The neo-open reading frame peptides that comprise the tumor framome are a rich source of neoantigens for cancer immunotherapy

Michael V. Martin[1], Salvador Aguilar-Rosas[1], Katka Franke[1], Mark Pieterse[1], Jamie van Langelaar[1], Renée Schreurs[1], Maarten F. Bijlsma[5,15], Marc G. Besselink[13,15], Jan Koster[5], Wim Timens[2], Mustafa Khasraw[4], David M. Ashley[11], Stephen T. Keir[4], Christian H. Ottensmeier[8], Emma V. King[9], Joanne Verheij[18], Cynthia Waasdorp[5], Peter J. M. Valk[12], Sem A. G. Engels[16], Ellen Oostenbach[16], Jip T. van Dinter[16], Damon A. Hofman[16], Juk Yee Mok[17], Wim J.E. van Esch[17], Hanneke Wilmink[14,15], Kim Monkhorst[10], Henk M.W. Verheul[19], Dennis Poel[6], T. Jeroen N. Hiltermann[3], Léon C.L.T. van Kempen[2, 20], Harry J.M. Groen[3], Joachim G.J.V. Aerts[7], Sebastiaan van Heesch[16], Bob Löwenberg[1], Ronald Plasterk[1], and Wigard P. Kloosterman[1,21]

[1]CureVac Netherlands B.V., Science Park 106, Amsterdam, the Netherlands
[2]Department of Pathology and Medical Biology, University of Groningen, University Medical Center Groningen, the Netherlands
[3]Department of Pulmonary Diseases, University of Groningen, University Medical Center Groningen, the Netherlands
[4]Duke University Medical Center, Duke University
[5]Amsterdam UMC location University of Amsterdam, Center for Experimental and Molecular Medicine, Laboratory for Experimental Oncology and Radiobiology, Meibergdreef 9, Amsterdam, the Netherlands
[6]Department of Medical Oncology, Radboud University Medical Center, Nijmegen, the Netherlands
[7]Erasmus Medical Center, 6993, Pulmonary Medicine, Rotterdam, Zuid-Holland, the Netherlands
[8]Liverpool Head and Neck Centre, Institute of Systems, Molecular and Integrative Biology, University of Liverpool & Clatterbridge Cancer Center NHS Foundation Trust, Liverpool, United Kingdom
[9]Department of Otorhinolaryngology, Head & Neck Surgery, Poole Hospital, Poole, United Kingdom
[10]Netherlands Cancer Institute, Amsterdam, the Netherlands
[11]Preston Robert Tisch Brain Tumor Center, Department of Neurosurgery, Duke University
[12]Department of Hematology, Erasmus University Medical Center, Rotterdam, the Netherlands
[13]Amsterdam UMC, location University of Amsterdam, Department of Surgery, De Boelelaan 1117, Amsterdam, the Netherlands
[14]Amsterdam UMC, location University of Amsterdam, Department of Medical Oncology, De Boelelaan 1117, Amsterdam, the Netherlands

[15]Cancer Center Amsterdam, Imaging and Biomarkers, Amsterdam, the Netherlands
[16]The Princess Máxima Center for Pediatric Oncology, Utrecht, the Netherlands
[17]Sanquin Reagents,Sanquin, Amsterdam, the Netherlands
[18]Amsterdam UMC, location University of Amsterdam, Department of Pathology, De Boelelaan 1117, Amsterdam, the Netherlands
[19]Department of Medical Oncology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands
[20]University of Antwerp, Antwerp University Hospital, 2650 Edegem, Belgium
[21]Correspondence: wigard.kloosterman@curevac.com

Running Title: The tumor framome is a source of neoantigens

Keywords: neoantigen, neo-open reading frame, structural genomic variation, cancer genomics

Funding Information: Eurostars European Commission: grant no E! 114238 to Wigard P. Kloosterman, European Fund for Regional development: grant KvW-00279 to Wigard P. Kloosterman and Ronald Plasterk, Top Consortia for Knowledge and Innovation: Public Private Partnership grant EMCLSH20002 to Joachim G.J.V. Aerts and Wigard P. Kloosterman.

Disclosures: Mustafa Khasraw obtained Institutional research funding from AbbVie, Bristol Myers Squibb, Celldex and Specialized Therapeutics, and honoraria for consultancy/advisory roles with AbbVie, Bristol Myers Squibb, Eli Lilly, Ipsen, Janssen and Janssen, Pfizer, Roche, Voyager Therapeutics and the JAX lab for genomic research. Ronald H.A. Plasterk is shareholder and former employee of CureVac Netherlands B.V. and holds restricted stock units of CureVac. Bob Löwenberg is shareholder of CureVac and former consultant of CureVac Netherlands B.V.. Wigard P. Kloosterman, Michael V. Martin, Salvador Aguilar-Rosas, Katka Franke, and Mark Pieterse are employees of CureVac Netherlands B.V. and hold restricted stock units of CureVac.

Abstract

Identification of immunogenic cancer neoantigens as targets for therapy is challenging. Here, we integrate whole genome and long-read transcript sequencing of cancers to identify the collection of neo-open reading frame peptides (NOPs) expressed in tumors. We termed this collection of NOPs the tumor framome. NOPs represent tumor-specific peptides that are different from wild-type proteins and may be strongly immunogenic. We describe a class of hidden NOPs that derive from structural genomic variants involving an upstream protein coding gene driving expression and translation of non-coding regions of the genome downstream of a rearrangement breakpoint, i.e. where no gene annotation or evidence for transcription exists. The entire collection of NOPs represents a vast number of possible neoantigens particularly in tumors with many structural genomic variants and a low number of missense mutations. We show that NOPs are immunogenic and epitopes derived from NOPs can bind to MHC class I molecules. Finally, we provide evidence for the presence of memory T cells specific for hidden NOPs in peripheral blood from a patient with lung cancer. This work highlights NOPs as a major source of possible neoantigens for personalized cancer immunotherapy and provides a rationale for analyzing the complete cancer genome and transcriptome as a basis for detection of NOPs.

Synopsis

Identifying immunogenic cancer neoantigens for cancer vaccine design is challenging. The authors uncover NOPs as a widespread source of neoantigens derived from structural genomic variants and indels. The findings open new avenues towards personalized immunotherapies for cancer.

Introduction

Identifying and selecting the most immunogenic cancer neoantigens is a key challenge to maximizing the potency of personalized immunotherapy (1, 2). Several studies have combined whole exome sequencing (WES) with short-read RNA sequencing to identify neoantigens derived from expressed somatic mutations in coding regions of the genome (1, 3). Although missense mutations are the most frequent type of protein-altering coding mutation found in cancer genomes (4), they represent a small difference with respect to wild-type (WT) epitopes. There is mounting evidence that the degree of epitope dissimilarity to WT self-peptides is a critical determining factor in neoantigen immunogenic potential (5, 6), which justifies a search for alternative sources of neoantigens.

Neoantigens arising from neo-open reading frames (neo-ORFs) produce long stretches of completely foreign protein sequences referred to as neo-open reading frame peptides (NOPs) (6, 7). NOPs arising from genomic events have been ascribed to four main sources: structural genomic variants (SVs) (8–12), small insertions/deletions (indels) (6, 7), mutations affecting mRNA splicing (13–15), and mutations disrupting stop codons (16).

We integrated whole genome sequencing (WGS) with long- and short-read RNA sequencing to characterize the tumor framome, the set of all NOPs expressed by a tumor as a result of genomic mutations in *cis*. Our approach detects full-length transcripts encoding NOPs at single-molecule resolution, thereby accounting for isoform diversity. The approach was applied to 61 tumors across six cancer types, demonstrating a framome size of over 2,000 amino acids for some tumors. We identify a class of possible neoantigens, which we referred to as 'hidden' NOPs, in which a known protein coding gene drives transcription and translation of a usually non-coding region of the genome that has been placed downstream via an SV. The entire tumor framome represents an attractive target for personalized immunotherapy.

Materials and Methods

Patient samples and ethics

Fresh frozen tumor biopsies and corresponding blood samples or normal control tissue were obtained from 61 patients with cancer treated at different clinical centers as outlined in **Supplementary Table S1 -** *Sample Overview*. Written informed consent and ethical approval were obtained for each sample for studying tumor DNA and RNA sequencing information. Patient samples were obtained as part of the following clinical studies: OLS041-202100773 Framoma (Oncolifes, University Medical Center Groningen), AMC 2014 181 BioPAN (Amsterdam UMC), IRBdm21-018 (Netherlands Cancer Institute), 09H050190 (LREC, University of Liverpool), Pro000074343 (Duke University), PRIMA study (Erasmus Medical Center Rotterdam), NCT01792934 (Radboud University Medical Center).

Whole genome sequencing

Genomic DNA was isolated from tumor biopsies and control tissue (blood or adjacent normal tissue) using Qiagen DNeasy. As input, 50-200 ng of DNA was sheared to an average length of 450 bp using a Covaris ultrasonicator. Standard TruSeq Nano LT library preparation (Illumina) with 8 PCR cycles was performed. Barcoded libraries were sequenced on Illumina NovaSeq instruments with 2x151bp settings, to an average coverage depth of 100X (tumor samples) and 35X (control samples). FASTQ generation was done using Illumina bcl2fastq (v2.20.0.42). Sequencing reads were mapped to human reference genome GRCh37 using BWA (0.7.17) with settings:

```
bwa mem -Y -t ${THREADS} ${GENOME} ${R1} ${R2}
```

Somatic genomic variants were called from aligned sequencing data using a custom pipeline (4) (https://github.com/hartwigmedical/pipeline5/tree/master/cluster/src/main/java/com/hartwig/pipeline). Detailed whole genome sequencing statistics and variant calling methods are outlined in **Supplementary Table S2 -** *WGS Overview*.

Short-read RNA sequencing

Total RNA was isolated from fresh frozen tumor samples using NucleoSpin RNA isolation (Machery Nagel). cDNA library prep was performed with the KAPA RNA HyperPrep kit (Roche) with RiboErase, using 100 ng of total RNA, which was chemically sheared for 7 minutes. cDNA was PCR amplified for 15 cycles. Libraries were sequenced on an Illumina NovaSeq system to a minimal depth of 50M paired reads (100M tags) per cDNA library based on 2x151bp settings. FASTQ generation was done using Illumina bcl2fastq (v2.20.0.42). cDNA sequencing reads were mapped to the human reference genome GRCh37 using STAR
(2.7.9a) with settings:

```
STAR --runThreadN ${THREADS} \
--genomeDir ${STAR_GENOME} \
--readFilesCommand zcat \
--readFilesIn ${R1} ${R2} \
--outFileNamePrefix ${SAMPLE_NAME} \
--twopassMode Basic \
--outSAMunmapped Within \
--alignMatesGapMax 200000 \
--alignIntronMax 200000 \
--outSAMattrRGline ID:GRPundef \
--outFilterMultimapNmax 100 \
--outFilterIntronStrands None \
--outSJfilterIntronMaxVsReadN 200000 \
--outSAMtype BAM SortedByCoordinate \
--chimOutType Junctions WithinBAM SoftClip
```

Detailed short-read cDNA sequencing statistics per tumor sample are outlined in **Supplementary Table S3 -** *Short Read RNA Overview*.

Long-read RNA Nanopore sequencing

About 500ng to 2 microgram of total RNA, isolated as described above (see *Short-read RNA sequencing*), was used as input for double-stranded cDNA preparation using TeloPrime Full-Length cDNA Amplification kit V2 (Lexogen) according to the manufacturer's specifications. TeloPrime selects mRNA molecules containing a *5'* CAP and a *3'* poly-A tail. For some samples, polyA-selected RNA was used as input for TeloPrime cDNA preparation (see **Supplementary Table S4** *Long Read RNA Overview*). For those cases, selection of polyA mRNA was performed using Dynabeads mRNA Purification kit (Invitrogen) and between 20-100ng of polyA-selected mRNA was used as input for TeloPrime. Between 11-20 PCR cycles were performed for each sample. Double-stranded cDNA was used as input for preparation of a Nanopore sequencing library using SQK-LSK109. Libraries were sequenced on GridION or PromethION systems (Oxford Nanopore Technologies) to a depth of between 20M-100M reads. Long cDNA Nanopore reads were mapped to human reference genome GRCh37 using Minimap2 (2.17) with settings:

```
minimap2 -t
${THREADS} \ -ax
splice \
--split-prefix tmp \
-un \
-k14 \
--eqx \
${GENOME} ${FASTQ_READS}
```

Detailed long-read cDNA sequencing statistics per tumor sample are outlined in **Supplementary Table S4 –** *Long Read RNA Overview*.

NOP identification methods

All core steps in the NOP identification pipeline including genome reconstruction, RNA isoform identification, isoform translation prediction, and NOP identification were implemented in python (https://www.python.org/). Nextflow (17) was used to integrate these steps with RNA mapping and read extraction.

Tumor genome reconstruction

To identify neo-ORFs and corresponding NOPs, a tumor-specific reference genome was generated for each sample onto which long- and short-read RNA could be aligned. These tumor-specific reference genomes consisted of collections of contigs that captured the local effects of somatic mutations. For SVs, these contigs were identified by taking sets of ungapped chimeric RNA alignments and attempting to explain their transcript structure at the genome level through SVs.

This RNA-guided approach starts with the alignment of RNA to a base reference genome as outlined above and proceeds as illustrated in **Supplementary Fig. S1**. Let R be the set of RNA reads with at least one alignment within ε_{sv} base pairs (default 200 kB) of an SV breakend and that has at least one supplementary or secondary alignment. Let A_r be the set of primary, supplementary, and secondary alignments of read $r \in R$. For given alignment $a_i \in A_r$ let a_{iqs} be the start position within the query sequence of a_i as measured from the 5′ end of the RNA read r. Similarly, let a_{iqe} be the query end position. Let a_{is}, a_{irs}, and a_{ire} be the reference alignment strand, strand-specific reference start, and strand-specific reference end of a_i, respectively where $a_{irs} \leq a_{ire}$ if and only if a_{is} is positive. A set Q_r of ordered sets p of alignments in A_r can be defined as:

$$Q_r = \{\boldsymbol{q} \mid q_i \in A_r, q_{i_{qs}} < q_{i+1_{qs}} \text{ and } |q_{i+1_{qs}} - q_{i_{qe}}| < \epsilon_p \forall i \in (1, \dots, \|\boldsymbol{q}\| - 1)\}$$

The elements of Q_r represent collections of consecutive segments of the read r that are non-linearly aligned to the reference genome. A gap or overlap buffer of ε_p is utilized to allow for soft or hard-clipping, erroneous indels, and homology at the beginning and ends of the alignments. To arrive at a non-redundant (excluding prefix/suffix paths) set of chimeric RNA paths for read r, the set P_r can be defined as:

$$P_r = \{\boldsymbol{p} \mid \boldsymbol{p} \in Q_r, \boldsymbol{p} \not\subset \boldsymbol{q} \forall \boldsymbol{q} \in Q_r \setminus \{\boldsymbol{p}\}\}$$

To find possible underlying tumor contig regions from which the proposed chimeric RNA structures within P_r may have arisen, it is necessary to find paths of SVs that connect the beginning and ends of consecutive chimeric alignments, referred to as chimeric introns. Each chimeric RNA path $\boldsymbol{p}$ in P_r contains a set M_p of size $\|\boldsymbol{p}\| - 1$ such chimeric introns m where m_L and m_H represents the lower and upper alignments on each side of the chimeric intron. This set M_p can be defined as:

$$M_{\boldsymbol{p}} = (m | m_L = \boldsymbol{p}_i \text{ and } m_H = \boldsymbol{p}_{i+1} \forall i \in (1, \dots, \|\boldsymbol{p}\| - 1))$$

