Last Updated on 2024-08-04

Supplementary Materials

This file contains information for the supplementary tables as `csv` files for: **Proteogenomics guided identification of functional neoantigens in non-small cell lung cancer** [1]. This data is available at repository https://github.com/ab604/lung-neoantigen-supplement and https://zenodo.org/doi/10.5281/zenodo.12820423

The column names and contents of the `csv` files in the `tables` folder are described below.

Supplementary Material 1: Patient information

Supplementary Material 1 is Table S1, a `csv` file containing patient information with 24 rows and 21 column variables. Each row in Table S1 represents observations for a single patient.

Table 1 provides descriptions of the values contained in each column of Table S1.

Table 1: Patient information Table S1 variables

Column name	Description
`accel_id`	CRUK Accelerator patient identifier
`target_lung_id`	Targeted Lung Health Check patient identifier
`tissue`	NSCLC type: Adenocarcinoma or Squamous cell carcinoma
`n_somatic_variants`	Total number of somatic variants identified by whole exome sequencing
`mut_burden_per_mb`	Mutational burden: mutations per million bases of DNA. Exome target size was 35.7 Mb

obs_class_I	Number of observed HLA I peptides by mass spec. immunopeptidomics
obs_class_II	Number of observed HLA II peptides by mass spec. immunopeptidomics
HLA	Class I and II HLA allotypes identified by genomic sequencing
wet_weight	Wet weight of tumour tissue
tumour_purity	Tumour purity as calculated from WES by ASCAT
tumour_ploidy	Tumour ploidy as calculated from WES by ASCAT
til_status	Tumour infiltrating T-cell status by immunohistochemistry: Low, Moderate, High or NA
weeks_post_surgery	Number of weeks since surgery
status_as_of_2021_01_19	Status since 2021-01-19: Alive, Deceased or NA
sex	Patient sex
date_of_diagnosis	Date of diagnosis
age_at_diagnosis	Age at diagnosis
smoking_status	Smoking status
notes_2	Notes about smoking history

Supplementary Material 2: NSCLC mutations

Supplementary Material 2 is Table S2, a compressed `csv` file containing all the mutations (variant calls) from the WES comparing tumour to normal adjacent tissue. It has 106,285 rows with 16 columns comprising the variants from 24 donors. Variant types are single nucleotide

variant, insertion, deletion and complex variant. Table 2 contains the description of the column variables.

Table 2: NSCLC VCF Table S2 variables

Column variable	Description
`accel_id`	CRUK Accelerator patient identifier
`vid`	Unique variant identifier
`chrom`	Chromosome
`pos`	Genomic coordinate
`ref`	Reference base
`alt`	Variant base
`info`	Information field from VCF file
`format`	Format of VCF variable columns
`sample_1`	Reference sample VCF variable values corresponding with format
`sample_2`	Tumour sample VCF variable values corresponding with format
`type`	Variant type: `snv`, `ins` , `del` or `complex` . Single nucleotide variant, insertion, deletion and complex variant respectively
`ensembl`	Ensembl gene identifier
`gene_name`	HGNC gene name
`vaf`	Variant allele frequency
`tissue`	Lung tumour tissue type: `Squamous` or `Adenocarcinoma`
`cell_compartment`	Cell compartment of protein product of gene,

Supplementary Material 4 and 4: pVACseq predicted neoantigens

Supplementary Material 3 and 4 are Tables S3 and S4. These are `csv` files containing all the pVACseq [2] predicted neoantigen peptides and their wildtype equivalents, Table 3 contains descriptions of the values contained in each column. Each row in Tables S3 and S4 represents one set of predictions i.e. one mutation and predicted neoantigen peptide per row.

Table S3 has 27,446 rows and 59 columns. Table S4 has 127,015 rows and 59 columns.

