treatment after chemotherapy and/or targeted therapy against EGFR, ALK, or ROS oncogene mutations. While there are marginal differences in chemotherapeutic options between LUAD and LUSC, LUAD generally has more favourable survival odds. The longitudinal NSCLC TRACERx (TRAcking Cancer Evolution through therapy (Rx)) study has identified evolutionary processes that help explain treatment resistance. Whole genome doubling is common in NSCLC due to tobacco smoke and cytidine deaminase activity, serving to protect the tumour against the effects of high numbers of mutations and chromosomal instability [6]. Smoking mutations are truncal, whereas branch

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mutations tend to be caused by cytidine deaminases. These two major categories of mutations lead to extensive intratumour heterogeneity in NSCLC. The degree of heterogeneity was found to be prognostic for disease recurrence or death, but confound the utility of biomarkers used to predict immunotheraputic responses. [7]. In heterogeneous tumours the expression or secretion levels of putative biomarkers may be unrepresentative of the whole, thus prognostic tests may not be sensitive enough to detect them. The mutational evolution of NSCLC tumours is mirrored by a parallel evolution of T-cell receptors and tumour infiltration by T-cells. Tumour mutations shape the T-cell repertoire via their effects on human leukocyte antigen (HLA) heterozygosity, antigen processing machinery and the neoantigen peptides generated from the cancer genome, the mutanome, which are presented at the cell surface by HLA molecules. Tumour mutations can have opposing effects on immune function, depending on when and how T-cells encounter the neoantigens. Recognition by early-differentiated T-cells may lead to effective tumour control. However, chronic exposure to these neoantigens can drive T-cells into dysfunctional states. Likewise, as mutations accumulate, late-differentiated T-cells may out-compete early-differentiated T-cells and dominate the tumour microenvironment [10]. These findings have important implications for the development of more effective, personalized treatment strategies that can overcome these evolutionary consequences.

An attractive strategy for NSCLC treatment is vaccination targeting on HLA presented neoantigens. This approach assumes neoantigens can be identified that expand tumour killing T-cell populations and/or modulate the tumour microenvironment to make T-cell infiltration or checkpoint inhibitors more effective. The personalised nature of neoantigens minimise the risk of off-target effects and autoimmunity. However, direct identification of neoantigens is rare [12] and most approaches to neoantigen discovery rely on predicting that a given mutation leads to protein synthesis, antigen processing and HLA presentation. Direct observation is rare in part due to limits in the sensitivity of the mass spectrometry proteomic detection of HLA ligands,

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

known as immunopeptidomics. Moreover, it is estimated that only a small fraction of mutations are actually presented, possibly as low as 0.5% of non-silent mutations [13]. For example, a NSCLC tumour with 600 missense variants might yield only 3 presented neoantigens amongst a lung tissue immunopeptidome of around 60,000 unique class I and II HLA peptides [14]. A typical experiment may identify 3,000 of these peptides. Assuming a hypergeometric distribution, the probability of observing one class I or II HLA neoantigen is about 14%. Or to put it another way, there is around an 86% chance of not seeing any neoantigens in any single mass spectrometry proteomics experiment.

Given these odds, much effort has been put into the in silico prediction of mutations that will give rise to neoantigens that would make effective vaccines. There are well established algorithms that can predict the likelihood of a peptide of given amino acid sequence binding to an HLA molecule, and immunopeptidomic evidence of the peptide length preference of peptide for different HLA allotypes [15], and preferential regions of proteins favourable for presentation [18]. However, even with this knowledge prediction is stymied by the number of potential neoantigen candidates each mutation might yield, creating large lists of candidate peptides. Moreover, the key biochemical and structural parameters of immunogenic neoantigens remain unknown. The best neoantigen prediction models have a success rate such that around 6% of their putative neoantigens are T-cell reactive [20], although recent machine learning models claim to have increased this predictive power [22].

Here we adopted an alternative approach where, rather than trying to predict whether neoantigens would be presented on the basis of various characteristics alone, we would instead use immunopeptidomic data as evidence that the source protein of predicted neoantigens could be processed and presented on HLA-I and -II. Thus, immunopeptidomics was used as circumstantial evidence of the biological availability of a mutated protein for presentation by HLA. First we mapped the mutational and immunopeptidome landscapes of a cohort of LUAD

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and LUSC patients. We then predicted HLA-restricted neoantigens using existing algorithms and used immunopeptidomic data from their individual tumours to filter those predictions on the basis of evidence that they could be presented. We identified neoantigens in five out of the six patients we tested our predictions by functional assay. Our overall success rate was 13% of predicted neoantigens yielded positive functional assay tests. For one LUAD donor we were able to use scRNAseq to further explore the specificity of our neoantigens and identify cognate CD8+ and CD4+ T-cell receptors.

These proof-of-concept results demonstrate how the information contained within the immunopeptidome has the potential to enhance proteogenomics strategies for identifying neoantigens for every patient, and thus truly personalised vaccination strategies for NSCLC.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

A proteogenomics workflow for neoantigen identification

Our NSCLC cohort consisted of 24 patients, 15 LUAD (8 female, 7 male) and 9 LUSC (5 female, 4 male). Median age at diagnosis was 69 (See Table 1 and Supplementary Table S1). Tumour tissue and PBMCs were used for HLA typing, whole exome sequencing, RNA sequencing and mass spectrometry proteomics of the HLA immunopeptidome (Figure 1). To identify candidate neoantigens for each patient we developed a workflow that surveyed both the genomic and immunopeptidomic landscapes. Somatic missense variants called from the whole exome sequencing (WES) were used to generate a mutanome for each individual against which the HLA immunopeptidome could be searched for direct observation of neoantigens. Variants, gene expression and the patient HLA allotypes were also used for the prediction of putative neoantigens using existing tools [23].

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

Donor[1]	Age at diagnosis	Sex	Smoking status	Cancer subtype
A113	67	Male	Current smoker	Squamous
A114	77	Female	Ex smoker	Adenocarcinoma
A115	59	Male	Ex smoker	Squamous
A116	62	Female	Never smoker	Squamous
A117	83	Male	Current smoker	Adenocarcinoma
A118	59	Female	Ex smoker	Adenocarcinoma
A119	71	Male	Ex smoker	Adenocarcinoma
A120	73	Male	Ex smoker	Adenocarcinoma
A133	82	Female	Ex smoker	Squamous
A134	61	Female	Current smoker	Squamous
A136	72	Male	Never smoker	Adenocarcinoma
A137	72	Female	Ex smoker	Adenocarcinoma

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A139	76	Female	Ex smoker	Adenocarcinoma
A140	66	Female	Ex smoker	Squamous
A141	77	Male	Ex smoker	Adenocarcinoma
A142	78	Male	Current smoker	Adenocarcinoma
A143	55	Female	Ex smoker	Adenocarcinoma
A144	72	Male	Ex smoker	Squamous
A145	69	Male	Ex smoker	Squamous
A146	67	Female	Current smoker	Adenocarcinoma
A147	55	Female	Current smoker	Adenocarcinoma
A148	69	Male	Ex smoker	Adenocarcinoma
A152	69	Female	Ex smoker	Squamous
A153	66	Female	Ex smoker	Adenocarcinoma

[1] Summary details to be decided

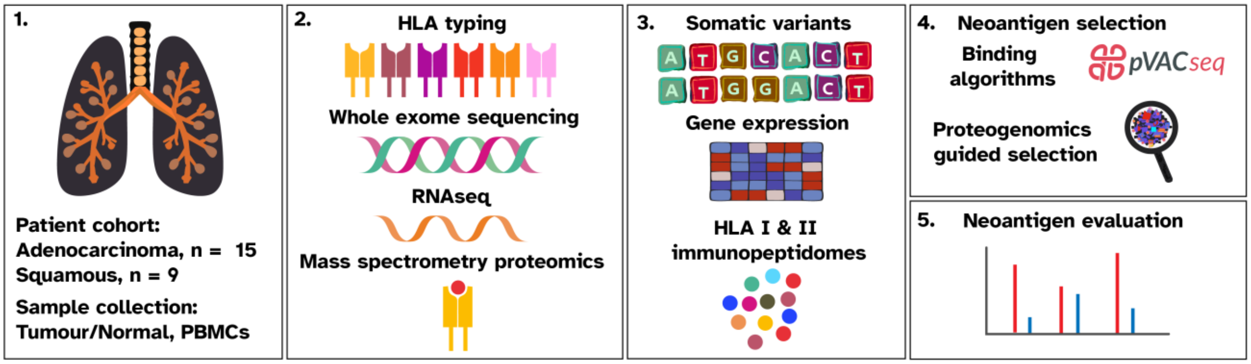

Figure 1: Integrated proteogenomics workflow. HLA typing, whole exome sequencing, RNASeq and mass spectrometry-based proteomic of the HLA immunopeptidome were collected for 24 lung cancer patients providing mutational, gene expression and immunopeptidomic data from which to identify candidate neoantigens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

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The mutational landscape of NSCLC in the studied cohort

To assess the likelihood of identifying HLA presented neoantigens we first examined the mutational landscape of the NSCLC cohort and found it to be consistent with previous reports [24,25]. Somatic variants were identified by WES of tumour and matched normal adjacent tissues. Tumour mutational burden (TMB) quantifies the number of mutations per million bases (Mb). From WES it is calculated as the number of variants divided by the size of the exome targets; here the target size was 35.7 Mb and the number of variants were either the total number of all variants, or only the protein coding missense variants: $N_{vars}/(35.7) = N_{vars}/Mb$.