A chimeric RNA path is considered supported by somatic genomic events if there is a conceivable path through the tumor genome that connects the end of the first chimeric intron alignment to the start of the second chimeric intron alignment for each chimeric intron in the path. To determine this for each path, a directed graph $G_{\boldsymbol{p}}$ is constructed that represents all possible connections within the tumor genome. The end/start loci of each chimeric intron can then be anchored onto $G_{\boldsymbol{p}}$ in order to find a valid path across the chimeric intron. To construct this graph, let the sample SVs be represented by a set B of breakends b where b_c, b_p, b_s, b_m are the breakend chromosome, position, strand, and mate breakend, respectively. Let the vertex set V ($G_{\boldsymbol{p}}$) consist of vertices v where v_c, v_p, v_s correspond to chromosome, position, and strand of genomic loci. Let two identical sets of breakend vertices be V_{source} and V_{sink} be defined as:

$$V_{sources} = \{v \mid v_c = b_c \text{ and } v_p = b_p \text{ and } v_s = b_s \forall b \in B\}$$
$$V_{sinks} = \{v \mid v_c = b_c \text{ and } v_p = b_p \text{ and } v_s = b_s \forall b \in B\}$$

Let the sets of lower-alignment and upper-alignment chimeric intron vertice sets be defined as:

$$V_L = \{v \mid v_c = m_{L_c} \text{ and } v_p = m_{L_p} \text{ and } v_s = m_{L_s} \forall \boldsymbol{p} \in P_r \forall m \in M_{\boldsymbol{p}}\}$$
$$V_H = \{v \mid v_c = m_{H_c} \text{ and } v_p = m_{H_p} \text{ and } v_s = m_{H_s} \forall \boldsymbol{p} \in P_r \forall m \in M_{\boldsymbol{p}}\}$$

The vertex set $V(G_{\boldsymbol{p}})$ is then:

$$V(G_{\boldsymbol{p}}) = V_{sources} \cup V_{sinks} \cup V_L \cup V_H$$

Two types of connections between genomic loci are possible within the rearranged tumor genome: those which occur between points on the same strand of the same chromosome in the normal reference genome and those which occur due to SVs. The edge set E_{WT} represents WT connections which point from source vertices to sink vertices:

$$E_{WT^+} = (v,u) \mid v \in V_{sources}, u \in V_{sinks}, v_c = u_c, v_s = +, v_s = -, v_p \leq u_p$$
$$E_{WT^-} = \{(v,u) \mid v \in V_{sources}, u \in V_{sinks}, v_c = u_c, v_s = -, v_s = +, v_p \geq u_p\}$$
$$E_{WT} = E_{WT^+} \cup E_{WT^-}$$

The edge set E_{SV} represents connections between breakpoints due to SVs which point from sink vertices to their partner breakend source vertices:

$$E_{SV} = \{(v,u) \mid v \in V_{sinks}, u \in V_{sources}, v_c = b_c, v_p = b_p, v_s = b_s, u_c = b_{m_c}, u_p = b_{m_p}, u_s = b_{m_s} \forall b \in B\}$$

The edge set E_M represents the connections between lower chimeric intron alignments to sink breakend loci as well as the connections between source breakend loci to upper chimeric intron alignments:

$$E_{L^+} = (v,u) \mid v \in V_L, u \in V_{sinks}, v_c = u_c, v_s = +, v_s = -, v_p \leq u_p$$
$$E_{L^-} = \{(v,u) \mid v \in V_L, u \in V_{sinks}, v_c = u_c, v_s = -, v_s = +, v_p \geq u_p\}$$
$$E_{H^+} = \{(v,u) \mid v \in V_{sources}, u \in V_H, v_c = u_c, v_s = +, v_s = +, v_p \leq u_p\}$$
$$E_{H^-} = \{(v,u) \mid v \in V_{sources}, u \in V_H, v_c = u_c, v_s = -, v_s = -, v_p \geq u_p\}$$
$$E_M = E_{L^+} \cup E_{L^-} \cup E_{H^+} \cup E_{H^-}$$

The edge set $E(G_{\boldsymbol{p}})$ can now be specified as:

$$E(G_{\boldsymbol{p}}) = E_{WT} \cup E_{SV} \cup E_M$$

Let the weight of edge tuples in $E(G_{\boldsymbol{p}})$ be defined as the genomic distance between each loci vertex, where connections between mate breakends have a distance of zero:

$$w(e) = \begin{cases} |e_{1_p} - e_{2_p}|, & e_2 \in V_{sink} \\ 0, & e_2 \in V_{source} \end{cases}$$

Together $V(G_{\boldsymbol{p}})$, $E(G_{\boldsymbol{p}})$, and $w: e \rightarrow \mathbb{N}_0$ fully define $G_{\boldsymbol{p}}$. This RNA-SV graph was built using the python *networkx* package (18), and Dijkstra's algorithm was used to find the shortest weighted genomic path between every m_L to m_H chimeric intron vertices through an alternating set of *sink* and *source* breakend vertices. The genomic paths of each chimeric intron were appended in the order of appearance in each path $\boldsymbol{p}$ to produce a contig starting at the first chimeric intron start anchor and ending at the final chimeric intron end anchor. The contigs specified by the set of these shortest chimeric intron paths were padded at the beginning and end by prepending/appending enough sequence to encompass the full chimeric RNA alignment at the start/end of the contig and any annotated genes overlapping these start/end alignments. Additionally, to better resolve RNA structures with exons overlapping a structural variant (a minority of cases), 500 bps on either side of the breakpoints were added as buffer to allow for structural breakpoint position uncertainty. This set of contigs was collapsed by removing all contigs whose sequence was a strict subset of another. This set of non-redundant contigs was appended to the tumor-specific reference genome. Of note, the cell line NOPs were identified by an earlier version of the approach that also appended genic regions affected by SVs to the tumor reference genome without prior chimeric RNA support.

Small variants predicted to lead to NOPs were also used as a basis for tumor-specific contig construction. To identify all indels possibly leading to NOPs, indels within the bounds of protein coding genes were identified. If the indel was within the exonic boundaries of any protein coding exon, it was selected for inclusion in variants used for reconstruction. If the indel was in a non-protein coding region of the gene such as an intron or UTR, the variant was included if there was at least one long RNA read which covered the indel locus. Stoploss variants were identified by selecting variants which disrupted an annotated known stop codon. Mutations leading to novel splice junctions were also selected for inclusion in the reconstruction. A portion of the reference chromosome containing each variant was extracted to include the entire region of any genes and/or long reads overlapping each variant position. The genomic change specified by each small variant was then applied to this contig with each variant producing a contig that was appended to the tumor-specific reference genome.

Novel Splice Junction Identification

Short-read RNA splice junctions were considered novel and tumor-specific if they were absent in the healthy tissues sequenced as part of the GTEx database (19) and were associated with a predicted causal somatic variant. The pre-compiled STAR splice junctions for GTEx v6 were downloaded from the Recount2 webserver and used as the normal tissue splice junction database (20). Two general classes of variants are considered to potentially cause novel splice junctions. The first class consists of variants near un-annotated splice sites of expressed splice junctions. These splice-gain variants are known to often lead to the formation of more-canonical splicing signals (13). The second class of genomic variants leading to novel RNA splice junctions are variants that disrupt annotated splice sites by changing the genomic context of an annotated splice donor or acceptor. This splice site disruption may lead to full exon skipping or partial intron retention/truncation. The effect

zone of these splice-disrupting variants was therefore taken as the *5'* start of the exon before the variant-affected exon up through the *3'* end of the exon after the variant-affected exon, including intronic regions. Any tumor-specific splice junction with splice points within this genomic range was considered caused in *cis* by the splice-disrupting variant.

Tumor-specific RNA isoform identification

After alignment to the reconstructed tumor genome, tumor-specific RNA isoforms were identified through a combination of high-accuracy short reads and long but error prone long reads. Short-read junctions were used to correct the splice points of long-read alignments via a novel Bayesian splice-correction model illustrated in (**Supplementary Fig. S2**). Long-read splice sites were corrected to short-read splice sites using short-read junctions in which both the *5'* and *3'* splice sites were within a ε_{splice} basepair (default 15) window of the respective long read splice sites. For cases in which multiple short-read junctions satisfied this criteria for a given long-read junction, the most likely short read junction was chosen via a Bayesian model in which the posterior probability that an observed long read junction arose from an mRNA with a given short read junction was calculated according to:

$$P(s_i \mid F_i, T_i) = \frac{P(F_i, T_i \mid s_i)P(s_i)}{P(F_i, T_i)}$$

where the event s_i is the long read arising from the splice junction i, and the event F_i, T_i is the observation of a long read having a given *5'* or *3'* distance pair from its underlying original splice sites. The prior probability that a long read arose from an RNA molecule with splice junction i was calculated according to:

$$P(s_i) = \frac{R_i}{R}$$

where R_i is the number of short reads supporting junction i and R is the total spliced reads within the long read splice site window. The probability of observing the splice offset pair F_i, T_i given that the long read arose from an RNA molecule with splice junction i was calculated according to:

$$P(F_i, T_i \mid s_i) = \frac{N_{F_i T_i}}{N}$$

where N_{FiTi} is the number of times the given offset pair occurred in all other long-read splice junction corrections which were unambiguous because a single short-read junction was present within the correction window and N is the total number of unambiguously corrected junctions. Both N_{FiTi} and N were calculated for each sample based on mapping of the short- and long-read RNA to the base reference genome. The total probability of observing the long-read offset pair F_i, T_i irrespective of any given short read junction can be calculated according to:

$$P(F_i, T_i) = \sum_{j=1}^{n} P(F_i, T_i \mid s_j)P(s_j)$$

where the summation is taken over the n splice junctions within the long-read junction window. Combining these expression gives:

$$P(F_i, T_i) = \frac{N_{F_i T_i} R_i}{\sum_{j=1}^{n} N_{F_j T_j} R_j}$$

Splice junctions with the highest probability were chosen, and long-read splice junctions for which no short-read junctions had a correction probability of at least p_{splice} (default 0.9) were considered uncorrected. Reads that had one or more uncorrected junctions were not considered further for isoform identification. Splice corrected long-read tumor-genome alignments were collapsed into RNA isoform structures by grouping reads with identical splice junctions together if their start loci and end loci were within $\varepsilon_{isoform}$ basepairs (default 10) of each other.

A summary of the application of this algorithm to all long-read splice junctions for the samples considered in this work is provided in (**Supplementary Fig. S2D**). For most samples, more than 90% of long-read splice junctions were within the vicinity (15 bps) of a corresponding short-read junction. In >75% of these cases only a single splice junction was present leading to an unambiguous correction of the long-read splice coordinates to the short-read coordinates. For the remaining junctions the correction algorithm was employed, and in the vast majority (>99%) of cases a short-read splice junction could be confidently selected leading to a correction.

Translation prediction

Known protein coding transcript structures were used to predict the translation start sites of RNA isoforms. ENSEMBL gene annotations were parsed using the *pyensembl* python package (21). These annotations were transposed onto the reconstructed tumor reference genome. For each RNA isoform, the set of most consistent transcript structures was identified by selecting the structures that had the most contiguous matching splice junctions, starting from the most *5'* transcript splice site. If a unique translation start site overlapping the RNA isoform could be identified for this collection of transcript structures, the protein sequence of the RNA isoform was predicted. If more than one translation start site was consistent with the transcript structure, the protein sequence of the isoform was considered ambiguous and a translation prediction was not performed. If the most consistent transcript structure was of a non-coding biotype, the RNA isoform was annotated as non-coding.

NOP identification

Once full-length protein isoforms arising from RNA aligned to the reconstructed reference genome were identified, the tumor-specific portions of each peptide were annotated as NOPs. Each amino acid of each protein coding isoform was annotated as *novel* or *WT* based on the following set of criteria, and strings of consecutive novel amino acids were considered distinct NOPs. For an amino acid to be considered novel it has to:

1. not overlap in-frame with a known WT protein coding isoform
2. be a part of at least one 8, 9, 10, or 11mer amino-acid sequence that was not in the set of known WT peptides
3. arise from a position in the RNA isoform that was downstream of the first potentially causal variant position

The first criteria was considered to be satisfied if the first nucleotide of the amino acid's codon did not align to a genomic position that was a known WT P-site. To rapidly check this for each amino acid in each protein isoform, a P-site genome was pre-compiled by annotating each position of each reference chromosome as either not overlapping with any known P-site, overlapping a P-site in the sense strand, overlapping a P-site in the antisense strand, or overlapping in both strands. Pyensembl (21) with ENSEMBL reference version 75 (GRch37) was used to determine the P-site status of each position in the reference genome. This P-site genome was compiled in a coded string format and stored as a fasta file that was loaded for each sample. This format can easily be extended to include other gene references or WT P-sites from other sources such as RiboSeq experiments.

Lack of overlap with a WT P-site indicated that a portion of the genome was being translated that was not known to be translated. However, homology between the novel translated region and other normally translated regions of the genome may exist for a portion of an otherwise novel protein isoform. To avoid considering these portions as part of NOPs, each amino acid had to be a part of at least one k-mer that was not present in the set of known WT peptides. As NOPs represent potentially interesting neoantigen targets, the k-mer sizes corresponding to potential MHC-I epitopes were chosen. A pre-compiled WT k-mer database was compiled by decomposing all peptides in ENSEMBL and RefSeq protein databases into all possible 8-11mers. This set was made unique and stored as a flat file that was loaded as a set for each sample run. For each amino acid in each isoform, all possible 8-11mers that contained the amino acid in that peptide (max 38) were screened against the WT k-mer set. If all of the 8-11mers were contained within the WT set, the amino acid was not considered novel.

To be considered novel, an amino acid had to also arise from a codon that was downstream of the first variant that would potentially be driving tumor-specific translation. For indels, stoploss, and structural variants the amino acid had to be downstream of the first variant spanned by the RNA isoform. For SV NOPs, the donor and acceptor loci of the first SV-spanning splice junction had to be at least 200 kB away in the base reference genome and then less than 200 kB away in the reconstructed tumor reference. For splice NOPs, the amino acid had to be downstream of the first novel splice junction. To avoid considering amino acids novel due to likely un-annotated splice isoforms that were not altered by the underlying somatic variants, the first exon downstream of the first novel splice junction had to contain at least one novel amino acid for any of the amino acids in the peptide isoform to be considered novel. Additionally, amino acids in peptides spanning SVs were not considered novel if they were within the boundary of the anchor gene that was driving translation. NOPs that had no stop codon but were exact prefixes of other NOPs were removed to avoid duplication in NOP isoform counting due to truncated long reads. Of

note, the cell line NOPs were identified by an earlier version of the approach that allowed for chimeric introns under 200kB as long as they were explained by a structural variant and where hidden NOPs and fusion genes could arise from the same parent gene.

EasyFuse comparison

EasyFuse version 1.3.5 was run using default input parameters. The authors of the EasyFuse tool analyzed a normal tissue set (N=136) and observed a high fraction of the *cis*-near fusion genes identified in tumor samples by the tools that underlie their methodology as also appearing in normal tissue samples. This led them to conclude that these events are often not tumor-specific and thus their tool aimed to improve *trans*-like fusion gene calls as a means of producing reliable neoantigen identifications. Accordingly, these *cis*-near fusion genes were discarded from the final EasyFuse fusion gene peptide set to leave only the ostensibly tumor-specific "neo frame" (coding gene onto non-coding gene fusion) and "out of frame" (coding to coding fusion gene with frame mismatch) neoantigen calls.