Table 3: pVACseq predictions Tables S3 and S4 variables

Column Name	Description
`sample`	CRUK Accelerator patient identifier
`Chromosome`	The chromosome of this variant
`Start`	The start position of this variant in the zero-based, half-open coordinate system
`Stop`	The stop position of this variant in the zero-based, half-open coordinate system
`Reference`	The reference allele
`Variant`	The alt allele
`Transcript`	The Ensembl ID of the affected transcript
`Transcript Support Level`	The transcript support level (TSL) of the affected transcript. `NA` if the VCF entry doesn't contain TSL information.

Ensembl Gene ID	The Ensembl ID of the affected gene
Variant Type	The type of variant. missense for missense mutations, inframe_ins for inframe insertions, inframe_del for inframe deletions, and FS for frameshift variants
Mutation	The amnio acid change of this mutation
Protein Position	The protein position of the mutation
Gene Name	The Ensembl gene name of the affected gene
HGVSc	The HGVS coding sequence variant name
HGVSp	The HGVS protein sequence variant name
HLA Allele	The HLA allele for this prediction
Peptide Length	The peptide length of the epitope
Sub-peptide Position	The one-based position of the epitope within the protein sequence used to make the prediction
Mutation Position	The one-based position of the start of the mutation within the epitope sequence. 0 if the start of the mutation is before the epitope
MT Epitope Seq	The mutant epitope sequence
WT Epitope Seq	The wildtype (reference) epitope sequence at the same position in the full protein sequence. NA if there is no wildtype sequence at this position or if more than half of the amino acids of the mutant epitope are mutated
Best MT Score Method	Prediction algorithm with the lowest mutant ic50 binding affinity for this epitope

Best MT Score	Lowest ic50 binding affinity of all prediction algorithms used
Corresponding WT Score	ic50 binding affinity of the wildtype epitope. NA if there is no WT Epitope Seq.
Corresponding Fold Change	Corresponding WT Score / Best MT Score. NA if there is no WT Epitope Seq.
Best MT Percentile Method	Prediction algorithm with the lowest binding affinity percentile rank for this epitope
Best MT Percentile	Lowest percentile rank of this epitope's ic50 binding affinity of all prediction algorithms used (those that provide percentile output)
Corresponding WT Percentile	binding affinity percentile rank of the wildtype epitope. NA if there is no WT Epitope Seq.
Tumor DNA Depth	Tumor DNA depth at this position. NA if VCF entry does not contain tumor DNA readcount annotation.
Tumor DNA VAF	Tumor DNA variant allele frequency (VAF) at this position. NA if VCF entry does not contain tumor DNA readcount annotation.
Tumor RNA Depth	Tumor RNA depth at this position. NA if VCF entry does not contain tumor RNA readcount annotation.
Tumor RNA VAF	Tumor RNA variant allele frequency (VAF) at this position. NA if VCF entry does not contain tumor RNA readcount annotation.

Normal Depth	Normal DNA depth at this position. NA if VCF entry does not contain normal DNA readcount annotation.
Normal VAF	Normal DNA variant allele frequency (VAF) at this position. NA if VCF entry does not contain normal DNA readcount annotation.
Gene Expression	Gene expression value for the annotated gene containing the variant. NA if VCF entry does not contain gene expression annotation.
Transcript Expression	Transcript expression value for the annotated transcript containing the variant. NA if VCF entry does not contain transcript expression annotation.
Median MT Score	Median ic50 binding affinity of the mutant epitope across all prediction algorithms used
Median WT Score	Median ic50 binding affinity of the wildtype epitope across all prediction algorithms used. NA if there is no WT Epitope Seq.
Median Fold Change	Median WT Score / Median MT Score. NA if there is no WT Epitope Seq.
Individual Prediction Algorithm WT and MT Scores (multiple)	ic50 binding affintity for the MT Epitope Seq and WT Eptiope Seq for the individual prediction algorithms used. Four binding algorithms were used for class I predictions (MHCflurry, MHCnuggetsI, NNalign, NetMHC, PickPocket) and four for class II predictions