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

Approximately one third of all nucleotide transitions and transversions were C>A transversions in both LUSC and LUAD (Figure 2 B), a known mutational signature of smoking [24]. Of the 51,810 LUAD and 32,344 LUSC single nucleotide variants, approximately 20% were missense variants (10,565 LUAD, 6,772). These missense SNVs along with approximately 15% insertion/deletion variants (9,697 LUAD, 6,780 LUSC) predict amino acid changes at the protein level and are therefore potential sources of HLA neoantigens (Figure 2 C).

For each cancer subtype, patterns of single base substitutions created by the somatic mutations were extracted to identify mutational signatures that were fitted to those identified in COSMIC [26–28] (Figure 2 D-F). LUAD signature A fit SBS36 indicating base excision repair deficiency characterised by C>A transversions. LUAD signature C fit SBS2, which is common in lung cancer and thought to indicate APOBEC cytidine deaminase activity as characterised by C>T

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

In addition to examining the potential for neoantigen generation at the exon level, we sought to examine the potential for neoantigen recognition within the tumours using bulk gene expression data from RNAseq to assess the fractions of immune cells present in the tumours[29] (Figure 2 G). This estimation also provides an indication of the tumour sample purity. All expressed genes not used as markers for immune cells are labelled as ‘otherCells’ and we would expect this catergory to comprise the largest proportion of cells in a tumour sample. Therefore if a sample has a low proportions of ‘otherCells’ it is indicative of a less pure tumour sample. For LUAD and LUSC, the median proportions of ‘otherCells’ are one third. In cases with very low proportions of ‘otherCells’ such as A134 and A145, the corresponding histology reports indicate these were fibrotic samples consistent with the very high proportions of cancer associated fibroblasts identified by RNAseq. However at the cohort level, proportions of T-cells estimated capable of responding to neoantigens presented by HLA were estimated with similar medians for CD4+ T-cells of 18% and 15% for LUAD and LUSC respectively, and medians for CD8+ T-cells of 2% and less than 1% for LUAD and LUSC respectively.

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

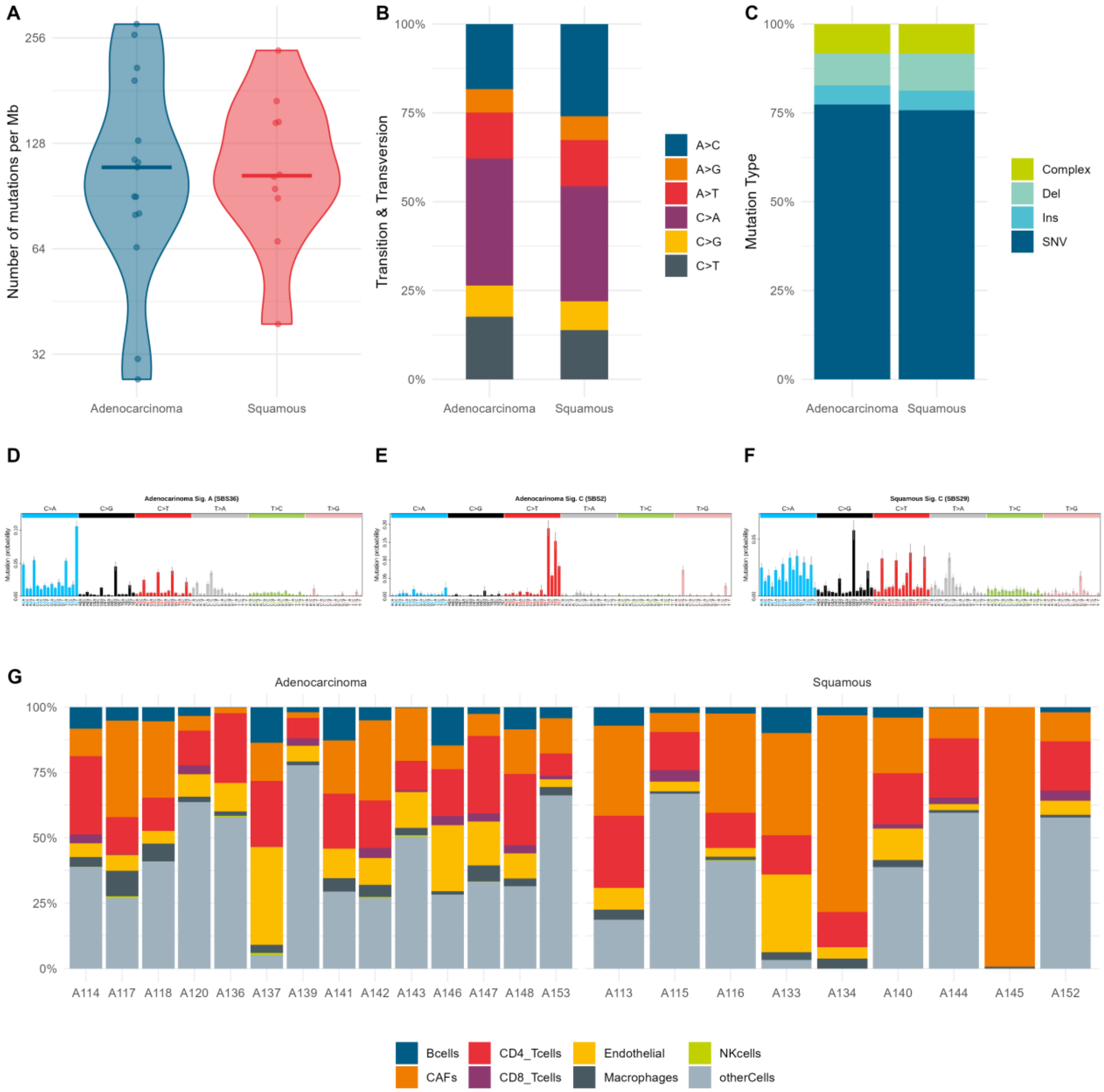

Figure 2: The mutational landscape of lung cancer in the studied cohort. (A) The mutational burden of each cancer type: squamous (n=9) and adenocarcinoma (n=15). (B) Mutation frequency of six transition and transversion categories for each cancer type. (C) Mutation frequencies each cancer type. (D-F)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

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The peptidome landscape of NSCLC in the studied cohort

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

The lung cancer peptidome resembles the healthy lung tissue peptidome

We compared the distinct source protein populations yielding the class I and II HLA peptidomes between our LUAD, LUSC samples and healthy lung tissues from the Human HLA Ligand Atlas [14] to examine their similarities and differences (Figure 3 B-C) , considering only proteins present in at least two-thirds of our samples peptidomes. Our analysis suggests that healthy lung and tumour tissues immunopeptidomes sample largely the same protein populations. 92% of HLA-I proteins and 52% of HLA-II proteins were common to all three tissue types. The remaining proteins most likely represent experimental variation.