To compare the long-read and WGS NOP with the EasyFuse fusion gene tumor specificity, the 17,350 healthy short-read RNA tissue samples obtained from the Genotype-Tissue Expression (GTEx) database were processed through STAR (2.7.9a) to extract chimeric and non-chimeric RNA junctions with the following settings:

```
STAR –runThreadN ${THREADS} \
--genomeDir ${STAR_GENOME} \
--readFilesCommand zcat \
--readFilesIn ${R1} ${R2} \
--outFileNamePrefix ${SAMPLE_NAME} \
--outSAMtype None \
--chimSegmentMin 15 \
--chimOutType Junctions \
--alignMatesGapMax 200000 \
--alignIntronMax 200000 \
--twopassMode Basic \
--outSJfilterIntronMaxVsReadN 200000 \
--chimMultimapNmax 20 \
--chimSegmentReadGapMax 3 \
--chimMultimapScoreRange 10
```

Both the GRCh37 and GRCh38 genomes were used as a reference in order to produce junctions that could be cross referenced to junctions arising from EasyFuse (GRCh38) and to this manuscript's NOP identification methodology (GRCh37). For the SV support analysis, EasyFuse RNA junction coordinates were converted from GRCh38 coordinates to GRCh37 coordinates using the *liftover* python package (https://github.com/jeremymcrae/liftover) version 0.4. A small fraction of junctions (2%) was not able to be lifted over, and these were considered unsupported. The SV genome graph approach described above was then used to find a path from the donor to the acceptor splice junction site less than 200 kB in length.

Simulated data generation and analysis

Samples were selected to form the basis of the simulated data sets based on high structural variant content. Artificial target NOP sequences were generated for each sample using the sample's actual observed genomic variants. To generate target indel NOPs, the indels in coding regions were inserted into the overlapping ENSEMBL transcript structures and which were then translated. To generate fusion genes, each gene in the vicinity of an SV was paired with another randomly selected gene and random isoform structures were selected from each gene. The isoform structures were randomly truncated at an exon–intron boundary and fused in the proper orientation. If a path through the samples SVs could be identified that connected the donor truncation site to the acceptor truncation site in less than 200 kB, then the isoform structure was translated. If this produced a NOP, the fusion gene was accepted. Each gene near an SV was given 100 such attempts at forming a fusion gene with a partner, with the first success ending the trials for that gene. Hidden NOP RNA structures were generated in a similar fashion, but the downstream novel structure was randomly generated by creating a transcript with a number of exons taken from the distribution of exon counts of the normal human reference transcriptome with intron distances also taken from the distribution of annotated introns. If a structural breakpoint was within the range of a randomly generated intron, it was crossed with 50% probability. Exon lengths were also selected from the normal exon length distribution, with edges trimmed to converge on canonical GT–AG splice sites to not interfere with RNA mapping approaches that use such heuristics. In total, 911 hidden NOPs, 722 fusion genes, and 162 indel NOPs were simulated.

The expression levels of these NOP RNA structures were then determined by multiplying the anchor-gene expression level as determined by Kallisto for that sample by either 0.1, 1, or 10 with equal probability to represent various levels of mixing of NOP RNA sequences with normal background gene transcripts. These expression levels along with the normal background gene Kallisto quantification were combined with the short- and long-read RNA library sizes sequenced for the given sample to generate a target number of reads for each RNA species. The short reads were simulated with Rsubread 2.12.3 (22) simReads module that capped total simulated reads at 100 million. The short-read quality scores were randomly selected from the LUN029 non-synthetic short-read dataset. The long reads were simulated with Trans-NanoSim (23) using an error model trained on the most recently sequenced sample (LUN029) so as to provide a best view of the error rates expected in the future, but with a read-length and truncation module trained on a sample of more average RNA integrity (LUN022).

The resultant RNA FASTQ files were combined with the non-synthetic genomic variants for NOP identification. The usual long read–guided NOP identification was carried out as described above. The reconstructed tumor genome generated during this approach was then used for the short read–based NOP identification. The short-read assembly-based approach was based on feeding the short-reads mapped to the reconstructed tumor genome into Stringtie 2.2.1 (24) for assembly using the following settings:

```
stringtie -o ${SAMPLE}_stringtie.gtf \
   -t \
```

```
-p ${THREADS} \
-M 1 \
-j 0.0000001 \
-v \
-f 0.0000001 \
-c 0.001 \
-u ${BAM}
```

The input settings were chosen in an attempt to provide higher sensitivity to very lowly expressed isoforms. These isoforms were then translated according exactly as corrected long-read isoforms as described above. For the RNA junction-centric NOP identification, all splice junctions identified by STAR during the mapping of the short reads to the reconstructed tumor genome that spanned an SV and that had a base-genome donor acceptor distance of over 200kB were identified. For the junctions that started and ended at or within the exon boundaries of annotated transcripts, all possible upstream and downstream isoform structures were merged to form fusion gene isoforms. For junctions that started in a protein coding gene exon and ended in an intergenic region, a pileup of the short-read RNA in the downstream region was taken and the downstream hidden NOP exon was taken to extend from the novel junction acceptor site up to the first non-covered base. These SV NOP isoforms were then translated and annotated as novel according to the methods described for the long-read approach.

MHC-binding prediction

Polysolver (25) was used to predict HLA types using WGS data. NetMHCpan4.1 (26) was used to predict MHC-binding using an EL percentile rank score cutoff of 2 for binders.

Self similarity

Self similarity of epitopes was computed as described in (27). As a normal reference the ENSEMBL GRCH38 proteome was used. To generate random epitopes, random strings of 1,000 nucleotides were generated with a GC content of 40.9% to match the bias of the human genome (28). These nucleotide strings were translated and a random 9-mer epitope was selected from the collection of resultant 9-mers. This process was repeated until enough random epitopes were generated to match the number of NOP epitopes.

In-silico vaccine design

To construct patient-specific framome vaccine designs of a given amino-acid length the longest NOPs were chained together with the remainder of the vaccine consisting of a NOP portion. Missense vaccines of a given length were constructed by chaining together 21 amino acid long sequences with the variant amino acid in the center. Any remaining required length consisted of an amino-acid sequence of the required length with the missense mutation in the middle to provide the most potential CD8 epitopes. For both classes the minimum number of amino acids appended was 8. The number of possible CD8

epitopes was calculated as the number of tumor-relevant 8-11mers included within the design, not accounting for MHC-I binding predictions.

Cell culture

MCF-7 [https://www.atcc.org/products/htb-22, delivery date May 18 2021, grown in DMEM (Gibco, 41965039), 10% FCS. Passaged every 2 to 3 days in T75 culture flask 1:5 to 1:10.], A375 [https://www.atcc.org/products/crl-1619, delivery date June 23 2021, grown in DMEM (Gibco, 41965039), 10% FCS. Passaged every 2 to 3 days in T75 culture flask 1:3 to 1:8.] and 786O [https://www.atcc.org/products/crl-1932, delivery date June 29 2021, medium: RPMI 1640 (Gibco, 11875-093), 10% FCS. Passed every 2 to 3 days in T75 culture flask 1:4 to 1:12.] cells were obtained from ATCC. Mycoplasma testing was performed for the cell lines. Original vials were thawed and passaged for about 10 passages before processing for further analysis. Cell line authentication was ensured based on sequencing.

RiboSeq Analysis

Thawed MCF-7, A375 and 786O cells were lysed in 1 mL lysis buffer (final composition: 1X Mammalian Polysome Buffer (100 mM Tris-Cl pH 7,4 (Cat#T2663, Sigma-Aldrich), 750 mM NaCl (Cat#AM9760G, Fisher Scientific), 25 mM $MgCl_2$ (Cat#AM9530G, Fisher Scientific) diluted in nuclease free water (NFW) (Cat#11-05-01-14, IDT)), 1% Triton X-100 (Cat#T878750ML, Sigma-Aldrich), 0.1% Igepal CA-630 (Cat#I8896-50ML, Sigma-Aldrich), 1 mM DTT (Cat#64656310X.5ML, Sigma-Aldrich), 10 U/mL Dnase I (Cat#D9905K, Lucigen), 0.1 mg/mL Cycloheximide (Cat#C48591ML, Sigma-Aldrich) and NFW to final volume) on ice for 10 minutes. Samples were centrifuged at 20,000g for 10 minutes at 4°C for the precipitation of cell debris. Ribosome footprints were obtained by treating 200 ul lysate with Rnase I (Cat#N6901K, Lucigen) and purified using Sephacryl S400 columns (Cat#GE27-5140-01, Sigma-Aldrich). Next, RNA was extracted using TRIzol LS (Cat#10296028, Fisher Scientific) followed by the Direct-zol RNA Microprep kit (Cat#R2062, Zymo Research). Ribosomal RNA (rRNA) was removed using siTOOLs Ribo-Seq riboPOOL Homo Sapiens (RiboPOOL ID 0.52, siTOOLs BioTech) and subsequently ribosome footprints were polyacrylamide gel electrophoresis (PAGE) purified using Novex 15% TBE-Urea gel (Cat#EC68852BOX, Fisher Scientific). 3' linker ligation and Reverse Transcription (RT) reaction (EpiScript Rnase H – Reverse Transcriptase, Cat#ERT12925K, Lucigen) was performed before PAGE purification using Novex 10% TBE-Urea gel (Cat#EC68752BOX, Fisher Scientific). cDNA fragments were circularized using Circligase I (Cat#CL4115K, Lucigen), and amplified with 12 PCR cycles using 2X Phusion HiFi Master Mix (Cat#F531L, Fisher Scientific). After PAGE purification using Novex 8% TBE PAGE gel (Cat#EC62152BOX, Fisher Scientific), libraries were quantified using Qubit dsDNA HS Assay Kit (Cat#Q32854, Fisher Scientific). Last, quality and fragment size were assessed using the High Sensitivity DNA assay (Cat#5067-4626, Agilent Technologies) on the Bioanalyzer 2100 (Agilent Technologies). To perform multiplex sequencing, barcodes (Illumina TruSeq Small RNA Adapters (RPI series)) were used and sequencing pools were made containing an equal amount of each sample. Libraries were

sequenced on Illumina NextSeq2000 (1x50bp) to an average depth of 301 million reads per sample.

RiboSeq data were mapped to human reference GRCh37. Ribosomal P-site offsets were calculated using the *Ribo-SeQC* R package. Long-read Nanopore RNA sequencing was performed on A375, MCF-7, and 786O cells as described above. Short RNA reads for A375, MCF-7, and 786O (SRA accession numbers SRR8616020, SRR8615758 and SRR8615642 respectively) and SV calls (SvABA calls) were obtained from the CCLE (29). SV calls were converted to breakend format. Subsequently all neo-ORFs and corresponding NOPs for this cell line were determined. RiboSeq read mapping locations (P-sites) were intersected with the portions of neo-ORF long-read RNA mappings leading to hidden NOPs. The periodicity of RiboSeq read P-site coverage in these regions was identified using custom scripts available in the repository accompanying this manuscript (https://github.com/bioinformatics-papers/framome)

Protein mass spectrometry analysis

Protein mass spectrometry data were obtained for A375 cells by ProteiQ (www.proteiq.com). In brief, 3 replicate samples of A375 cells were grown in standard media. Cells were lysed in buffer containing 2% SDS and 100mM Tris. Protein was extracted from each sample and 50 µg was used for further analysis. Cysteines were reduced in 10 mM DTT at 45°C following by alkylation in 2 mM IAA for 30 minutes in the dark. Proteins were precipitated by addition of equal volume of ethanol on magnetic beads using a KingFisher Flex (Thermo). Protein pellets were washed 2x with 80% ethanol and 1x with acetonitrile, proteins were resuspended in 50mM ammonium bicarbonate and digested 4 hours at 37 using trypsin in a 1:50 ratio (trypsin:protein). For chymotrypsin digestion, proteins were resuspended in 50 mM ammonium bicarbonate, 10 mM $CaCl_2$ and digested for 1 hour at 25°C using chymotrypsin in a 1:20 ratio (chymotrypsin:protein). 27 peptides were purchased from JPT peptide technologies, resuspended in 20% acetonitrile, 100mM ammonium bicarbonate to a concentration of 1pmol/µl per peptide. The pool was spiked-in to the samples before the MS measurements were obtained. All samples were measured with a combination of a nanoLC ultraperformance liquid chromatography (UPLC) system (Dionex Ultimate 3000 UHPLC) and a Thermo Scientific Orbitrap Fusion Lumos Mass Spectrometer (Thermo Fisher Scientific, Waltham, MA, USA). Peptides were separated on a C18 50 cm analytical column (EasySpray PepMap ES903) with a flowrate of 300 nL/min (buffer A, 2% ACN in HPLC H_2O, 0.1% formic acid; buffer B, 98% ACN, 0.1% formic acid; 60 min run method). MS was run in positive ion mode. For MS data acquisition, the SureQuant (Thermo) approach was used. Survey Run was used to determine intensity of internal standards, to build spectral library and create SureQuant method for custom panel. For the targeted sample runs, MS1 resolution was set to 120,000 and MS2 resolution was set to 7,500; triggering intensity of the precursors was set to the 1% of the maximum intensity of every precursor. Identification of heavy and endogenous peptides was performed in Skyline v21.2 and peptide chromatograms were manually validated.

Immunopeptidomics analysis

The immunopeptidomic analysis was performed by Biognosys, Switzerland. For DDA LC-MS/MS measurements, peptides were injected to an in-house packed reversed phase column on a Thermo Scientific™ EASY-nLC™ 1200 nano-liquid chromatography system connected to a Thermo Scientific™ Orbitrap™ Exploris 480 ™ mass spectrometer operated in positive mode equipped with a Nanospray Flex™ ion source and a FAIMS Pro™ion mobility device (Thermo Scientific™). LC solvents were A: water with 0.1% FA; B: 80% acetonitrile, 0.1% FA in water. The LC gradient was 1 - 30% solvent B followed by a column washing step in 90% B, and a final equilibration step of 1% B in nano flow. MS1 precursor scans were followed with data-dependent MS2 scans and the cycle time was kept constant. Only precursors with charge state 2-6 were isolated for dependent MS2 scans. The mass spectrometric data was analyzed using SpectroMine and the false discovery rate was set to 1%. A human UniProt fasta database (Homo sapiens, 2022-01-07) combined with a list of A375 NOP protein sequences (including upstream WT portion) was used for the search engine, allowing for non-specifically digested peptides (no trypsin cleaved termini) with a length between 7 and 12 amino acids for MHC class I and between 8 and 23 amino acids for MHC class II, max. 2 variable modifications (N-term acetylation, phosphorylation (STY) for MHC class I and carbamidomethylation (C) for MHC class I and II) and requiring methionine oxidation as fixed modification. Downstream data analysis, including peptide length distribution, peptide motifs and protein group sequence coverage, was performed in SpectroMine. Isotopic peak pattern of specific peptides was extracted from the raw data using SpectroMine. Fragment ion annotation of the best MS2 spectrum and correlation of predicted versus measured fragment ion intensities were visualized using SpectroMine.

Tumor suppressor genes

NOPs were classified as arising from tumor suppressor genes if their gene of origin was in the TSGene database (30).

Validation of RNA junctions by PCR

Target template regions were taken as the 200 bps centered on the junction 100 bp upstream and downstream of the novel RNA junction leading to an NOP. PCR primers (**Supplementary Table S5**) for RNA junctions were designed using a python implementation of Primer3 software with default input parameters and a minimum distance from the center of the template of 10 bps. cDNA produced by TeloPrime Full-Length cDNA Amplification kit V2 (Lexogen) was used as template. PCR reactions were performed using Phusion high-fidelity polymerase (Thermo Scientific) according to manufacturer's specifications. PCR reactions (ProFlex PCR System, Applied Biosystems) were performed on cDNA derived from tumor samples with the identified junctions and an unrelated cDNA sample as a negative control. PCR products were analyzed on agarose gel and manually scored.