	(MHCnuggetsII, NetMHCIIpan, NNalign, SMMalign).
cterm_7mer_gravy_score	Mean hydropathy of last 7 residues on the C-terminus of the peptide
max_7mer_gravy_score	Max GRAVY score of any kmer in the amino acid sequence. Used to determine if there are any extremely hydrophobic regions within a longer amino acid sequence.
difficult_n_terminal_residue (T/F)	Is N-terminal amino acid a Glutamine, Glutamic acid, or Cysteine?
c_terminal_cysteine (T/F)	Is the C-terminal amino acid a Cysteine?
c_terminal_proline (T/F)	Is the C-terminal amino acid a Proline?
cysteine_count	Number of Cysteines in the amino acid sequence. Problematic because they can form disulfide bonds across distant parts of the peptide
n_terminal_asparagine (T/F)	Is the N-terminal amino acid an Asparagine?
asparagine_proline_bond_count	Number of Asparagine-Proline bonds. Problematic because they can spontaneously cleave the peptide
b_rank	Rank of binding score: 1/median neoantigen binding affinity . Lower is better
f_rank	Rank of fold change: the difference in median binding affinity between neoantigen and wildtype peptide (agretopicity). Higher is better.
m_rank	Ranks of mutant allele expression: the product of

	gene_expression and tumor_rna_vaf . Higher is better.
d_rank	Rank of the tumor_dna_vaf . Higher is better.
score	A score is calculated from the above ranks with the following formula: b_rank + f_rank + (m_rank * 2) + (d_rank/2) . Higher is better
rank_score	The score converted to a rank, with the best being 1, splitting ties by first. Lower is better
rank_percent	The percentage rank score. Lower is better.

Supplementary Material 5: Tested neoantigens

Supplementary Material 5 is Table S5, a csv file with 70 rows and 17 column variables for the neoantigen peptide predictions tested by IFN-γ ELISPOT using autologous PBMCs. Each row in Table S4 represents one neoantigen peptide and its wildtype equivalent and Table 4 contains descriptions of the values contained in each column of Table S5.

Table 4: Tested candidate neoantigen peptides Table S4 variables

Column name	Description
accel_id	CRUK Accelerator patient identifier
gene_name	Gene
mt_epitope_seq	Mutated (neoantigen) peptide sequeunce
wt_epitope_seq	Wildtype peptide sequence
peptide_length	Peptide length

table_name	Identifier in the form accel_id / predicted_hla_allotype / peptide_length e.g. A119/DRB1*04:04/15
mutation	The mutation From/To
protein_position	Location of the mutation in the source protein, UNIPROT sequence number.
Obs_I	The number of peptides from the source protein observed by mass spectrometry observed in HLA-I immunopeptidome
Obs_II	The number of peptides from the source protein observed by mass spectrometry observed in HLA-II immunopeptidome
median_mt_score	The median pVACseq predicted binding affinity of the neoantigen peptide
median_wt_score	The median pVACseq predicted binding affinity of the wildtype peptide
median_fold_change	The ratio between the median neoantigen affinity and wildtype peptide affinity
rank_percent	The overall rank percentage for the neoantigen from pVACseq for the peptide of that length and HLA allotype.
mean_sfc_mt	Mean IFN-**γ** ELISPOT spot forming cells per million cells for the neoantigen peptide
mean_sfc_wt	Mean IFN-**γ** ELISPOT spot forming cells per million cells for the wildtype peptide
elispot_response	ELISPOT response category: Strong, Weak or None

Table S6 List of patient samples selected for single-cell RNA and TCR sequencing and TotalSeq C antibodies (Biolegend).

Patient ID and condition	TotalSeq C Hashtag ID	Hashtag barcode
A119_PTPRT-12_**MUT**	C0255	AAGTATCGTTTCGCA
A119_PTPRT-12_**WT**	C0256	GGTTGCCAGATGTCA

References

1. Nicholas B, Bailey A, McCann KJ, Wood O, Currall E, Johnson P, et al. Proteogenomics guided identification of functional neoantigens in non-small cell lung cancer. 2024. Available: http://dx.doi.org/10.1101/2024.05.30.596609

2. 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. 2016;8: 11. doi:10.1186/s13073-016-0264-5