Across the cohort of 24 patients we identified a single missense variant product by direct mass spectrometric observation in the class I HLA immunopeptidome of one LUAD patient (A147) (Figure S1). This derived from a C>A variant in the ALYREF gene yielding an Asp10Tyr mutation in its protein product THO complex subunit 4 (Uniprot: Q86V81). This mutation yielded seven nested 15-18mer peptides with the mutation Y before the start of core sequence of MSLDDIIKL. No wild type peptides were observed for this protein in either the HLA I or II immunopeptidomes, suggesting this mutation altered either the binding affinity of these peptides or the source protein processing in the antigen processing pathway. The rarity of this observation is in keeping with estimates of frequencies in the order of 0.5% of missense variants encoding presented neoantigens [11,13]. The length of the ALYREF peptides

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

suggested these may be class II HLA peptides that we had captured by chance in this assay. The motif most closely matched the patients HLA-DRB1*03:01 allotype with peptide AYKMDMSLDDIIKLN predicted as a weakly binding peptide [30]. We identified 1135 missense mutations for patient A147 (Table S1) potentially yielding 6 neoantigens, representing 0.5% missense derived neoantigens, of which we observed one.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

234

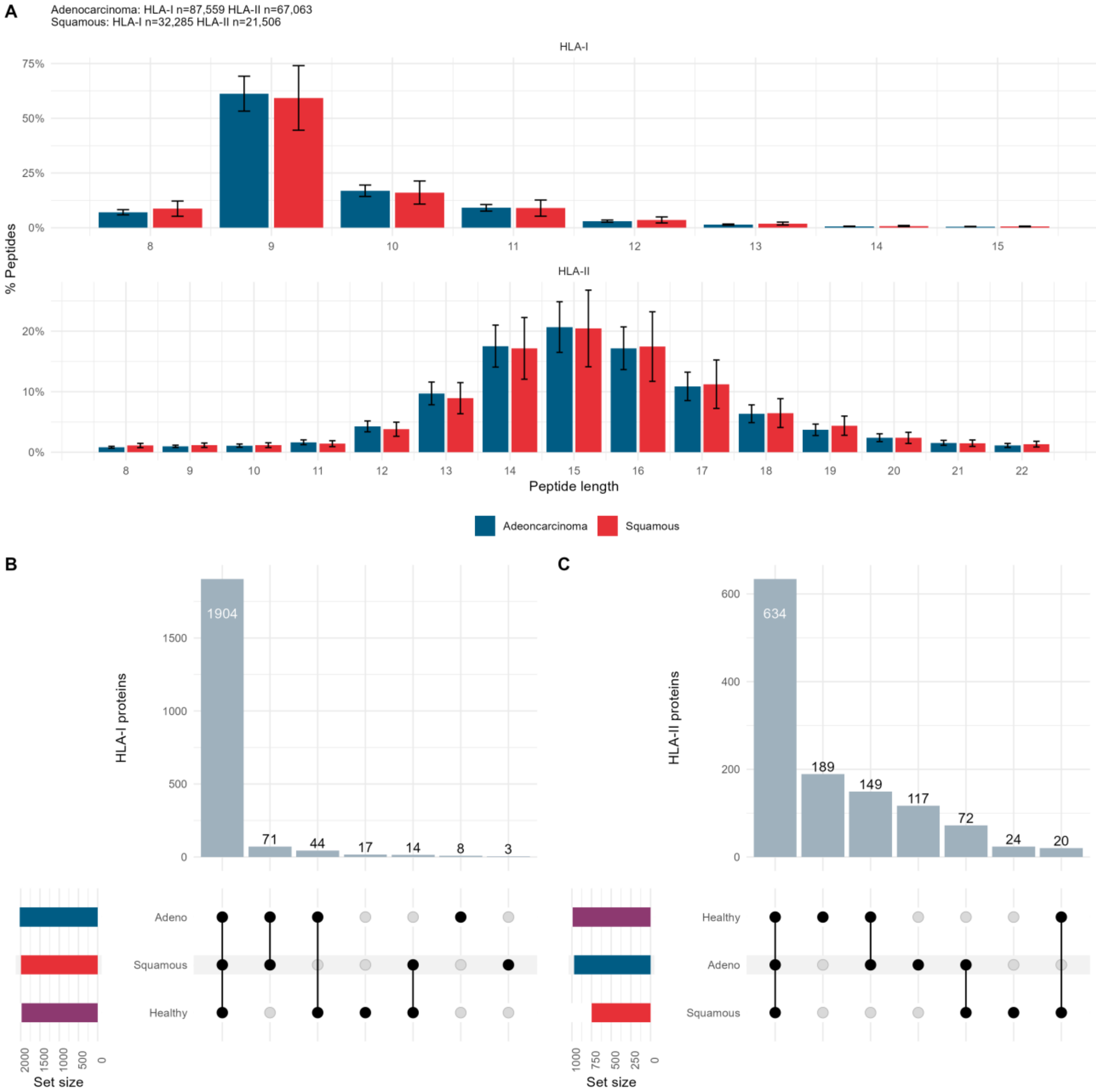

Figure 3: The peptidome landscape of lung cancer. (A) Length distributions of immunopeptides from tumour tissues. (B-C) Upset plots of proteins presented by HLA molecules (class I left, class II right) comparing proteins between cancer subtypes and healthy lung tissues from the HLA ligand atlas.

235

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A proteogenomics view of NSCLC in the studied cohort

Consistent with the view that non-silent mutations rarely encode HLA presented neoantigens [13], we observed that LUAD and LUSC driver genes [31] are not mutated and presented by HLA molecules with the same frequencies. Some drivers are frequently mutated, but rarely presented e.g. APC, whereas other are rarely mutated, but frequently present in the HLA immunopeptidomes e.g. KEAP1 (Figure 4 A). TP53 is both frequently mutated and presented in the class I HLA peptidomes of both NSCLC subtypes (Figure 4 A).

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

We also found that loss of class I HLA heterozygosity in the genome [32] is reflected in the peptidome. In heterozygous patients, immunopeptides identified as presented by HLA molecules from the retained allele were observed at higher proportions in the peptidome than from the lost allele for HLA-A and B allotypes (Figure 4 C).

These observations imply firstly that the likelihood of a putative neoantigen being presented by either HLA class is influenced by the cellular compartment origin of the source protein and secondly, putative neoantigens with motifs [33] for the retained class I HLA allotypes are more likely to be presented than those from the lost allotype.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

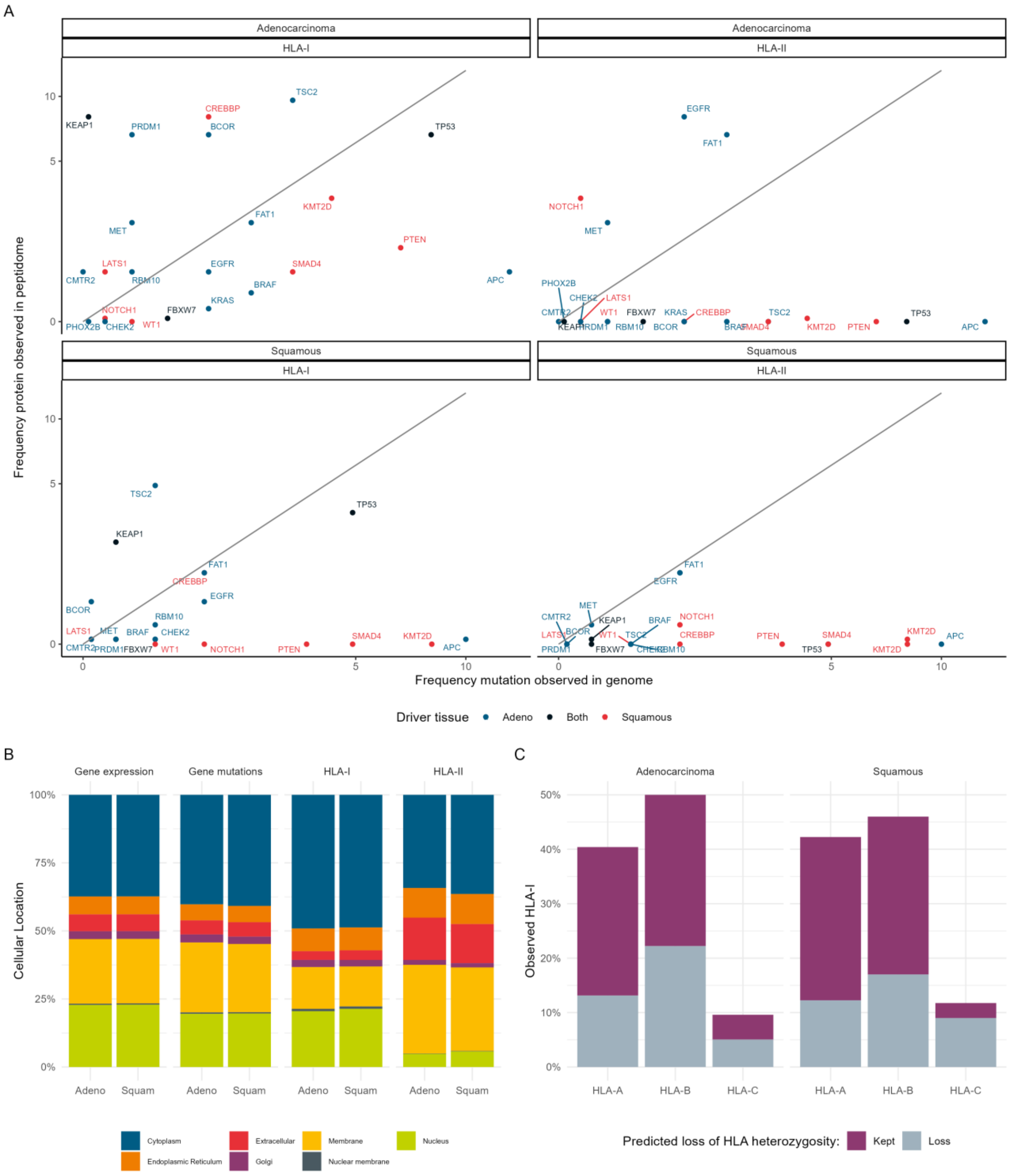

Figure 4: Integrating the mutational and immunopeptidome landscape reveals previously unclear relationships between mutations and peptide presentation. (A) The frequency of mutated initiating driver genes for each cancer type plotted against the frequency of

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

observed presentation from the corresponding protein in either the HLA-I (left) or HLA-II peptidome (right). The colour indicates which cancer type the gene was identified as an initiating driver in [6]. (B) Comparison of the frequency between the cellular compartments in which genes expression and somatic mutations occur, and those from which HLA peptides are observed in each cancer type. (C) The relative proportions of immunopeptides assigned to each of the two HLA-A, B and C allotypes for heterozygous patients. [11,33,34]. The colour represents whether an allele is predicted to have loss of heterozygosity in the genome [32].