In vitro immunogenicity assay

The immunogenicity of selected A375 NOP-derived epitopes was determined as described previously using previously published protocol (6), (31) with minor adaptations. PBMC from buffy coats of healthy individuals with at least one HLA-A*02:01 or HLA-A*01:01 allele (Sanquin Bloedvoorziening, Amsterdam, Netherlands) were isolated using Percoll gradient isolation and stored in liquid nitrogen for further use. On the day of the assay, the cells were thawed and plated in U-bottom 96-well plate at 10^5 cells/well in X-VIVO 15 medium (LONZA). The cells were first stimulated with a mixture of GM-CSF (1000UI/mL, Peprotech), IL-4 (500IU/mL, R&D Systems), and Flt-3L (50ng/mL, R&D Systems) for 24h. Next day, peptides (each peptide at $1\mu g$/mL, GeneCust) and adjuvants LPS ($0.1\mu g$/mL, Enzo), R848 ($10\mu M$, Enzo), and IL-1 (10ng/mL, R&D Systems) were added. As a negative control, cells cultured without any peptide were used. As a positive control CEFT mixture of viral peptides was used (1μg/mL, JPT). The day after, IL-2 (10IU/mL, Peprotech), IL-7 (10ng/mL, Peprotech), and IL-15 (10ng/mL, Peprotech) were added to the cell culture in R10 medium (RPMI, LONZA, 10% heat inactivated human serum, Sanquin Bloedvoorziening, Netherlands, HEPES 10mM, penicillin/streptomycin). Medium was refreshed on day 5 with fresh cytokines. On day 7, the cells were refreshed with R10 medium without cytokines. On day 9 the cells were harvested and plated in fresh R10 medium overnight. Next day, for HLA-A*02:01 donors, T2 cells (ATCC) were exogenously loaded with respective peptides or with myelin oligodendrocyte glycoprotein (MOG, negative control) to present epitopes to $CD8^+$ T cells, or the peptides were added directly to the cell culture. Anti-CD28 and anti-CD49d (both 0.5mg/mL, BD Biosciences) were added for co-stimulation. Where depicted, PMA (50ng/mL, Sigma Aldrich) and ionomycin ($1\mu g$/mL, Sigma Aldrich) were added as a positive control. After 1h, Brefeldin A and monensin (5mg/mL and 2mM, respectively, both BioLegend) were added, followed by an additional 5h of incubation. Subsequently, the cells were stained for surface and intracellular markers (CD3 FITC, clone UCHT1, cat#300405; CD8 PerCp Cy5.5, clone RPA-T8, cat#301031; CD4 APC/Fire 750, clone RPA-T4, cat#300559; IFNγ PE, clone B27, cat#506507; TNFα APC, clone Mab11, cat#502912; and IL-2 PE Cγ7, clone MQ1-17H12, cat#500326, all BioLegend). Acquisition of the data was performed on a BD Symphony (BD) and analysis performed using FlowJo software (FlowJo, version V10, https://www.flowjo.com/solutions/flowjo).

In vitro dextramer staining

FRM0469-specific MHC I Dextramer reagent (conjugated to PE) was custom-generated by Immudex. Immunogenicity assay was performed as described above using PBMCs from three healthy individuals (ImmunXperts) with at least one HLA-B*57:01 allele. PBMCs were either primed with peptide of interest FRM0469 or with irrelevant peptide FRM0460 (negative control) on day 2 of the immunogenicity assay. To perform FRM0469-dextramer staining, FRM0469-primed or negative control cells were resuspended in 50 uL FACS buffer (PBS with 3% human albumin [200 g/L; Sanquin], filtered through a 0.2 µm filter to remove particles) per 3 million cells to which 10 µL FRM0469-dextramer reagent was added. FRM0469-dextramer reagent was centrifuged prior to use [3,300 g, 4°C, 5 mins] and only

supernatant was used. Cells were incubated with the FRM0469-dextramer reagent for 30 mins on melting ice and in the dark, after which a cocktail of extracellular surface and viability markers was added (CD3-FITC [clone UCHT1]; CD8-PerCP-Cy5.5 [clone SK1]; Zombie Aqua Fixable viability dye; all BioLegend) and cells were incubated another 30 mins on melting ice and in the dark. After incubation, cells were washed twice with FACS buffer and pelleted via centrifugation for 5 mins, 500 g, break 6, and 4°C. Cells were subsequently fixated with 1X Cyto-Fast Fix/Perm Buffer (BioLegend) for 15 mins on melting ice after which they were washed twice again with FACS buffer and pelleted via centrifugation for 5 mins, 500 g, break 6, and 4C. Cells were resuspended in 600 uL FACS buffer and acquired on the BD FACS Symphony within 24 hours of staining with a maximum of 8,000 events/sec. FRM0469-dextramer-specific $CD8^+$ T cells were gated as described in **Supplementary Fig. S3**. Cells primed with irrelevant peptide FRM0460 were used as a negative control to determine the background staining of FRM0469-dextramer (**Supplementary Fig. S3**).

In vitro MHC binding and tetramer staining

Selected epitopes for the assessment of in vitro binding were synthesized by GeneCust (GeneCust). In vitro binding was performed as described previously (32). Briefly, a conditional HLA class I complex was stabilized through a photolabile peptide, which could be dissociated through UV irradiation. If the cleavage occurred in the presence of another HLA class I peptide, the reaction resulted in net exchange of the cleaved peptide, yielding an HLA class I complex with an epitope of choice. The peptide exchange efficiency was then analyzed using an HLA class I ELISA. The combined technologies allowed the identification of ligands for an HLA class I molecule of interest. HLA–peptide complexes with binding affinity > 40% were then used to prepare fluorescently labeled tetramers for combinatorial coding and phenotyping, as described previously (33).

Data Availability

WGS and long- and short-read RNA sequencing data are accessible via the European Genome Archive (accession number EGAS00001006021). Code used for NOP identification can be furnished upon reasonable request. Code used to produce all figures in this paper can be accessed at https://github.com/bioinformatics-papers/framome. This resource also contains html files with supporting data for all identified NOPs in this work.

Results

Whole-genome and full-length transcriptome sequencing of 61 human cancers

Recent studies have evaluated personalized neoantigen cancer vaccines in early-stage clinical trials with a primary focus on missense neoantigens identified from exome sequencing (34–36). To systematically extend the repertoire of possible neoantigens expressed in tumors, herein, we focused on the identification of NOPs derived from neo-ORFs resulting from *cis* genomic mutations, including genomic rearrangements, indel frameshifts, splice mutations, and stoploss mutations (6–11, 13–16).

As a basis for our analysis, we collected a series of 61 tumor samples from patients with non-small lung cancer, pancreatic ductal adenocarcinoma, head and neck squamous cell carcinoma, colorectal cancer, glioblastoma, and triple-negative breast cancer (**Supplementary Table S1 –** *Sample Overview*). Tumor samples and corresponding normal tissue or blood samples were whole genome sequenced (tumor WGS - 100X) to identify all classes of somatic genetic changes based on an existing and validated analysis pipeline (4). We identified on average 26,287 (208 - 418,406) single-nucleotide variants (SNVs); 1,847 (65 - 24,160) short indels, and 261 (3 - 2,417) SVs per tumor sample (**Supplementary Fig. S4**, **Supplementary Table S2 –** *WGS Overview*).

To characterize the effects of genomic changes on the tumor transcriptome, we performed RNA sequencing using a combination of short-read RNA sequencing and long-read sequencing of mRNA transcripts (**Supplementary Table S3 -** *Short Read RNA Overview,* **Supplementary Table S4 -** *Long Read RNA Overview***)**. We developed a method to extract intact mRNAs from tumor samples, by a cDNA preparation process involving 3'-polyA and 5'-CAP selection. Double-stranded cDNA was sequenced on Nanopore sequencing devices reaching a throughput of 1M-97M RNA sequences per sample. Up to 92.3% of long-read mRNA sequences mapping to a protein coding gene spanned a full transcript molecule known in the ENSEMBL database, indicating the strength of the long-read data to determine complete transcript sequences at the single molecule level (**Supplementary Fig. S5**). Despite the difference in the intended RNA content captured by each sequencing approach (fully mature vs total), the gene expression quantification was reasonably concordant between short and long reads (**Supplementary Fig. S6A**). Additionally, due to the minority of non-full length long-read RNA reads the *5'* and *3'* coverage biases were also similar in magnitude to the edge-effects of ribo-depleted short read RNA sequencing (**Supplementary Fig. S6B**).

Identification of expressed neo-ORFs

Identification of possible neoantigens from sequencing data is often limited to the detection of coding mutations, followed by analysis of their expression using short-read RNA sequencing (2). The neoantigenic peptide sequence is subsequently inferred from known annotated transcript structures. However, a preferred method would be to directly determine peptide sequences based on the repertoire of expressed transcript isoforms in the tumor. We leveraged the WGS and short- and long-read RNA sequencing data as input for an analysis workflow that maps complete tumor-specific transcript sequences caused by *cis* somatic mutations, including SVs, indels, and SNVs within and outside coding regions.

The analysis approach consists of four steps that integrate somatic mutation data with transcriptome sequences to identify all neo-ORFs and corresponding NOPs (**Fig. 1**). In a first step, the collection of somatic small and structural variants is combined with chimeric long-and short-read RNA mappings to construct tumor-specific contigs that together create a tumor-specific reference for each sample. In a subsequent step, short-read and long-read RNA sequences are mapped to the tumor-specific reference to identify transcripts in the vicinity of a somatic mutation. In this step, the short-read RNA sequences are used to correct (splice-junction) errors inherent to long-read single-molecule Nanopore sequencing

data. Subsequently, individual corrected transcript reads are used for *in silico* translation based on annotated translation start sites (Ensembl) to derive entire protein sequences. In a final step, NOPs (or neo-epitopes from the NOPs) are derived from the protein sequences by trimming of the WT portions of each protein sequence.

This analysis approach is what we believe to be the first to internally integrate full-length sample-specific transcript structures with variant protein effect prediction as well as what we believe to be the first method to directly couple WGS with long-read transcriptome sequencing for the discovery and validation of SV-driven tumor specific isoforms. We used the workflow to analyze neo-ORFs and corresponding NOPs caused by SVs, frame-shift indels, splice mutations, and stoploss mutations (**Supplementary Table S6 -** *NOP Annotations* and https://github.com/bioinformaticspapers/framome). PCR across identified tumor-specific RNA junctions confirmed 78% of junctions, supporting the validity of our approach (**Supplementary Table S5 -** *PCR Validation*). An example of each class of NOP is provided in **Fig. 2**. A known category of SV-driven NOPs is a fusion gene event (**Fig. 2A)**. While the canonical source of NOPs arising from fusion genes is through a mismatch in reading frame between the upstream and downstream genes, the example depicts translation initiation in the *CAMSAP1* gene, which is predicted to lead to a 27 amino acid NOP partially overlapping with the *5'*-UTR of the *URM1* gene. An example of a NOP derived from a canonical exonic indel frame-shift is depicted in **Fig. 2B**, which displays a 57 amino acid NOP in the *TP53* tumor suppressor gene in non-small cell lung cancer (NSCLC) sample LUN011. Frame-shift derived NOPs are commonly found in tumor suppressor genes (24 genes across 26 samples), which form a source of shared antigens (7). We also identified NOPs caused by either mutations affecting known splice sites or mutations introducing new splice sites (**Fig. 2C**), as well as NOPs derived from mutations in known stop codons (**Fig. 2D**).

Identification of hidden NOPs

Gene fusions represent a frequent outcome of somatic SVs in cancer genomes and in-frame gene fusions can be drivers of tumorigenesis (37). However, the majority of gene fusions are out-of-frame, creating a novel gene encoding a NOP (9) (**Fig. 2A**). We observed between 0 and 76 unique expressed NOPs corresponding to out-of-frame gene fusions per tumor sample, representing a substantial class of potentially neoantigenic sequences, particularly in tumors with high SV loads. An additional, yet largely uncharacterized configuration of genomic rearrangements involves the fusion between the *5'* part of a known gene and a non-coding genomic region, i.e. where no gene annotation or evidence for transcription exists. Based on *in silico* annotation of SV breakpoint junctions in the tumor samples, we observed that 3,002 (19%) of somatic SV breakpoint junctions involved a 5'-part of a known gene fused to a non-coding genomic region downstream of the SV breakpoint junction (**Fig. 3A**). Next, we analyzed the RNA sequences overlapping the SV breakpoint junctions in such regions and we observed that for 8% (245) of such SV junctions involving the 5'-end of a known gene and a non-coding region, breakpoint junction–spanning transcripts were identified aligning with the 5'-end of a known gene and with a 3'-end that involved one or more novel cryptic exons. We have termed these chimeric transcripts and their resulting tumor-specific peptide products 'hidden NOPs'. An example of a hidden NOP identified in NSCLC sample LUN004 is depicted in **Fig. 3B**, which shows the fusion of the 5'

exons of gene *TIMM8B*, located on chromosome Chr11, coupled to cryptic exons encoded by a genomic region on Chr2. The novel chimeric transcript was confirmed by 21 long and corrected transcript reads.

To understand the translation of tumor-specific chimeric transcripts that encode hidden NOPs, we performed analysis of WGS and transcriptome sequencing data of human cell lines A375 (melanoma), MCF7 (breast adenocarcinoma), and 786O (renal cell adenocarcinoma) resulting in the identification of 7, 51, and 6 hidden NOPs, respectively. We complemented the analysis with ribosome profiling (RiboSeq), which enables the identification of translation of mRNA sequences, including the reading frame in which an mRNA is translated (38). An example of a highly-expressed hidden NOP identified in this way is depicted in (**Fig. 3C**); it shows RiboSeq reads in the expected reading frame across the novel exons triggered by a set of genomic SVs in MCF7. For the majority of the hidden NOPs in these three cell lines, RiboSeq coverage was observed for the expressed novel exons derived from intergenic (non-coding) genomic regions, with the majority of the RiboSeq reads indicating the expected reading frame that was inferred from the translation start site of the partner gene (**Fig. 3D**). To complement these results and further demonstrate that hidden NOPs lead to proteins expressed by cancer cells, we performed protein mass spectrometry analysis of intracellular proteins of A375 cells. Using spike-in of 27 heavy labeled peptides corresponding to 11 different NOP sequences, we observed the presence of 5 unique peptides corresponding to both hidden NOPs and fusion gene NOPs (**Supplementary Fig. S7**, **Supplementary Table S7 -** *A375 MS Peptides*). Three of these identified peptides arose from three different hidden NOP protein isoforms anchored in the *AHCY* gene (**Fig. 3E**), further supporting the presence of NOP protein in A375 cells.

Whole-genome and full-length transcriptome sequencing improves quality of NOP identification

Existing methods for identifying neoantigens arising from SVs are not guided by full-length transcript sequences or knowledge of how the tumor genome has been rearranged. Additionally, most available tools are canonical fusion-gene callers that focus on identifying aberrant RNA structures bridging two annotated genes while ignoring gene-intergenic hidden NOPs. To compare the structural variant NOPs identified via full-length transcriptome sequencing and WGS with a state-of-the-art fusion gene identification tool, EasyFuse (39) was run on 39 samples (**Supplementary Table S8 -** *EasyFuse Results*).

The overlap in the potential neoantigens identified by EasyFuse and the approach presented here was quantified by breaking down the full-length peptides into non-reference protein amino acid 9mers in order to allow for partial protein isoform overlap. Only frame-shift and neo-frame non *cis*-near EasyFuse fusion genes were considered. EasyFuse and the long-read RNA and WGS approach identified 15,613 and 12,968 9mer-sample pairs, respectively. The overlap between the peptide sets derived from both approaches was low (1,424 (5.2%) sample-9mer pairs). This overlap is on-par with the overlap observed between the individual fusion gene callers underlying EasyFuse (39).

While it is recognized that there is a lack of tumor specificity of *cis*-near fusion genes identified by short-read RNA sequencing using EasyFuse, the size and scope of the normal-tissue reference was limited in that study (N=136) (39). To gain a better understanding of

the tumor specificity of short-read only fusion gene calls as compared to NOPs identified through a combination of long-read RNA and WGS, we searched for evidence of the identified RNA structures underlying these neoantigen calls in the 17,350 healthy tissue samples in the GenotypeTissue Expression (GTEx) database (19). To do this, we remapped all GTEx short-read RNA samples to the human reference genome and extracted the chimeric and non-chimeric RNA junctions identified by the STAR aligner. We then cross-referenced these exact junction coordinates with the breakpoint coordinates (junctions) of the EasyFuse fusion gene calls and of the NOPs identified via long-read RNA and WGS (**Supplementary Fig. S8A**). This showed that 22% (229 out of 1054) of the EasyFuse junctions were present in the GTEx normal tissue samples at support levels of up to thousands of junction-spanning reads per sample while only 0.9% (5 out of 543) of the long-read RNA and WGS NOP junctions were present (**Supplementary Fig. S8B**, **Supplementary Table S9 -** *GTEx Junction Overlap*). To quantify how wide-spread such potential false positives were in the neoantigen calls generated by each approach, the amino-acid sample-9mer pairs were classified according to whether they arose from a protein with a defining novel RNA junction present in GTEx. The results shown in **Supplementary Fig. S8C** illustrate that 44.3% of potential neoantigen sequence content identified by EasyFuse arose from breakpoint RNA junctions that can also be found in healthy tissues while only 1.4% of long-read RNA and WGS identified NOP sample-9mers fell into this category. Additionally, sample-9mers were classified by whether a path from the donor to the acceptor site of the defining RNA junction could be found through SVs. While an SV event can be found that supports all SV-NOP protein sequences identified in our methodology, only 33% of the EasyFuse sample-9mer pairs were supported by WGS. Overall, the combination of WGS with full-length transcriptome sequencing appears to dramatically increase the tumor specificity of identified SV neoantigens.

To understand the effect of full-length transcriptome sequencing on NOP identification and partially separate its contribution from that of WGS, simulated short- and long-read RNA sequencing datasets containing *a priori* designed ground-truth NOPs were generated. NOPs were identified from the simulated data using the described combined long- and short-read methodology as well as by two short-read only approaches. For the first approach, short-reads remapped to the reconstructed tumor genome were assembled and the resulting RNA isoforms were translated into NOPs using the same technique as for the long-reads. Secondly, novel RNA junctions spanning structural variant breakpoints in the reconstructed tumor genome were identified and the long-range upstream and downstream structures were assumed from annotations and coverage (similar to the method employed by traditional fusion gene callers). The results of each approach were then compared to the ground-truth via amino acid 9mer overlap. The results presented in **Supplementary Fig. S9A** illustrate that the overall sensitivity of the long-read guided approach is essentially the same as for the short-read junction translation, with both methods reaching a high recall of ~70% of sample-9mers. The short-read assembly approach is less sensitive, possibly due to an inability of the assembler to detect minority NOP isoforms mixed with obfuscating normal gene expression. The short-read junction approach identified more false-positive neoantigens than the long-read guided approach

(~10% vs ~1% FDR). This difference is especially true for the fusion gene NOPs (**Supplementary Fig. S9B**). Comparing the target expression level of ground-truth NOPs versus the recall rate, in **Supplementary Fig. S9C** all methods perform poorly on lowly expressed (<1 TPM) sequences but that the sensitivity is essentially unchanged beyond this level of expression. Additionally, filtering called NOPs based on estimated expression level can reduce FDR but at a high cost to sensitivity. These results illustrate the power of long-read RNA sequencing to achieve high sensitivity while avoiding erroneous neoantigen calls resulting from the transcript structure ambiguity inherent to short-read RNA sequencing.

Many tumors have large framomes

We identified 998 unique NOPs amongst the 61 tumor samples described in this work (**Supplementary Table S6 –** *NOP Annotations*), and we classified the NOPs according to their genomic origin (**Fig. 2**, **Fig. 3A**). On average we identified 16 NOPs with a combined length of 383 amino acids per tumor sample (**Fig. 4A**). We found that hidden NOPs were the major source of NOPs with 49% of novel amino acid sequence arising from this source. Fusion genes and indels contributed 37% and 12% respectively, with the remaining 2% made up of the infrequent stoploss and splice NOPs. Taken together, hidden NOPs and fusion genes made up 86% of all NOPs identified, highlighting the importance of SVs as a source of neoantigens. Different cancer types express NOP classes at different frequencies. We observed that glioblastomas often express hidden NOPs and gene-fusion NOPs (97% of novel amino acids), as a result of the high load of somatic SVs and a low number of exonic indels. For NSCLC, we found a higher amount of indel-derived NOPs (20% of novel amino acids) than average, which reflects the relative amount of frame-shift indels and SVs in this cancer type.

We have termed the entire collection of NOPs expressed by a specific tumor sample 'the framome'. Representative examples of tumor framomes are given in **Fig. 4B,C**. Glioblastoma sample GBM005 expressed 86 unique NOPs, for a total of 1,806 amino acids, almost all of which are derived from somatic SVs. In contrast, the framome of NSCLC sample LUN013 represented 1,196 amino acids across 49 NOPs, many of which resulted from canonical frame-shift indels.

Expression level and clonality are important features for selection of neoantigens as immunotherapy targets (40, 41). The expression levels of mRNAs encoding NOPs were measured based on the long-read RNA sequencing data generated for each tumor sample and quantified as transcripts per million (TPM) where the denominator of the calculation was taken as the total number of full-length translatable long-reads (**Fig. 4D**). Expression levels of NOPs largely fell into the distribution of the expression levels observed for other genes (**Supplementary Fig. S10A**). A slight shift in the average expression of NOPs compared to missense variants was observed for some tumor samples, which is likely an effect of nonsense-mediated decay (**Supplementary Fig. S10B**) (42). The variant allele frequency of the underlying somatic genetic changes was determined from the whole genome sequencing data and corrected for tumor purity **Fig. 4D** (4). We observed that variants encoding NOPs have an average purity corrected VAF of 0.37, similar to other somatic variants in the analyzed tumor genomes (**Supplementary Fig. S10C**). Of note, SVs underlying hidden NOPs and fusion NOPs have a lower average VAF, likely representing

technical difficulties in accurately estimating the number of discordant read-pairs supporting such variants.

Complex chromosomal rearrangements, such as chromothripsis, are a frequent phenomenon in cancer genomes (43). For 18% of the genomic events leading to hidden NOPs or out-of-frame gene fusions, the genomic connection between the *5'*-end of the known gene and the non-coding genomic segment or downstream out-of-frame gene was formed by more than one genomic breakpoint-junction (**Fig. 4E**, **Supplementary Fig. S11**). The analysis of single full-length transcript molecules using our approach enabled us to identify the entire spectrum of transcript isoforms encoding a hidden NOP. The majority (60%) of SVs leading to hidden NOPs involved transcripts that encoded a single unique NOP. However, we observed multiple instances of hidden NOPs that were caused by different transcript isoforms derived from the same genomic SV. For example, a hidden NOP in a triple-negative breast tumor involved multiple splice isoforms encoding 4 different unique NOP sequences (**Supplementary Fig. S12**). Isoform diversity may thus enlarge the neoantigenic potential of hidden NOPs.

To understand the amount of possible HLA-binding epitopes among NOPs expressed in tumors, we performed *in silico* characterization of HLA class I binding. While the overall percentage of candidate epitopes predicted to bind was similar between NOPs and missense mutations (2.20% and 2.28%, respectively), **Supplementary Fig. S13A** illustrates striking tumor-type specific differences in the absolute number of predicted binders arising from the two neoantigen sources. In most glioblastoma, pancreas, and triple-negative breast tumors, NOPs represented a larger source of binding epitopes than missense mutations, while for NSCLC, head and neck squamous cell carcinoma, and colon tumors, the number of binders arising from the two classes was more even except for several outliers with high missense mutation loads. These results indicate NOPs may increase the immunotherapy target landscape, especially in indications such as glioblastoma which have traditionally been considered to have a low neoantigen burden.

To understand whether a cancer vaccine based on NOPs would be advantageous with respect to the number of possible MHC class I epitopes, as compared to vaccines based on commonly used missense variants, we generated cancer vaccine designs in silico. In **Supplementary Fig. S13B** a comparison is made between the number of potential MHC class-I sized kmers (k=8 through 11)for the two classes of antigens in the context of a neoantigen-based personalized therapeutic cancer vaccine designed *in silico* for each of the tumors reported in this study. This analysis shows that for many tumors an ~2 fold increase in targeted sequence can be achieved through the use of NOPs compared to missense variants. Targeting NOPs may also allow for a superior quality of each epitope as there is increasing evidence that neoantigen dissimilarity to self proteins is important for effective immune response (27). The long out-of-frame peptide sequences represented by NOPs are, in principle, fully tumor-specific and the same sequences should not be expressed in normal (non-tumor) cells. We determined the similarity to self for all 9-mers derived from NOPs and missense variants expressed by each of the 61 tumors analyzed in our study (**Supplementary Fig. S13C**). This demonstrated that NOP epitopes were nearly as dissimilar from self as completely random epitopes (mean 0.7 vs 0.74), while missense

epitopes which differ from wild-type epitopes by only a single point mutation were highly self-similar (mean 0.86).

Framome-derived epitopes are presented on tumor cells and can induce polyfunctional CD8+ T-cell responses in vitro

To validate the presence of HLA-binding epitopes derived from NOPs, we performed immunopeptidomics analysis of A375 cells. We identified 10,537 peptides eluted from A375 MHC class I molecules, matching to 4,824 protein groups. For 2 peptides, a match was found to two hidden NOPs, each representing a different transcript isoform encoded by the same SV (**Fig. 5A** and **Supplementary Fig. S14**), demonstrating presentation of hidden NOP–derived epitopes on cancer cells. For one of the identified peptides, FRM0469 with high predicted affinity to HLA-B*57:01, we generated HLA-B*57:01 fluorescently labeled dextramers that we used to stain PBMCs isolated from three individual donors carrying this allele. We detected a low frequency of antigen-specific CD8+ T cells in two out of three donors at baseline and we observed their expansion upon a 10-day culture in the presence of FRM0469 peptide (**Fig. 5B** with gating strategy in **Supplementary Fig. S3**).

To further complement these findings, the immunogenicity of framome-derived antigens was assessed by *in vitro* immunogenicity assays using PBMCs of healthy donors with at least one HLA-A*02:01 or HLA-A*01:01 allele. This assay is suitable to rapidly prime naive T cells (adapted from (6)). The cells were primed with CD8+ T-cell epitopes derived from NOPs identified in melanoma cell line A375 and predicted to bind to HLA-A*02:01 or HLA-A*01:01 using NetMHCpan (**Supplementary Table S10 -** *A375 Tested Epitopes*). After the expansion phase the cells were re-stimulated either with relevant peptides or a negative control (MOG). Antigen-specific CD8+ T-cell responses were measured as IFNɣ and TNFα production using intracellular cytokine staining and detected by flow cytometry. A total of 8 HLA-A*02:01 peptides and 4 HLA-A*01:01 peptides were tested leading to reactivity in three out of ten donors (**Fig. 5C**). In total, NOP-specific CD8+ T cells were identified in five out of 13 tested donors (**Fig. 5D**). This confirms that NOPs can contain immunogenic epitopes with the potential to induce (polyfunctional) antigen-specific CD8+ T-cell responses.

NOP epitopes bind to MHC and are recognized by memory CD8+ T cells of an NSCLC patient

To further characterize the immunogenic properties of NOPs, we assessed the affinity of framome-derived epitopes to various HLA-A and -B alleles by performing *in vitro* HLA-binding assays. First, we selected ~30 epitopes derived from the framomes of each of three patients with advanced NSCLC (LUN024, LUN026, and LUN029) and we tested the binding of these epitopes using *in vitro* binding assays. As the framomes of patients LUN026 and LUN029 provided many predicted epitopes, we limited our selection to those with the highest predicted affinity (EL rank score below 2) for each relevant HLA allele (**Supplementary Table S11 -** *Lung Sample NetMHCpan*). *In vitro* binding analysis revealed a number of epitopes binding to HLA-A and HLA-B specific for each patient (**Fig.6A**). In the next step, we generated fluorescently labeled HLA tetramers carrying epitopes with at least

40% binding affinity, as determined by the *in vitro* binding assays (**Supplementary Table S12 -** *In Vitro Binding*). These tetramers, specific to each patient's HLA allele, along with relevant positive and negative controls, were then used to stain low frequency antigen-specific $CD8^+$ T cells within the PBMC population of each patient's peripheral blood using combinatorial coding. Additionally, antigen-specific $CD8^+$ T cells recognizing individual epitopes were phenotyped to determine their antigen experience status (for the gating strategy see **Supplementary Fig.S15**). We detected the presence of effector memory type ($CD8^+CD45RA^-CD27^{-/dim}$) T cells in the blood of patient LUN029, specific for two epitopes, FRM0417 and FRM0433 (**Fig.6B)**. Each of the epitopes originated from a different hidden NOP (**Fig.6C)**. No $CD8^+$ T cells could be detected by staining with any of the other tetramers carrying different NOP-derived peptides from the three lung cancer patients. Finally, we used the FRM0417 and FRM0433 peptides to expand $CD8^+$ T cells from two healthy donors. We observed expansion of $CD8^+$ T-cell populations upon priming with FRM0417 and FRMA0433 peptides in both healthy donors (**Supplementary Fig.S16**). These data confirm that some of the antigens derived from hidden NOPs can bind to various HLA alleles and can induce antigen specific immune responses in vitro using PBMCs from healthy donors and can be detected in a patient with cancer.

Discussion

The work described here provides a technological and bioinformatics framework to exploit the full potential of NOPs encoded in the tumor genome as a result of *cis*-acting somatic mutations. Identification of the full spectrum of expressed NOPs in tumors requires WGS as a basis complemented with RNA sequencing to map mutated transcripts. Only WGS captures the complete catalogue of somatic mutations arising in cancer genomes (4). Although commonly used as an efficient technology for detection of exonic mutations (e.g., frameshift indels), exome sequencing falls short with respect to identification of intronic and intragenic variants and SVs (e.g. splicesite creating mutations) (14). The approach outlined in our manuscript will become more feasible with the decreasing costs and efforts needed for WGS and long-read RNA sequencing. Internally, we are capable of generating good quality genome and transcriptome sequencing data within 10 days of obtaining a thin needle biopsy of a tumor. This makes the approach potentially applicable for routine clinical practice, e.g. in the context of personal cancer vaccine trials. Application of our approach to FFPE tumor specimens would likely results in suboptimal detection of NOPs, because of lower RNA integrity.