Proteogenomics guided NSCLC neoantigen selection and testing

In selecting neoantigens we initially used the pVACseq tool to create a list of putative neoantigens for each patient and HLA allotype and different peptide lengths [35]. Briefly, we used pVACseq with the whole exome and transcriptome outputs and patient HLA allotypes to predict 8-11mer peptides for class I HLA and 15-mer peptides for class II HLA-DRB allotypes across eight binding algorithms. This combined genomic and binding score creates an overall score for each peptide (Details in Section 1.8.0.9). For our 24 patients this comprises 524 HLA class I tables and 74 HLA class II tables of ranked predictions. Discarding any prediction for a peptide with >500 nM binding affinity, pVACseq yielded 27,466 class I HLA and 127,015 class II HLA predicted neoantigen peptides (Supplementary Tables S2 and S3)

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to HLA peptide length preferences [36]. This reduced the number of candidates to a few hundred peptides for each patient. We finally manually curated the ranked peptide candidates for biological relevance using auxiliary information from the literature, the Human Protein Atlas and COSMIC. (Figure 1, Section 1.8.0.9).

Our exploratory filtering process for candidate neoantigens can be summarised as: Does a missense mutation exist? Is there evidence that the mutated gene product enters the antigen processing pathway for presentation, and if so in which HLA pathway? Is the candidate neoantigen of the preferred HLA allotype length? Is the candidate neoantigen predicted to bind to the HLA allotype according to pVACseq? Is there any additional information available publicly to preferentially support one neoantigen candidate over another?

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

For six patients, 3 LUAD and 3 LUSC, we selected 9 to 14 putative neoantigens per patient (70 in total) and synthesised the specific putative HLA-I or HLA-II peptides in the mutant neoantigen and wildtype forms (Supplementary Table S4). We identified nine strong neoantigen specific responses to putative neoantigens in five out of six patients, including for the directly observed ALYREF peptide (Figure 5 A-F, Table 2). This represents a 13% response rate, twice the genomics-based peptide prediction rate of 6% reported in the literature [19]. We observed responses to both class I and class II HLA candidate neoantigens in LUAD (Figure 5 A-C), but only class II HLA candidate neoantigens in LUSC (Figure 5 E-F). LUSC patient A116 yielded no responses (Figure 5 D).

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

294

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.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

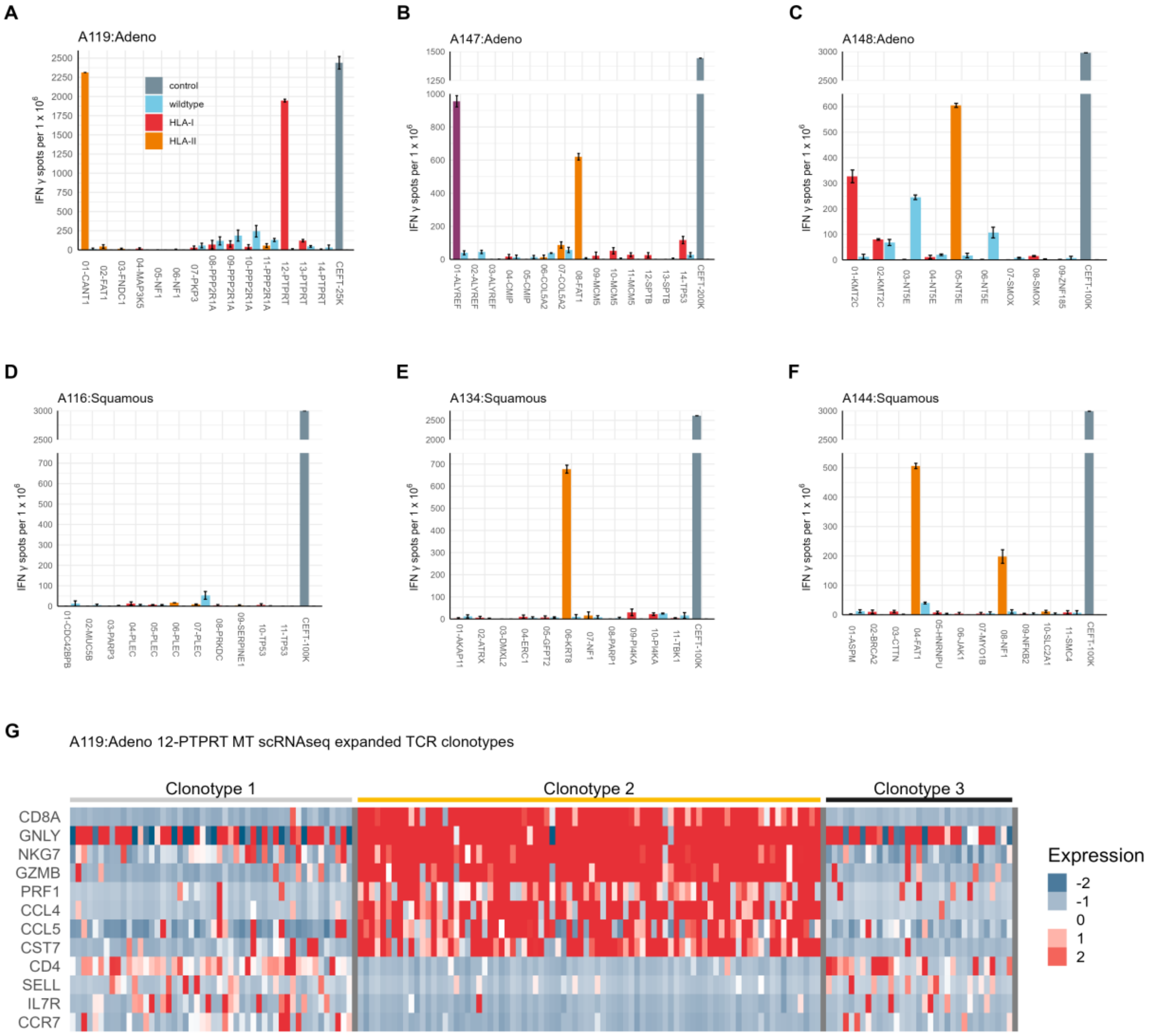

Figure 5: Proteogenomics guided NSCLC neoantigen selection identifies nine strong candidates (A-F) IFN-γ ELISPOT of putative neoantigens three LUAD and three LUSC patients. Wildtype peptides are represented by blue bars, putative HLA-I neoantigens by red bars, putative HLA-II neoantigens by orange bars, the observed ALYREF neoantigen in purple and the CEFT control peptide mix in grey. (G) Heatmap of expression of genes associated with CD8 and CD4 effector phenotypes for three clonotypes identified by scRNAseq from PBMCs from patient A119 exposed to putative HLA-I neoantigen 12-PTPRT. Each column represents a single cell.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

As an exploration of the specificity of our predictions, we performed scRNAseq to identify the corresponding cognate T-cell receptors for patient A119 candidate class I HLA neoantigen peptide 12-PTPRT (Figure 5 A, Table 2). Following stimulation of PBMCs with either the mutant or wildtype peptide we identified three clonotypes expanded only following exposure to the putative PTPRT neoantigen (Figure Figure 5 G). The most abundant clone, clonotype 2, had high expression for genes consistent with a CD8+ effector memory T-cell phenotype (CD8A, NKG7, GZMB, CCL5), whilst clonotypes 1 and 2 had gene expression patterns consistent with CD4+ effector memory T-cell phenotypes (CD4, SELL, CCR7)[37].