Our work demonstrates that SVs provide a rich source of possible cancer neoantigens, beyond well-described neoantigenic sequences derived from fusion genes (10). We find that SVs often drive expression of non-coding genomic regions via fusion with the *3'*-end of a known gene. We designate these as hidden NOPs as their existence cannot be identified from genome sequencing alone, but requires the integrated analysis of cancer transcripts sequences with somatic SVs. We observed that ~50% of the amino acid sequences contributed by NOPs are derived from hidden NOPs caused by SVs. Personalized neoantigen based immunotherapy strategies targeting tumors with a high level of SVs (e.g., glioblastoma, triple-negative breast), or with both high SV and high indel count (e.g. NSCLC)

would benefit from a neoantigen discovery approach as outlined here. We propose that a complete analysis of the cancer genome will enable optimal design of personalized cancer vaccines.

In addition to genomic analysis of the tumor, faithful mapping of mutation-derived transcripts encoding possible neoantigens is required to precisely determine tumor-specific peptide sequences. The wide diversity of transcript isoforms encoded by the human genome has become apparent through full-length transcript sequencing (44). Direct mapping of the isoforms of a gene would be a preferred approach to infer neoantigenic peptide sequences, rather than the commonly used approach to use existing transcript annotations. The combined approach of whole genome and long-read transcriptome sequencing enables analysis of neoantigenic sequences derived from individual transcript sequences based on the identification of translation start sites and accurate transcript structure and sequence. Our current approach involves the use of short-read RNA sequencing to refine transcript splice-junction sequencing, but we expect that future generations of long-read sequencing will make such an approach obsolete.

In conclusion, we here present a universally applicable workflow that enables systematic identification of neo-ORFs and corresponding NOPs resulting from somatic cancer mutations. Our work provides the foundation for developing new generations of personalized immunotherapies.

Acknowledgements

The authors would like to sincerely thank Hugo Olsman, Lisa van der Made, and Finn Edwards for their contribution and expertise in performing the immunological assays described in this work.

References

[1] Richters MM, Xia H, Campbell KM, Gillanders WE, Gri th OL, Gri th M. Best practices for bioinformatic characterization of neoantigens for clinical utility. Genome medicine. 2019;11(1):1-21.

[2] Shemesh CS, Hsu JC, Hosseini I, Shen BQ, Rotte A, Twomey P, et al. Personalized cancer vaccines: clinical landscape, challenges, and opportunities. Molecular Therapy. 2021;29(2):555-70.

[3] Garcia-Garijo A, Fajardo CA, Gros A. Determinants for neoantigen identification. Frontiers in immunology. 2019;10:1392.

[4] Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature. 2019;575(7781):210-6.

[5] Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, et al. Insertion-and-deletionderived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. The lancet oncology. 2017;18(8):1009-21.

[6] Roudko V, Bozkus CC, Orfanelli T, McClain CB, Carr C, O'Donnell T, et al. Shared immunogenic poly-epitope frameshift mutations in microsatellite unstable tumors. Cell. 2020;183(6):1634-49.

[7] Koster J, Plasterk RH. A library of Neo Open Reading Frame peptides (NOPs) as a sustainable resource of common neoantigens in up to 50% of cancer patients. Scientific reports. 2019;9(1):1-8.

[8] Rathe SK, Popescu FE, Johnson JE, Watson AL, Marko TA, Moriarity BS, et al. Identification of candidate neoantigens produced by fusion transcripts in human osteosarcomas. Scientific reports. 2019;9(1):111.

[9] Fotakis G, Rieder D, Haider M, Trajanoski Z, Finotello F. NeoFuse: predicting fusion neoantigens from RNA sequencing data. Bioinformatics. 2020;36(7):2260-1.

[10] Yang W, Lee KW, Srivastava RM, Kuo F, Krishna C, Chowell D, et al. Immunogenic neoantigens derived from gene fusions stimulate T cell responses. Nature medicine. 2019;25(5):767-75.

[11] Mansfield AS, Peikert T, Smadbeck JB, Udell JB, Garcia-Rivera E, Elsbernd L, et al. Neoantigenic potential of complex chromosomal rearrangements in mesothelioma. Journal of Thoracic Oncology. 2019;14(2):276-87.

[12] Kosari F, Disselhorst M, Yin J, Peikert T, Udell J, Johnson S, et al. Tumor Junction Burden and Antigen Presentation as Predictors of Survival in Mesothelioma Treated With Immune Checkpoint Inhibitors. Journal of Thoracic Oncology. 2021.

[13] Jung H, Lee KS, Choi JK. Comprehensive characterisation of intronic mis-splicing mutations in human cancers. Oncogene. 2021;40(7):1347-61.

[14] Jayasinghe RG, Cao S, Gao Q, Wendl MC, Vo NS, Reynolds SM, et al. Systematic analysis of splicesite-creating mutations in cancer. Cell reports. 2018;23(1):270-81.

[15] Shiraishi Y, Kataoka K, Chiba K, Okada A, Kogure Y, Tanaka H, et al. A comprehensive characterization of cis-acting splicing-associated variants in human cancer. Genome research. 2018;28(8):1111-25.

[16] Dhamija S, Yang CM, Seiler J, Myacheva K, Caudron-Herger M, Wieland A, et al. A pan-cancer analysis reveals nonstop extension mutations causing SMAD4 tumour suppressor degradation. Nature cell biology. 2020;22(8):999-1010.

[17] Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nature biotechnology. 2017;35(4):316-9.

[18] Hagberg A, Swart P, S Chult D. Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Lab.(LANL), Los Alamos, NM (United States); 2008.

[19] Consortium G, Ardlie KG, Deluca DS, Segr`e AV, Sullivan TJ, Young TR, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348(6235):64860.

[20] Leonardo CT, Abhinav N, Kai K. Ellis Shannon E, Taub Margaret A, Hansen Kasper D, Jaffe Andrew E, Langmead Ben, Leek Jeffrey T. Reproducible RNA-seq analysis using recount2. Nature Biotechnology. 2017;35(4):319-21.

[21] Rubinsteyn A, Kodysh J, Hodes I, Mondet S, Aksoy BA, Finnigan JP, et al. Computational pipeline for the PGV-001 neoantigen vaccine trial. Frontiers in immunology. 2018;8:1807.

[22] Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic acids research. 2019;47(8):e47-7.

[23] Hafezqorani S, Yang C, Lo T, Nip KM, Warren RL, Birol I. Trans-NanoSim characterizes and simulates nanopore RNA-sequencing data. Gigascience. 2020;9(6):giaa061.

[24] Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology. 2015;33(3):290-5.

[25] Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, Lawrence MS, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nature biotechnology. 2015;33(11):1152-8.

[26] Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic acids research. 2020;48(W1):W449-54.

[27] Wood MA, Paralkar M, Paralkar MP, Nguyen A, Struck AJ, Ellrott K, et al. Population-level distribution and putative immunogenicity of cancer neoepitopes. BMC cancer. 2018;18(1):1-15.

[28] Piovesan A, Pelleri MC, Antonaros F, Strippoli P, Caracausi M, Vitale L. On the length, weight and GC content of the human genome. BMC research notes. 2019;12(1):1-7.

[29] Ghandi M, Huang FW, Jan´e-Valbuena J, Kryukov GV, Lo CC, McDonald ER, et al. Next-generation characterization of the cancer cell line encyclopedia. Nature. 2019;569(7757):503-8.

[30] Zhao M, Kim P, Mitra R, Zhao J, Zhao Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic acids research. 2016;44(D1):D1023-31.

[31] Bozkus CC, Blazquez AB, Enokida T, Bhardwaj N. A T-cell-based immunogenicity protocol for evaluating human antigen-specific responses. STAR protocols. 2021;2(3):100758.

[32] Rodenko B, Toebes M, Hadrup SR, Van Esch WJ, Molenaar AM, Schumacher TN, et al. Generation of peptide–MHC class I complexes through UV-mediated ligand exchange. Nature protocols. 2006;1(3):1120-32.

[33] Hadrup SR, Bakker AH, Shu CJ, Andersen RS, Van Veluw J, Hombrink P, et al. Parallel detection of antigen-specific T-cell responses by multidimensional encoding of MHC multimers. Nature methods. 2009;6(7):520-6.

[34] Ott PA, Hu-Lieskovan S, Chmielowski B, Govindan R, Naing A, Bhardwaj N, et al. A phase Ib trial of personalized neoantigen therapy plus anti-PD-1 in patients with advanced melanoma, non-small cell lung cancer, or bladder cancer. Cell. 2020;183(2):347-62.

[35] Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature. 2017;547(7662):217-21.

[36] Sahin U, Derhovanessian E, Miller M, Kloke BP, Simon P, L¨ower M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature. 2017;547(7662):222-6.

[37] Mertens F, Johansson B, Fioretos T, Mitelman F. The emerging complexity of gene fusions in cancer. Nature Reviews Cancer. 2015;15(6):371-81.

[38] Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell. 2011;147(4):789-802.

[39] Weber D, Ibn-Salem J, Sorn P, Suchan M, Holtstr¨ater C, Lahrmann U, et al. Accurate detection of tumor-specific gene fusions reveals strongly immunogenic personal neo-antigens. Nature Biotechnology. 2022:1-9.

[40] McGranahan N, Furness AJ, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science. 2016;351(6280):1463-9.

[41] Westcott PM, Sacks NJ, Schenkel JM, Ely ZA, Smith O, Hauck H, et al. Low neoantigen expression and poor T-cell priming underlie early immune escape in colorectal cancer. Nature cancer. 2021;2(10):107185.

[42] Litchfield K, Reading JL, Lim EL, Xu H, Liu P, Al-Bakir M, et al. Escape from nonsense-mediated decay associates with anti-tumor immunogenicity. Nature communications. 2020;11(1):1-11.

[43] Cort´es-Ciriano I, Lee JJK, Xi R, Jain D, Jung YL, Yang L, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nature genetics. 2020;52(3):331-41.

[44] De Paoli-Iseppi R, Gleeson J, Clark MB. Isoform Age-Splice Isoform Profiling Using Long-Read Technologies. Frontiers in Molecular Biosciences. 2021;8.

Figure Legends

Figure 1: Overview of tumor NOP identification workflow. Tumor specific variants are identified from tumor/normal WGS and used in combination with short- and long-read RNA sequencing to reconstruct the tumor genome. RNA is remapped to this tumor specific reference to produce translatable full-length isoforms and a database of WT peptide k-mers and P-sites are used to identify which portions of these predicted peptides are novel. These NOPs are extracted to produce the framome.

Figure 2: Examples of NOP categories identified by our workflow. Reconstructed tumor contigs are shown as thick purple/green lines. Annotation isoforms from ENSEMBL are shown below the contigs. Full-length isoforms created through correction/collapsing of long-reads are shown above the contigs. The known/predicted protein coding structure of each isoform is provided with green for *5'*-UTRs, brown for WT coding, red for NOP, multi-colored for zoomed-in NOP amino acids, and blue for *3'*-UTRs. Non-coding isoforms are shown in grey. (A) An 8 Mb inversion within chromosome 9 leads to a fusion gene between the *CAMSAP1* and *URM1* genes in the glioblastoma sample GBM002. Beginning translation at the *CAMSAP1* start site gives an NOP partially overlapping the *5'*-UTR of *URM1*. (B) A basepair deletion in an exon of the *TP53* gene in lung sample LUN011 leads to out-of-frame translation of a portion of an exon. The 57 amino acid NOP represents a truncation of the normal protein isoform. (C) A point mutation in the head and neck tumor HAN001 leads to a splicing signal in the intron of the *MLLT10* gene. This splicing leads to a partial 3' intron retention and drives translation of a 10 amino acid NOP. (D) A point mutation within the stop codon of the *CHCHD6* gene in the head-and-neck sample HAN002 leads to a translation elongation and a 15 amino acid NOP.

Figure 3: Hidden NOPs are a frequent result of genomic structural variants. (A) Schematic outline of the origin of hidden NOPs. A somatic genomic breakpoint junction involving the 5'-end of a protein coding gene is fused to a non-coding genomic region. Transcription is driven by the promoter of the 5'-gene and continues across the structural variant breakpoint. The resulting transcript is spliced leading to a novel open reading frame encoding a tumor-specific NOP. (B) Example of a hidden NOP identified in LUN004, involving the *TIMM8B* gene. (C) Example of RiboSeq fragments across a hidden NOP involving the *BCAS4* gene in MCF7 cells. (D) Barplot indicating RiboSeq signal for three different open reading frame phases for hidden NOPs identified in MCF7, A375 and 786O cancer cell lines. The RiboSeq fragment counts were obtained from the novel sequence (3' portion) of the mRNA transcript encoding the hidden NOP. (E) Results of protein mass spectrometry analysis of NOPs for A375 cells. Each horizontal bar represents a hidden NOP peptide sequence driven by the *AHCY* gene, with each amino acid indicated with a different

color. The blue bar underneath each NOP sequence depicts the peptide identified by mass spectrometry.

Figure 4: Analysis of NOPs across cancer types. (A) Framome sizes, as measured in number of amino acids across 61 tumor samples included in this study. Different categories of NOPs are indicated. (B) and (C) Examples of the framomes of a lung tumor (LUN013) and glioblastoma (GBM005). Each horizontal bar represents the amino acid sequence of a single NOP expressed by the tumor. Different amino acids are depicted using different colors. The NOP sequences are sorted by length. (D) NOP expression plotted against NOP genomic variant allele frequency. Each dot represents one NOP. Dot size indicates NOP length in amino acids. Variant allele frequencies were adjusted for tumor purity. NOP expression is measured as transcripts per million (TPM). A small amount of jitter has been added to the points to avoid overplotting. (E) Overview of the origin of hidden NOPs and out-of-frame gene fusions categorized by SV type. Each item on the y-axis represents a (complex) structural variant event leading to the expression of one or more NOPs. The left barplot illustrates the number of unique NOPs arising from the given event colored by whether the NOP was a fusion gene or hidden NOP. The right barplot illustrates the number of genomic junctions involved in the structural variant event, with colors denoting junction class. Intergenic = SV breakpoint junctions for which there is no gene in the strand of the breakpoint on either side of the junction. Intragenic = SV breakpoint junctions where the breakpoints are located in the same gene. Gene-Intergenic = SV breakpoint junctions involving a 5'-end of a gene coupled to a genomic region where there is no gene in the strand of the breakpoint. Gene-Gene = SV breakpoint junctions involving a 5'-gene fused to a 3'-gene in the correct orientation to possibly form a fusion transcript.

Figure 5: Immunopeptidomics and in *vitro* immunogenicity of NOP-derived antigens in A375 cells. (A) Illustration of the A375 framome, the hidden NOP derived epitopes identified as presented via immunopeptidomics, and the number of framome peptides arising from various source genes selected for in-vitro immunogenicity testing in the context of the HLA on which they were predicted to bind. B) Dot plots showing percentage of FRM0469-specific CD8 T cells at the baseline, after stimulation with FRM0460 (negative control peptide), and with FRM0469 for the three tested donors. C) Lower panel: The total percentage of NOP-derived epitope-specific $CD8^+$ T cells of six HLA-A02:01 donors (left) or four HLA-A01:01 donors (right). Donors D2 and D3 were tested twice in two independent experiments. The values were corrected for the background staining in MOG re-stimulated condition. Upper panel: A representative flow cytometry analysis of IFNɣ and TNFα production by antigen specific CD8+ T cells of donor D3 and D9 for priming and re-stimulation with NOP-derived peptides (turquoise dots) or re-stimulation with negative control peptide MOG (orange dots). D) Summary of in *vitro* immunogenicity and dextramer staining responses across the 13 tested donors.

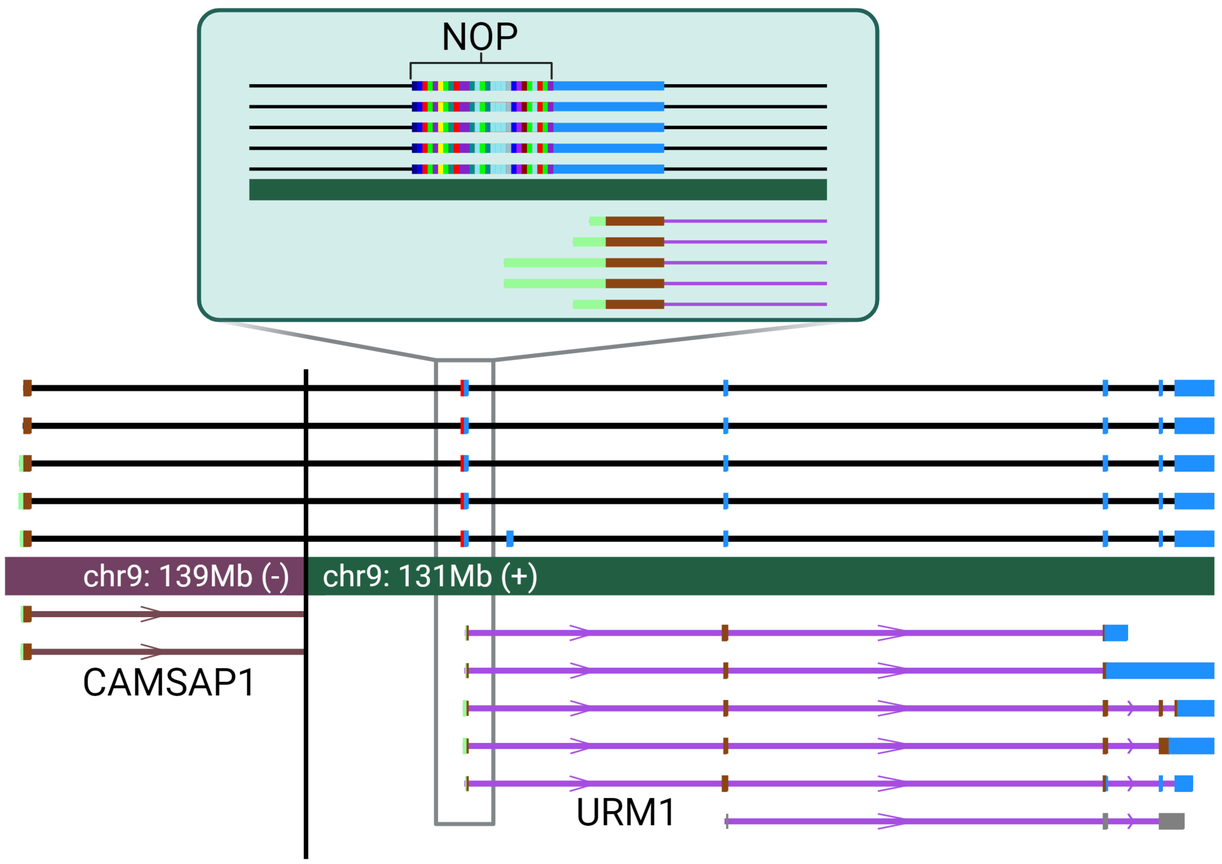
Figure 6: HLA binding and in vivo immunogenicity of NOPs. A) NOP-derived epitopes from three patients with NSCLC bound to patient specific HLA-A or B alleles as determined by *in vitro* binding assays. Positive control for each allele with high affinity was used to set 100% binding affinity and binding affinity of each peptide was calculated as relative to the positive control. Epitopes with at least 40% binding affinity are depicted. B) Analysis of PBMCs of patient LUN029 using combinatorial coding and immune phenotyping. The cells double positive for NOP-derived epitope-HLA tetramer complex are considered specific for the epitope. Phenotyping of epitope-tetramer complex double positive cells (black dots) was performed by staining with anti-CD45RA, anti-CD27, and anti-PD-1 to determine antigen experience status of the cells. C) Genomic origin of hidden NOPs identified in patient LUN029 for which $CD8^+$ T cells were identified in patient PBMCs.

Figure 1

Tumor
Normal
RNA
WGS
WGS
Variant Calling
Long Read
Mapping
Chimeric Long Reads
Short Read
Mapping
Short Read Splice Junctions
SV
SNV
Indel
splice
stoploss
Tumor Genome Graph
Normal Tissue Splice Junction Database
GTEx
Novel Splice Junction Identification
SV Contigs
Splice Contigs
Indel/Stoploss Contigs
Reconstructed Tumor Genome
SV
Splice
Stoploss
Indel
Remap RNA
Short Reads
Long Reads
Correct, collapse, and translate isoforms
WT k-mer Database
Isoforms
Genes
Identify NOPs
P-site genome
Isoforms
Genes
Compile NOPs into Framome

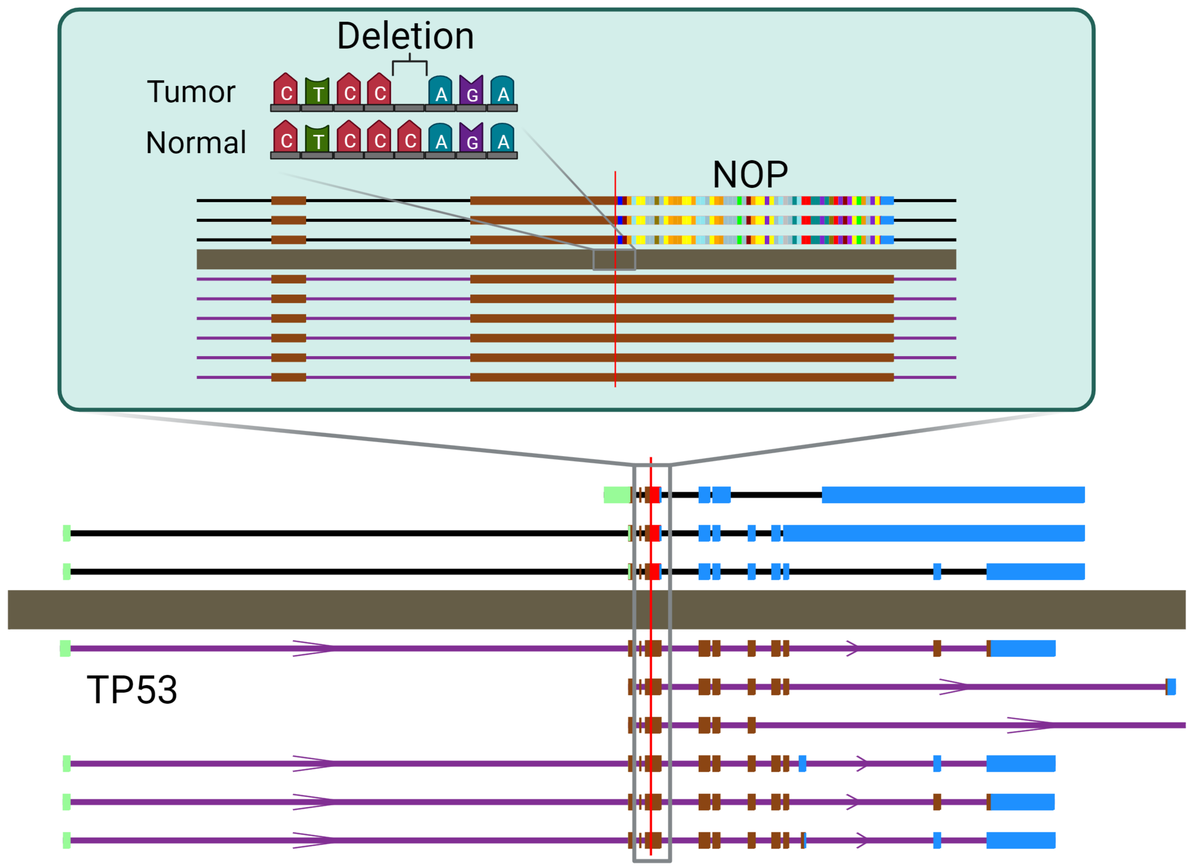
A

Structural Variant



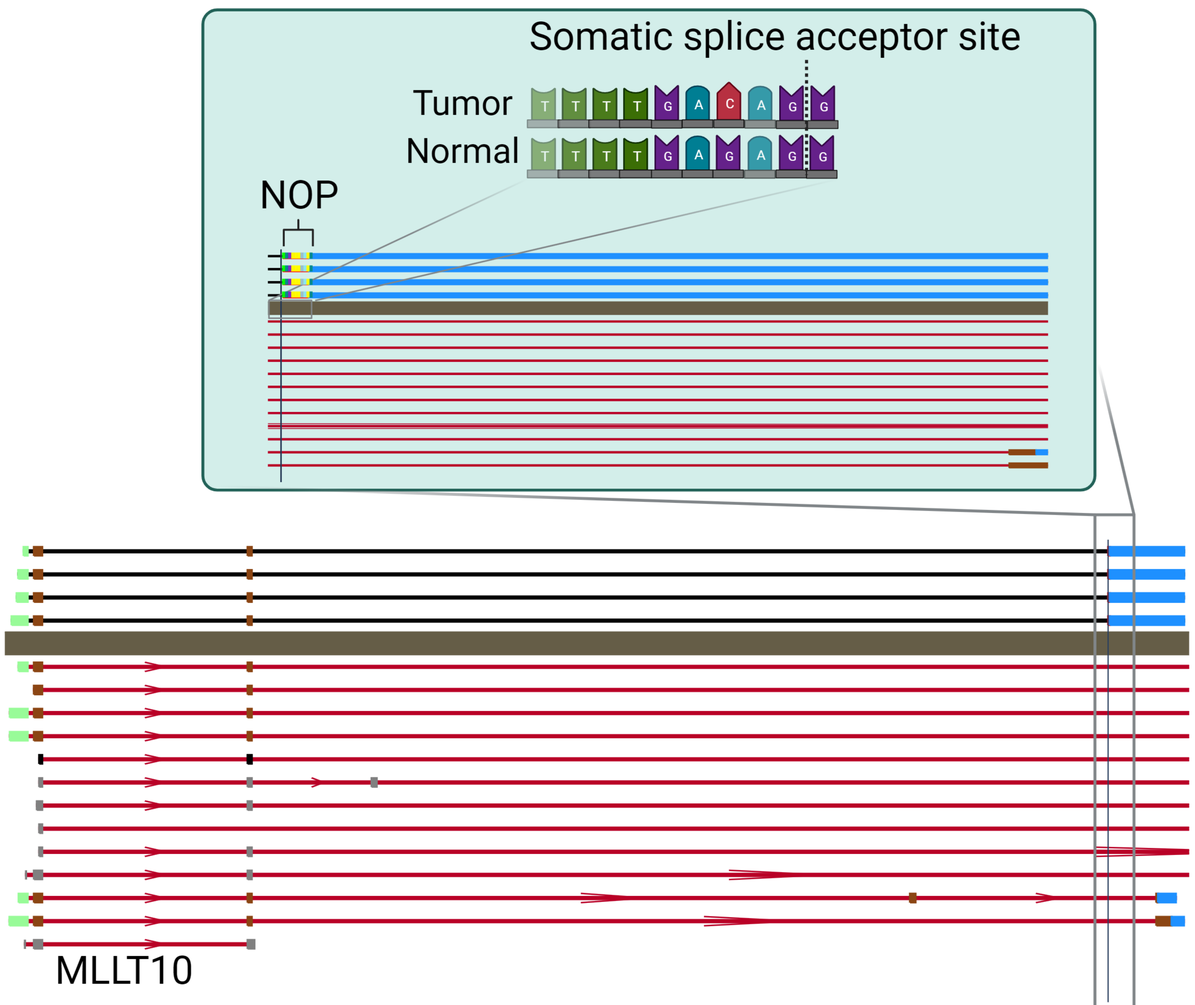
B

Indel



C

Splice



D

Stoploss

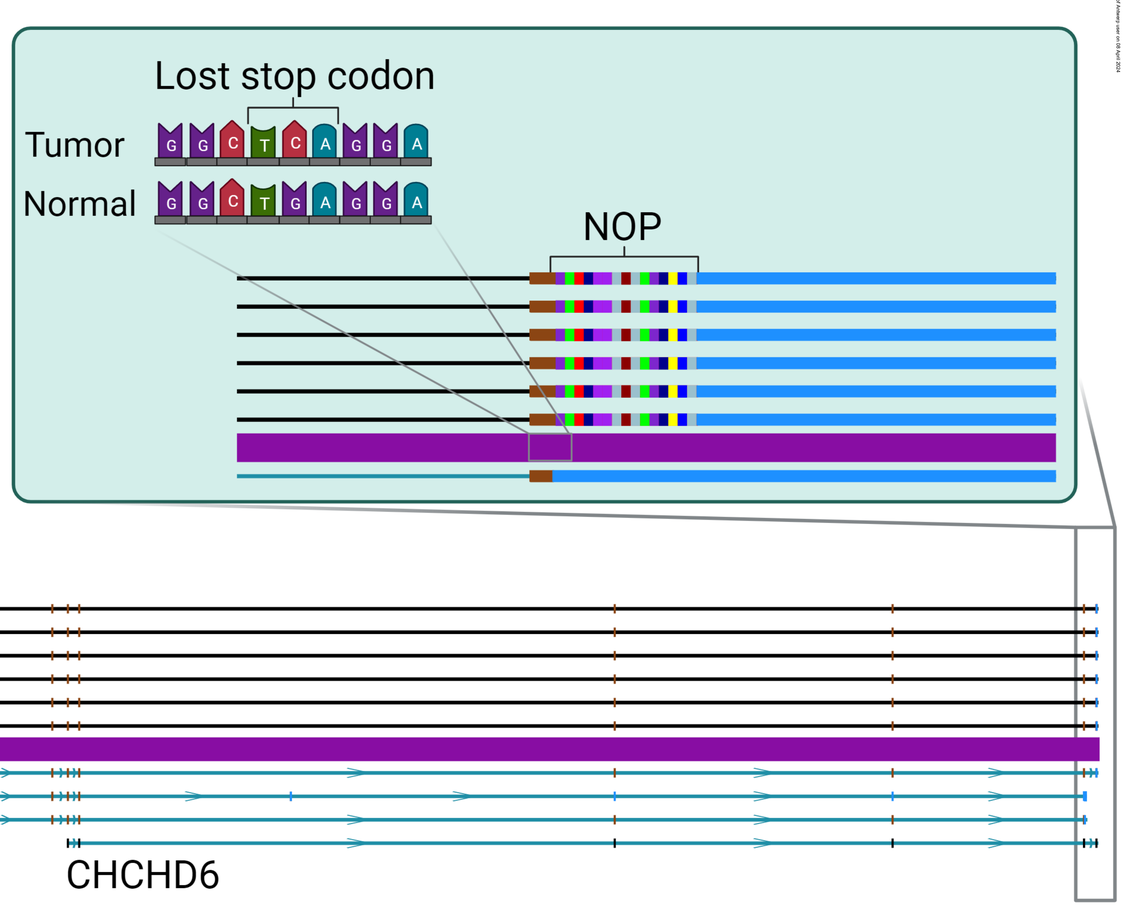
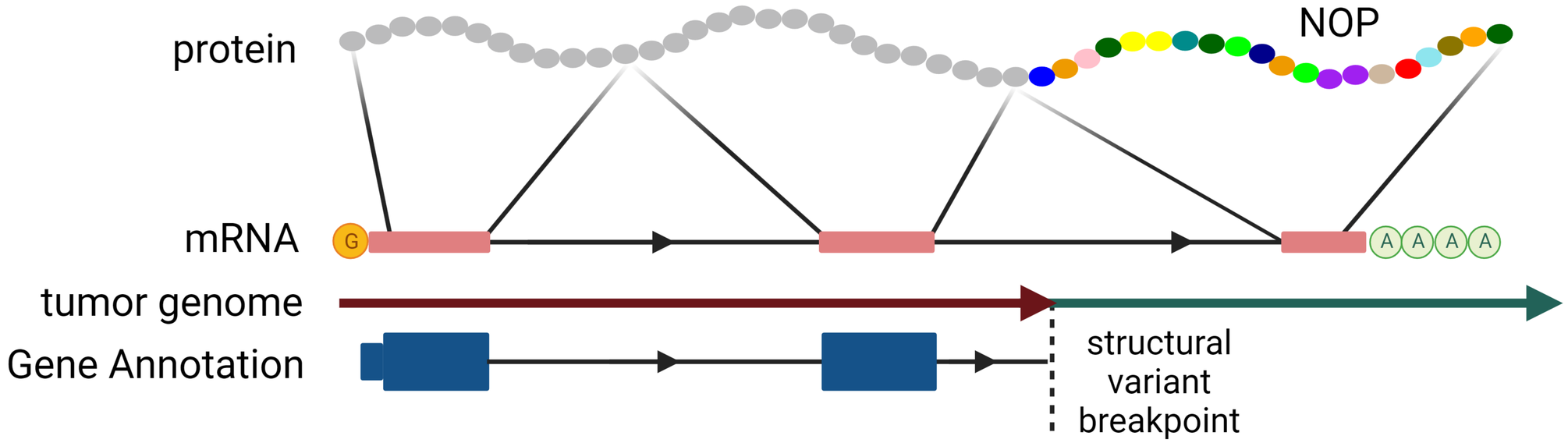
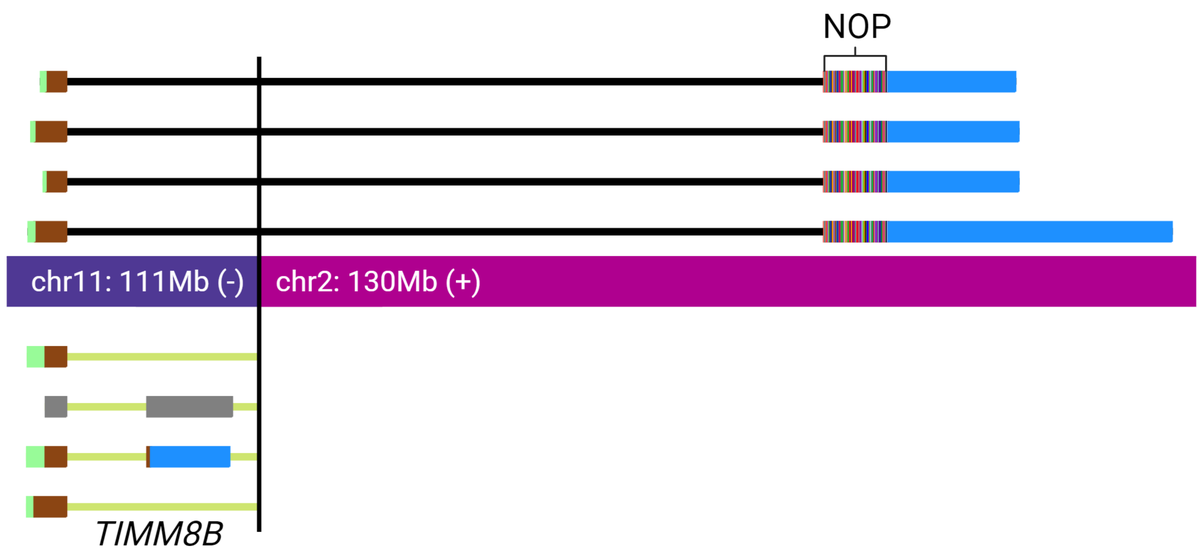