Discussion

The strategy of using HLA presented peptides as a basis for immunotherapy is long standing [39]. Researchers have sought to identify either peptides common to a cancer type, so-called tumour associated antigens, or peptides unique to a patient's tumour, so called neoantigens. Here we sought to identify HLA presented neoantigens in two NSCLC sub-types using a proteogenomics approach that combines exome sequencing, transcriptomics and mass spectrometry immunopeptidomics. In tumour samples from cohort of 24 NSCLC patients we found relatively high mutational burdens with exonic mutations characterised by a predominance of C>A transversions and containing small populations of T-cells. Consistent with previous reports [12] using mass spectrometry immunopeptidomics, we only directly identified one neoantigen amongst tens of thousands of peptide identifications. However, we utilised the remaining observations to inform our selection of neoantigens from ranked lists generated by in silico prediction algorithms [35] to the extent that we were able to identify positive functional assay neoantigens for 5 out of the 6 patients we were able to test. This included a positive response for the directly observed neoantigen. These findings represent a two-fold improvement over previous reports for neoantigen prediction and identification [20]. For one patient we were

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

able to test the specificity of the predictions, identifying a CD8+ T-cell clonotype that expanded only when exposed to the specific CD8+ neoantigen.

Although identifying neoantigens was the main aim of our study, the data included some interesting related observations: With the exception of TP53 and FAT1 in the HLA-I and HLA-II immunopeptidomes respectively, there was no correlation between driver gene mutation frequency and their peptide presentation frequency. This provides some circumstantial support for these two genes as sources for neoantigens [41]. TP53 mutations can be either truncal or late-stage [42], but a third of TP53 mutations occur in so-called hotspot regions [43], making them of interest both for early detection and as targets for immunotherapy [44,45]. Overall we found that the tumour peptidomes contained peptides derived from the same source proteins as healthy tissue. Furthermore, although somatic mutations are not random, as seen in the mutational signatures and driver genes, their distribution amongst cell compartments corresponds with gene expression frequency. There is no enrichment for mutations in genes expressing proteins in specific cell compartments. This implies a connection between the cell compartment from where the source protein derives and the subsequent HLA antigen processing pathway it primarily feeds. TP53 is a predominately a nuclear protein, whilst FAT1 is predominately extracellular, hence their higher frequencies in HLA I and II immunopeptidomes respectively. Whilst this might seem tautological, it does indicate that it would be unwise to preferentially select class I neoantigen predictions for FAT1 and vice versa for TP53, and yet this is not explicitly considered in existing neoantigen prediction algorithms. Hence we chose source protein cell compartment as a relevant neoantigen parameter.

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

There are various limitations in our study which might be addressed by future studies. Perhaps most significantly from a methodological perspective, mass spectrometry as a methodology does not have the amplification step found in many genomic sequencing methodologies. Therefore the strength of the input signal arises almost entirely from sample quality and preparation and sensitivity is determined by the mass spectrometer itself. The complexity of the input mixture and the differential ability of peptides to ionise, along with their relative abundances all affect what fraction of the immunopeptidome is identified. Various single molecule technologies are being developed that may address this problem, of which pore-based technologies, possibly in combination with fluorescence fingerprinting, seem well suited to identification of short peptides [48]. Sequencing peptides using pore technologies offers the tantalising prospect of providing much greater coverage of the immunopeptidome, and therefore direct observation of neoantigens. There are many challenges to this approach, not least post-translational modifications and the non-polar nature of protein peptides, but much progress has already been made [51].

In our study we only considered canonical neoantigens arising from missense variants. This was a limitation largely arising from choosing whole exome sequencing, but there is increasing evidence for non-canonical neoantigens arising from non-coding regions of the genome [54].

Here we have identified potential candidates for personalised vaccines that elicit strong positive responses in functional T-cell assays, but this doesn’t guarantee they would be effective as vaccines. The stage of the cancer at which the patient receives the vaccine may be crucial for efficacy. Chronic neoantigen exposure driving T-cells to dysfunctional states, late-differentiated T-cells dominating the tumour microenvironment, and loss of HLA heterozygosity are all reasons NSCLC may become harder to treat with neoantigen vaccines at later stages [10]. Heterogeneity in NSCLC tumours is likely to influence the efficacy of neoantigen based vaccines [31]. Differences between tumour cell immunopeptidomes raises the possibility of a

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

partial vaccine response. In the worst case this could create an evolutionary niche if slower growing tumour cells were destroyed, leaving more malignant tumour cells without competition .

Personalised neoantigen vaccines are already being trialled for the treatment of melanoma, glioblastoma and pancreatic cancer [57]. These trials rely on the delivery of mRNA containing a number of long sequences predicted to be processed into the final HLA presented neoantigens. The vaccine response rate is in the order of 50% of patients, so whilst these results are extremely promising, there is clearly room for improvement, including in the neoantigen selection process. Immunogenic peptides are identified by algorithms that incorporate machine learnt parameters such as peptide binding affinity [58] or proteosomal cleavage [59], or more recently using machine learning to identify features such as protein hotspots from large mass spectrometry immunopeptidomics datasets [22].

The principal difference in our approach is one of tactics rather than strategy, our tactical difference being to look at which proteins yield peptides presented by HLA molecules and then manually identifying supporting evidence for each neoantigen candidate protein in the literature. This tactic has some similarity to the 'Tübingen approach' for identification of tumour associated neoantigens which uses mass spectrometry proteomics identifications of HLA peptides to rank candidates [60], as used in the glioblastoma vaccine [56]. Whilst still far from successful, 87% of our predictions failed, it was twice as good than the current machine learning models. Our intention was to understand the direction of travel for better predictions, and our data strongly suggests that knowledge about the HLA peptides presented on each tumour is an important parameter in a neoantigen selection workflow.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Materials and Methods

Ethics statement

Ethical approval was obtained from the local research ethics committee (LREC reference 14-SC-0186 150975) and written informed consent was provided by the patients.

Tissue preparation

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

HLA typing

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

DNA and RNA extraction

DNA and RNA were extracted from tumor tissue that had been obtained fresh and immediately snap frozen in liquid nitrogen. Ten to twenty 10 µm cryosections were used for nucleic acid extraction using the automated Maxwell® RSC instrument (Promega) with the appropriate sample kit and according to the manufacturer’s instructions: Maxwell RSC Tissue DNA tissue kit and Maxwell RSC simplyRNA tissue kit, respectively. Similarly, DNA was extracted from snap frozen normal adjacent tissue as described above. DNA and RNA were quantified using Qubit fluorometric quantitation assay (ThermoFisher Scientific) according to the manufacturer’s instructions. RNA quality was assessed using the Agilent 2100 Bioanalyzer generating an RNA integrity number (RIN; Agilent Technologies UK Ltd.).

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Whole exome sequencing

The tumor and normal adjacent samples were prepared using SureSelect Human All Exon V7 library (Agilent, Santa Clara USA). 100 bp paired end reads sequencing was performed using the Illumina NovaSeq 6000 system by Edinburgh Genomics (Edinburgh, UK) providing ~100X depth. Reads were aligned to the 1000 genomes project version of the human genome reference sequence (GRCh38/hg38) using the Burrows-Wheeler Aligner (BWA; version 0.7.17) using the default parameters with the addition of using soft clipping for supplementary alignments. Following GATK Best Practices, aligned reads were merged [61], queryname sorted, de-duplicated and position sorted [62] prior to base quality score recalibration [63].

Somatic variant calling

Somatic variant calling was performed using three variant callers: Mutect2 (version 4.1.2.0) [64], Varscan (version 2.4.3) [65], and Strelka (version 2.9.2) [66]. For Mutect2, a panel of normals was created using 40 samples (20 male and 20 female) from the GBR dataset. Variants were combined using gatk GenomeAnalysisTK (version 3.8-1) with a priority order of Mutect2, Varscan, Strelka. Variants were then left aligned and trimmed, and multi-allelic variants split [67]. Hard filtering of variants was performed such that only variants that had a variant allele fraction > 5%, a total coverage > 20 and variant allele coverage > 5 were kept. Filtered variants were annotated using VEP (version 97) [68] and with their read counts (https://github.com/genome/bam-readcount) to generate the final filtered and annotated variant call files (VCF).

RNA sequencing

Samples were prepared TruSeq unstranded mRNA library (Illumina, San Diego, USA) and paired sequencing was performed using the Illumina NovaSeq 6000 system by Edinburgh Genomics (Edinburgh, UK). Raw reads were pre-processed to using fastp (version 0.20.0) [69].

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Filtered reads were aligned to the 1000 genomes project version of the human genome reference sequence (GRCh38/hg38 using hisat2 (version 2.1.0) [70], merged and then transcripts assembled and gene expression estimated with stringtie2 (version 1.3.5) [71] using reference guided assembly.

Mutanome generation

The annotated and filtered VCFs were processed using Variant Effect Predictor (version 97) [68] plugin ProteinSeqs to derive the amino acid sequences arising from missense mutations for each sample for use in immunopeptide analyses.

Neoantigen prediction

Variant call files were prepared for the pvacseq neoantigen prediction pipeline (version 1.5.10) [23,35] by adding tumor and normal DNA coverage, and tumor transcript and gene expression estimates using vatools (version 4.1.0) (http://www.vatools.org/). Variant call files of phased proximal variants were also created for use with the pipeline [72]. Prediction of neoantigens arising from somatic variants was then performed using pvacseq with the patient HLA allotypes to predict 8-11mer peptides for class I HLA and 15-mer peptides for class II HLA-DRB allotypes. Four binding algorithms were used for class I predictions (MHCflurry, MHCnuggetsI, NetMHC, PickPocket) and four for class II predictions (MHCnuggetsII, NetMHCIIpan, NNalign, SMMalign). Unfiltered outputs were post-processed in R [73] and split into individual tables for each peptide length and HLA allotype for each patient, and each table was then ranked according to the pvacseq score, where:

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

Immunopeptidomics

Snap frozen tissue samples were briefly thawed and weighed prior to 30s of mechanical homogenization (Fisher, using disposable probes) in 4 mL lysis buffer (0.02M Tris, 0.5% (w/v) IGEPAL, 0.25% (w/v) sodium deoxycholate, 0.15mM NaCl, 1mM EDTA, 0.2mM iodoacetamide supplemented with EDTA-free protease inhibitor mix). Homogenates were clarified for 10 min at 2,000g, 4°C and then for a further 60 min at 13,500g, 4°C. 2 mg of anti-MHC-I mouse monoclonal antibodies (W6/32) covalently conjugated to Protein A sepharose (Repligen) using DMP as previously described [74,75] were added to the clarified supernatants and incubated with constant agitation for 2 h at 4°C. The captured MHC-I/β_2m/immunopeptide complex on the beads was washed sequentially with 10 column volumes of low (isotonic, 0.15M NaCl) and high (hypertonic, 0.4M NaCl) TBS washes prior to elution in 10% acetic acid and dried under vacuum. The MHC-I-depleted lysate was then incubated with anti-MHC-II mouse monoclonal antibodies (IVA12) and MHC-II bound peptides were captured and eluted in the same conditions.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Immunopeptides were separated from MHC-I/β_2m or MHC-II heavy chain using offline HPLC on a C18 reverse phase column, as previously described [74]. Briefly, dried immunoprecipitates were reconstituted in buffer (1% acetonitrile,0.1% TFA) and applied to a 10cm RP-18e 100-4.6 chromolith column (Merck) using an Ultimate 3000 HPLC equipped with UV monitor. Immunopeptides were then eluted using a 15 min 0-40% linear acetonitrile gradient at a flow rate of 1 mL/min. Peptide fractions were eluted and pooled at between 0 and 30% acetonitrile, and the β_2m and MHC heavy chains eluted at >40% acetonitrile.

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

Proteomic data analysis

Raw spectrum files were analyzed using Peaks Studio 10.0 build 20190129 [76,77] and the data processed to generate reduced charge state and deisotoped precursor and associated product

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ion peak lists which were searched against the UniProt database (20,350 entries, 2020-04-07) plus the corresponding mutanome for each sample (~1,000-5,000 sequences) and contaminants list in unspecific digest mode. Parent mass error tolerance was set a 5ppm and fragment mass error tolerance at 0.03 Da. Variable modifications were set for N-term acetylation (42.01 Da), methionine oxidation (15.99 Da), carboxyamidomethylation (57.02 Da) of cysteine. As previously described, carbamidomethylated cysteines were treated as variable modifications due to the low concentration of 0.2 mM of iodoacetamide used in the lysis buffer to inhibit cysteine proteases [78]. A maximum of three variable modifications per peptide was set. The false discovery rate (FDR) was estimated with decoy-fusion database searches [76] and were filtered to 1% FDR. Downstream analysis and data visualizations of the Peaks Studio identifications was performed in R using associated packages [73,79].

Immunopeptide HLA assignment

Identified immunopeptides were assigned to their HLA allotype for each patient using motif deconvolution tools and manual inspection. For class I HLA peptides initial assignment used MixMHCp (version 2.1) [11,33] and for class II HLA peptides initial assignment used MoDec (version 1.1) [34]. Downstream analysis and data visualizations was performed in R using associated packages [73,79,80].

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,

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [73,79].

scRNAseq

Two peptide-expanded PBMC conditions were selected and prepared for combined single-cell RNAseq and TCRseq assays (10x Genomics, Table S5). Cells were thawed and counted; viability was >90%. Samples were incubated with TotalSeq C antibodies (Biolegend, Table 1), for 30 minutes to enable sample multiplexing. A maximum of 20,000 cells per condition were pooled into a 1.5mL low retention tube, with a maximum of 120,000 total PBMCs pooled. Following pooling, ice-cold PBS was added to make up to a volume of 1400uL. Cells were then centrifuged for 10 min (600g at 4C) and the supernatant was carefully removed. Sixty-six uL of resuspension buffer (0.22 um filtered ice-cold PBS supplemented with 10% foetal bovine serum, Sigma-Aldrich) was added to the tube and the pellet was gently but thoroughly resuspended. Following careful mixing, 66.6uL of the cell suspension was transferred to a PCR-tube for processing as per the manufacturer's instructions (10X Genomics). Briefly, single-cell RNA sequencing library preparation was performed as per the manufacturer's recommendations for the 10x Genomics 5' High-throughput Feature Barcode v2.0 (Dual Index) chemistry. Both initial

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

amplification of cDNA and library preparation were carried out with 13 cycles of amplification; V(D)J and cell surface protein libraries were generated using 9 and 8 cycles of amplification, respectively. Libraries were quantified and pooled according to equivalent molar concentrations and sequenced on Illumina NovaSeq6000 sequencing platform with the following read lengths: reads 1-101 cycles; reads 2 – 101 cycles; and i7 index – 8 cycles.

scRNAseq sequencing data was processed using cellranger-7.0.1 [81] using cellranger GRCh38 references for gene expression and VDJ sequences followed by post-processing using Seurat 5.0.1 [82] to filter for singlets only, percent mitochondrial genes < 12% and largest gene < 5%.

Data availability

EGA Study ID: EGAS00001005499

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[83] partner repository with the dataset identifier PXD028990 and 10.6019/PXD028990”. We would recommend you to also include this information in a much abridged form into the abstract itself, e.g. “Data are available via ProteomeXchange with identifier PXD028990.

Project Name: Immunopeptidomics guided identification of neoantigens in non-small cell lung cancer Project accession: PXD028990 Project DOI: 10.6019/PXD028990 Reviewer account details: Username: reviewer_pxd028990@ebi.ac.uk Password: dNbR5m6c

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgments

This study was supported by Cancer Research UK, Grant/Award Numbers: A21998, C328.

Instrumentation in the Centre for Proteomic Research is supported by the Biotechnology and Biological Sciences Research Council, Grant/Award Number: BM/M012387/1.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Types of lung cancer [Internet]. Available from: https://www.cancerresearchuk.org/about-cancer/lung-cancer/stages-types-grades/types

2. Cancer survival in england - office for national statistics [Internet]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancersurvivalinengland/stageatdiagnosisandchildhoodpatientsfollowedupto2018

3. Cancer Survival in England, cancers diagnosed 2016 to 2020, followed up to 2021 [Internet]. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/cancer-survival-in-england/cancers-diagnosed-2016-to-2020-followed-up-to-2021

4. Gainor JF, Shaw AT, Sequist LV, Fu X, Azzoli CG, Piotrowska Z, et al. EGFR Mutations and ALK Rearrangements Are Associated with Low Response Rates to PD-1 Pathway Blockade in NonSmall Cell Lung Cancer: A Retrospective Analysis. Clinical Cancer Research [Internet]. 2016;22:4585–93. Available from: https://clincancerres.aacrjournals.org/content/22/18/4585

5. Mazieres J, Drilon AE, Mhanna L, Milia J, Lusque A, Cortot AB, et al. Efficacy of immune-checkpoint inhibitors (ICI) in non-small cell lung cancer (NSCLC) patients harboring activating molecular alterations (ImmunoTarget). Journal of Clinical Oncology [Internet]. 2018;36:9010–0. Available from: https://ascopubs.org/doi/abs/10.1200/JCO.2018.36.15_suppl.9010

6. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, et al. Tracking the Evolution of Non–Small-Cell Lung Cancer. New England Journal of Medicine [Internet]. 2017 [cited 2021 Jun 1];376:2109–21. Available from: https://doi.org/10.1056/NEJMoa1616288

7. Neoantigen-directed immune escape in lung cancer evolution | nature [Internet]. Available from: https://www.nature.com/articles/s41586-019-1032-7#Sec4

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

8. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell [Internet]. 2015;160:48–61. Available from: https://www.cell.com/cell/abstract/S0092-8674(14)01639-0

9. Joshi K, Massy MR de, Ismail M, Reading JL, Uddin I, Woolston A, et al. Spatial heterogeneity of the T cell receptor repertoire reflects the mutational landscape in lung cancer. Nature Medicine [Internet]. 2019;25:1549–59. Available from: https://www.nature.com/articles/s41591-019-0592-2

10. Ghorani E, Reading JL, Henry JY, Massy MR de, Rosenthal R, Turati V, et al. The T cell differentiation landscape is shaped by tumour mutations in lung cancer. Nature Cancer [Internet]. 2020;1:546–61. Available from: http://dx.doi.org/10.1038/s43018-020-0066-y

11. Bassani-Sternberg M, Bräunlein E, Klar R, Engleitner T, Sinitcyn P, Audehm S, et al. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. Nature Communications [Internet]. 2016;7:13404. Available from: https://www.nature.com/articles/ncomms13404

12. Nicholas B, Bailey A, McCann KJ, Wood O, Walker RC, Parker R, et al. Identification of neoantigens in oesophageal adenocarcinoma. Immunology [Internet]. 2023;168:420–31. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/imm.13578

13. Newey A, Griffiths B, Michaux J, Pak HS, Stevenson BJ, Woolston A, et al. Immunopeptidomics of colorectal cancer organoids reveals a sparse HLA class I neoantigen landscape and no increase in neoantigens with interferon or MEK-inhibitor treatment. Journal for ImmunoTherapy of Cancer [Internet]. 2019;7:309. Available from: https://jitc.bmj.com/content/7/1/309

14. Marcu A, Bichmann L, Kuchenbecker L, Kowalewski DJ, Freudenmann LK, Backert L, et al. The HLA Ligand Atlas - A resource of natural HLA ligands presented on benign tissues. bioRxiv

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[Internet]. 2020 [cited 2021 Jun 1];778944. Available from: https://www.biorxiv.org/content/10.1101/778944v2

15. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: Application to the MHC class i system. Bioinformatics [Internet]. 2016;32:511–7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26515819

16. Juncker AS, Larsen MV, Weinhold N, Nielsen M, Brunak S, Lund O. Systematic Characterisation of Cellular Localisation and Expression Profiles of Proteins Containing MHC Ligands. Brusic V, editor. PLoS ONE [Internet]. 2009;4:e7448. Available from: http://dx.doi.org/10.1371/journal.pone.0007448

17. Müller M, Gfeller D, Coukos G, Bassani-Sternberg M. ‘Hotspots’ of antigen presentation revealed by human leukocyte antigen ligandomics for neoantigen prioritization. Frontiers in Immunology [Internet]. 2017;8. Available from: https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2017.01367

18. Gfeller D, Liu Y, Racle J. Contemplating immunopeptidomes to better predict them. Seminars in Immunology [Internet]. 2023;66:101708. Available from: http://dx.doi.org/10.1016/j.smim.2022.101708

19. Wells DK, van Buuren MM, Dang KK, Hubbard-Lucey VM, Sheehan KCF, Campbell KM, et al. Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. Cell [Internet]. 2020;183:818–834.e13. Available from: https://www.sciencedirect.com/science/article/pii/S0092867420311569

20. Buckley PR, Lee CH, Ma R, Woodhouse I, Woo J, Tsvetkov VO, et al. Evaluating performance of existing computational models in predicting CD8+ t cell pathogenic epitopes and cancer neoantigens. Briefings in Bioinformatics [Internet]. 2022;23:bbac141. Available from: https://doi.org/10.1093/bib/bbac141

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

21. Gartner JJ, Parkhurst MR, Gros A, Tran E, Jafferji MS, Copeland A, et al. A machine learning model for ranking candidate HLA class I neoantigens based on known neoepitopes from multiple human tumor types. Nature Cancer [Internet]. 2021;2:563–74. Available from: https://www.nature.com/articles/s43018-021-00197-6

22. Müller M, Huber F, Arnaud M, Kraemer AI, Altimiras ER, Michaux J, et al. Machine learning methods and harmonized datasets improve immunogenic neoantigen prediction. Immunity [Internet]. 2023;56:2650–2663.e6. Available from: https://www.cell.com/immunity/abstract/S1074-7613(23)00406-5

23. Hundal J, Kiwala S, McMichael J, Miller CA, Xia H, Wollam AT, et al. pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens. Cancer Immunology Research [Internet]. 2020;8:409–20. Available from: https://cancerimmunolres.aacrjournals.org/content/8/3/409

24. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature [Internet]. 2013;500:415–21. Available from: https://www.nature.com/articles/nature12477

25. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature [Internet]. 2013;502:333–9. Available from: https://www.nature.com/articles/nature12634

26. Gori K, Baez-Ortega A. sigfit: flexible Bayesian inference of mutational signatures [Internet]. 2020 Jan p. 372896. Available from: https://www.biorxiv.org/content/10.1101/372896v2

27. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature [Internet]. 2020;578:94–101. Available from: https://www.nature.com/articles/s41586-020-1943-3

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: The catalogue of somatic mutations in cancer. Nucleic Acids Research [Internet]. 2019;47:D941–7. Available from: https://doi.org/10.1093/nar/gky1015

29. Racle J, Jonge K de, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Valencia A, editor. eLife [Internet]. 2017;6:e26476. Available from: https://doi.org/10.7554/eLife.26476

30. Nilsson JB, Kaabinejadian S, Yari H, Kester MG D, Balen P van, Hildebrand WH, et al. Accurate prediction of HLA class II antigen presentation across all loci using tailored data acquisition and refined machine learning. Science Advances [Internet]. 2023;9. Available from: http://dx.doi.org/10.1126/sciadv.adj6367

31. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, et al. Tracking the evolution of nonsmall-cell lung cancer. New England Journal of Medicine [Internet]. 2017;376:2109–21. Available from: https://doi.org/10.1056/NEJMoa1616288

32. McGranahan N, Rosenthal R, Hiley CT, Rowan AJ, Watkins TBK, Wilson GA, et al. Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. Cell [Internet]. 2017 [cited 2021 Jun 1];171:1259–1271.e11. Available from: https://www.sciencedirect.com/science/article/pii/S0092867417311856

33. Gfeller D, Guillaume P, Michaux J, Pak H-S, Daniel RT, Racle J, et al. The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. The Journal of Immunology [Internet]. 2018; Available from: https://www.jimmunol.org/content/early/2018/11/13/jimmunol.1800914

34. Racle J, Michaux J, Rockinger GA, Arnaud M, Bobisse S, Chong C, et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. Nature Biotechnology. 2019;37:1283–6.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

35. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, et al. pVAC-seq: A genome-guided in silico approach to identifying tumor neoantigens. Genome Medicine [Internet]. 2016;8:11. Available from: https://doi.org/10.1186/s13073-016-0264-5

36. Gfeller D, Guillaume P, Michaux J, Pak H-S, Daniel RT, Racle J, et al. The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. The Journal of Immunology [Internet]. 2018; Available from: https://www.jimmunol.org/content/early/2018/11/13/jimmunol.1800914

37. Szabo PA, Levitin HM, Miron M, Snyder ME, Senda T, Yuan J, et al. Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. Nature Communications [Internet]. 2019;10:4706. Available from: https://www.nature.com/articles/s41467-019-12464-3

38. Rosenberg SA, Yang JC, Schwartzentruber DJ, Hwu P, Marincola FM, Topalian SL, et al. Immunologic and therapeutic evaluation of a synthetic peptide vaccine for the treatment of patients with metastatic melanoma. Nature Medicine [Internet]. 1998;4:321–7. Available from: https://www.nature.com/articles/nm0398-321

39. Weinschenk T, Gouttefangeas C, Schirle M, Obermayr F, Walter S, Schoor O, et al. Integrated functional genomics approach for the design of patient-individual antitumor Vaccines1. Cancer Research. 2002;62:5818–27.

40. Hsiue EH-C, Wright KM, Douglass J, Hwang MS, Mog BJ, Pearlman AH, et al. Targeting a neoantigen derived from a common *TP53* mutation. Science [Internet]. 2021;371. Available from: http://dx.doi.org/10.1126/science.abc8697

41. Castle JC, Kreiter S, Diekmann J, Löwer M, Roemer N van de, Graaf J de, et al. Exploiting the mutanome for tumor vaccination. Cancer Research [Internet]. 2012;72:1081–91. Available from: https://doi.org/10.1158/0008-5472.CAN-11-3722

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

42. Levine AJ, Jenkins NA, Copeland NG. The Roles of Initiating Truncal Mutations in Human Cancers: The Order of Mutations and Tumor Cell Type Matters. Cancer Cell [Internet]. 2019;35:10–5. Available from: http://dx.doi.org/10.1016/j.ccell.2018.11.009

43. Hollstein M, Sidransky D, Vogelstein B, Harris CC. p53 Mutations in Human Cancers. Science [Internet]. 1991;253:49–53. Available from: http://dx.doi.org/10.1126/science.1905840

44. Lin MJ, Svensson-Arvelund J, Lubitz GS, Marabelle A, Melero I, Brown BD, et al. Cancer vaccines: the next immunotherapy frontier. Nature Cancer [Internet]. 2022;3:911–26. Available from: http://dx.doi.org/10.1038/s43018-022-00418-6

45. Vadakekolathu J, Boocock DJ, Pandey K, Guinn B, Legrand A, Miles AK, et al. Multi-Omic Analysis of Two Common P53 Mutations: Proteins Regulated by Mutated P53 as Potential Targets for Immunotherapy. Cancers [Internet]. 2022;14:3975. Available from: http://dx.doi.org/10.3390/cancers14163975

46. Restrepo-Pérez L, Joo C, Dekker C. Paving the way to single-molecule protein sequencing. Nature Nanotechnology [Internet]. 2018;13:786–96. Available from: http://dx.doi.org/10.1038/s41565-018-0236-6

47. Alfaro JA, Bohländer P, Dai M, Filius M, Howard CJ, Kooten XF van, et al. The emerging landscape of single-molecule protein sequencing technologies. Nature Methods [Internet]. 2021;18:604–17. Available from: http://dx.doi.org/10.1038/s41592-021-01143-1

48. Lucas FLR, Versloot RCA, Yakovlieva L, Walvoort MTC, Maglia G. Protein identification by nanopore peptide profiling. Nature Communications [Internet]. 2021;12:5795. Available from: https://www.nature.com/articles/s41467-021-26046-9

49. Motone K, Nivala J. Not if but when nanopore protein sequencing meets single-cell proteomics. Nature Methods [Internet]. 2023;20:336–8. Available from: https://www.nature.com/articles/s41592-023-01800-7

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

50. Martin-Baniandres P, Lan W-H, Board S, Romero-Ruiz M, Garcia-Manyes S, Qing Y, et al. Enzyme-less nanopore detection of post-translational modifications within long polypeptides. Nature Nanotechnology [Internet]. 2023;18:1335–40. Available from: http://dx.doi.org/10.1038/s41565-023-01462-8

51. Motone K, Kontogiorgos-Heintz D, Wee J, Kurihara K, Yang S, Roote G, et al. Multi-pass, single-molecule nanopore reading of long protein strands with single-amino acid sensitivity [Internet]. 2023. Available from: http://dx.doi.org/10.1101/2023.10.19.563182

52. Laumont CM, Vincent K, Hesnard L, Audemard É, Bonneil É, Laverdure J-P, et al. Noncoding regions are the main source of targetable tumor-specific antigens. Science Translational Medicine [Internet]. 2018;10. Available from: http://dx.doi.org/10.1126/scitranslmed.aau5516

53. Chong C, Müller M, Pak H, Harnett D, Huber F, Grun D, et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. Nature Communications [Internet]. 2020;11. Available from: http://dx.doi.org/10.1038/s41467-020-14968-9

54. Ruiz Cuevas MV, Hardy M-P, Hollý J, Bonneil É, Durette C, Courcelles M, et al. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. Cell Reports [Internet]. 2021;34:108815. Available from: http://dx.doi.org/10.1016/j.celrep.2021.108815

55. Sahin U, Derhovanessian E, Miller M, Kloke B-P, Simon P, Löwer M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature [Internet]. 2017;547:222–6. Available from: https://www.nature.com/articles/nature23003

56. Hilf N, Kuttruff-Coqui S, Frenzel K, Bukur V, Stevanović S, Gouttefangeas C, et al. Actively personalized vaccination trial for newly diagnosed glioblastoma. Nature [Internet]. 2019;565:240–5. Available from: https://www.nature.com/articles/s41586-018-0810-y

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

57. Rojas LA, Sethna Z, Soares KC, Olcese C, Pang N, Patterson E, et al. Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer. Nature [Internet]. 2023;618:144–50. Available from: https://www.nature.com/articles/s41586-023-06063-y

58. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: Application to the MHC class i system. Bioinformatics [Internet]. 2016;32:511–7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26515819

59. Nielsen M, Lundegaard C, Lund O, Keşmir C. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. Immunogenetics [Internet]. 2005;57:33–41. Available from: https://doi.org/10.1007/s00251-005-0781-7

60. Singh-Jasuja H, Emmerich NPN, Rammensee H-G. The Tübingen approach: identification, selection, and validation of tumor-associated HLA peptides for cancer therapy. Cancer Immunology, Immunotherapy [Internet]. 2004;53:187–95. Available from: https://doi.org/10.1007/s00262-003-0480-x

61. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10.

62. Picard toolkit [Internet]. Broad Institute; 2019. Available from: http://broadinstitute.github.io/picard/

63. Genomics in the Cloud [Book] [Internet]. Available from: https://www.oreilly.com/library/view/genomics-in-the/9781491975183/

64. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2. bioRxiv [Internet]. 2019;861054. Available from: https://www.biorxiv.org/content/10.1101/861054v1

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

65. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research [Internet]. 2012;22:568–76. Available from: https://genome.cshlp.org/content/22/3/568

66. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nature Methods [Internet]. 2018;15:591–4. Available from: https://www.nature.com/articles/s41592-018-0051-x

67. Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, et al. HTSlib: C library for reading/writing high-throughput sequencing data. GigaScience [Internet]. 2021;10. Available from: https://doi.org/10.1093/gigascience/giab007

68. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biology [Internet]. 2016;17:122. Available from: https://doi.org/10.1186/s13059-016-0974-4

69. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics [Internet]. 2018;34:i884–90. Available from: https://doi.org/10.1093/bioinformatics/bty560

70. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology [Internet]. 2019;37:907–15. Available from: https://www.nature.com/articles/s41587-019-0201-4

71. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biology [Internet]. 2019;20:278. Available from: https://doi.org/10.1186/s13059-019-1910-1

72. Hundal J, Kiwala S, Feng Y-Y, Liu CJ, Govindan R, Chapman WC, et al. Accounting for proximal variants improves neoantigen prediction. Nature Genetics [Internet]. 2019;51:175–9. Available from: https://www.nature.com/articles/s41588-018-0283-9

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

73. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: https://www.R-project.org/

74. Purcell AW, Ramarathinam SH, Ternette N. Mass spectrometrybased identification of MHC-bound peptides for immunopeptidomics. Nature Protocols [Internet]. 2019;14:1687–707. Available from: https://www.nature.com/articles/s41596-019-0133-y

75. Bailey A, Nicholas B, Darley R, Parkinson E, Teo Y, Aleksic M, et al. Characterization of the class i MHC peptidome resulting from DNCB exposure of HaCaT cells. Toxicological Sciences [Internet]. 2020; Available from: https://doi.org/10.1093%2Ftoxsci%2Fkfaa184

76. Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, et al. PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. Molecular & Cellular Proteomics. 2012;11:M111010587.

77. Tran NH, Zhang X, Xin L, Shan B, Li M. De novo peptide sequencing by deep learning. Proceedings of the National Academy of Sciences of the United States of America. 2017;

78. Chong C, Marino F, Pak H, Racle J, Daniel RT, Müller M, et al. High-throughput and Sensitive Immunopeptidomics Platform Reveals Profound Interferonγ-Mediated Remodeling of the Human Leukocyte Antigen (HLA) Ligandome*. Molecular & Cellular Proteomics [Internet]. 2018;17:533–48. Available from: https://www.sciencedirect.com/science/article/pii/S153594762032260X

79. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. Journal of Open Source Software. 2019;4:1686.

80. Jessen LE. PepTools - an r-package for making immunoinformatics accessible [Internet]. 2018. Available from: https://github.com/leonjessen/PepTools

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.30.596609; this version posted June 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

81. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications [Internet]. 2017;8. Available from: http://dx.doi.org/10.1038/ncomms14049

82. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature Biotechnology [Internet]. 2023;42:293–304. Available from: http://dx.doi.org/10.1038/s41587-023-01767-y

83. Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Research [Internet]. 2021;50:D543–52. Available from: http://dx.doi.org/10.1093/nar/gkab1038