Figure 3

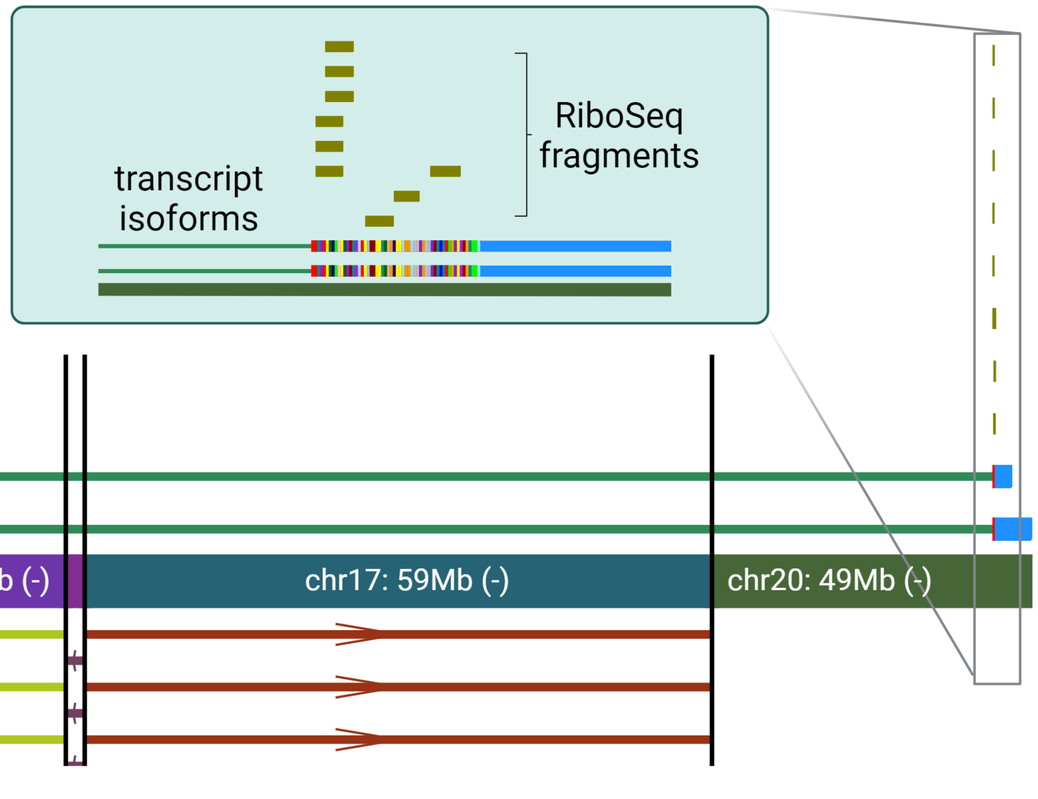
A



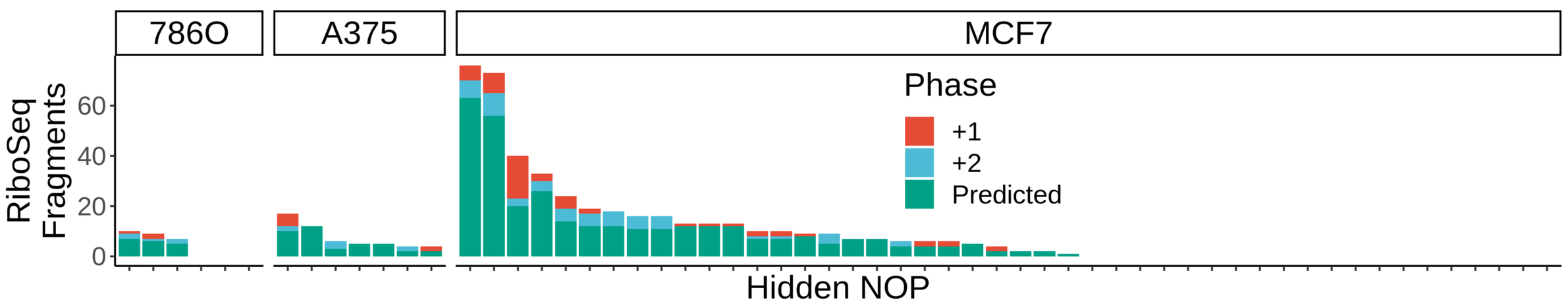
B



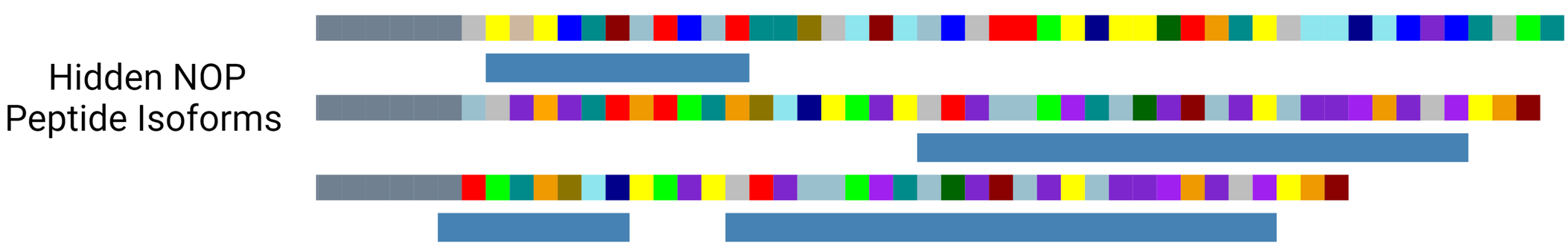
C



D



E



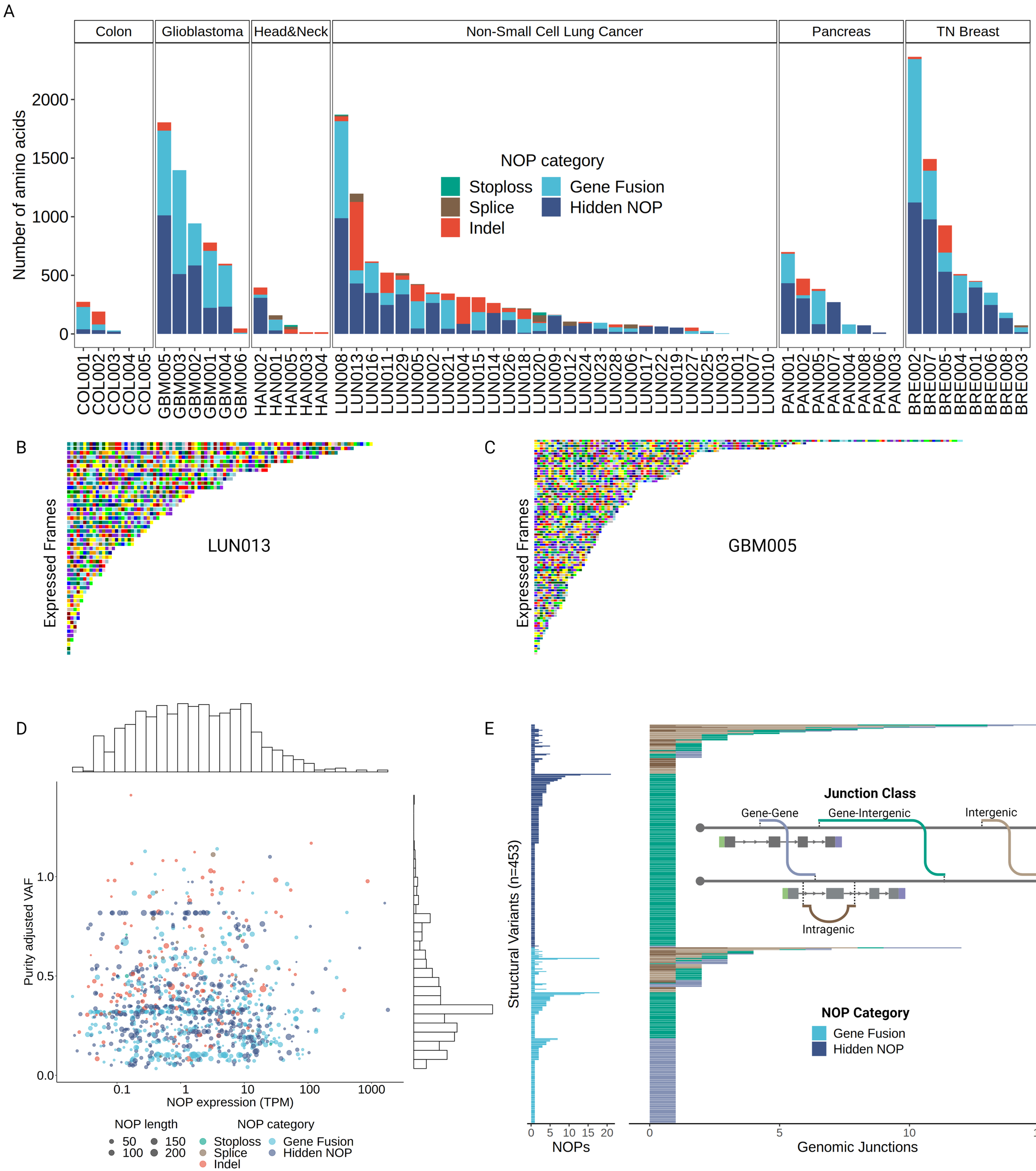

A
Colon
Glioblastoma
Head&Neck
Non-Small Cell Lung Cancer
Pancreas
TN Breast
Number of amino acids
2000
1500
1000
500
0
NOP category
Stoploss
Splice
Indel
Gene Fusion
Hidden NOP
COL001 COL002 COL003 COL004 COL005
GBM005 GBM003 GBM002 GBM001 GBM004 GBM006
HAN002 HAN001 HAN005 HAN003 HAN004
LUN008 LUN013 LUN016 LUN011 LUN029 LUN005 LUN002 LUN021 LUN004 LUN015 LUN014 LUN026 LUN018 LUN020 LUN009 LUN012 LUN024 LUN023 LUN028 LUN006 LUN017 LUN022 LUN019 LUN027 LUN025 LUN003 LUN001 LUN007 LUN010
PAN001 PAN002 PAN005 PAN007 PAN004 PAN008 PAN006 PAN003
BRE002 BRE007 BRE005 BRE004 BRE001 BRE006 BRE008 BRE003
B
Expressed Frames
LUN013
C
Expressed Frames
GBM005
D
Purity adjusted VAF
1.0
0.5
0.0
0.1
1
10
100
1000
NOP expression (TPM)
NOP length
50
100
150
200
NOP category
Stoploss
Splice
Indel
Gene Fusion
Hidden NOP
E
Structural Variants (n=453)
0 5 10 15 20
NOPs
Junction Class
Gene-Gene
Gene-Intergenic
Intergenic
Intragenic
NOP Category
Gene Fusion
Hidden NOP
0
5
10
15
Genomic Junctions

Figure 5

A

A375 Framome

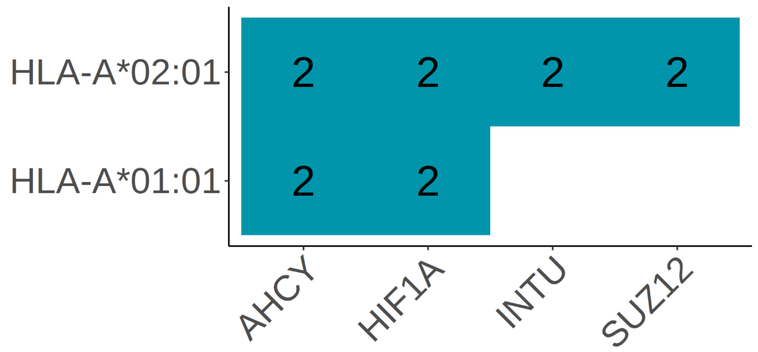
AHCY

Epitopes identified via immunopeptidomics

Peptide HSVAQAGVQW

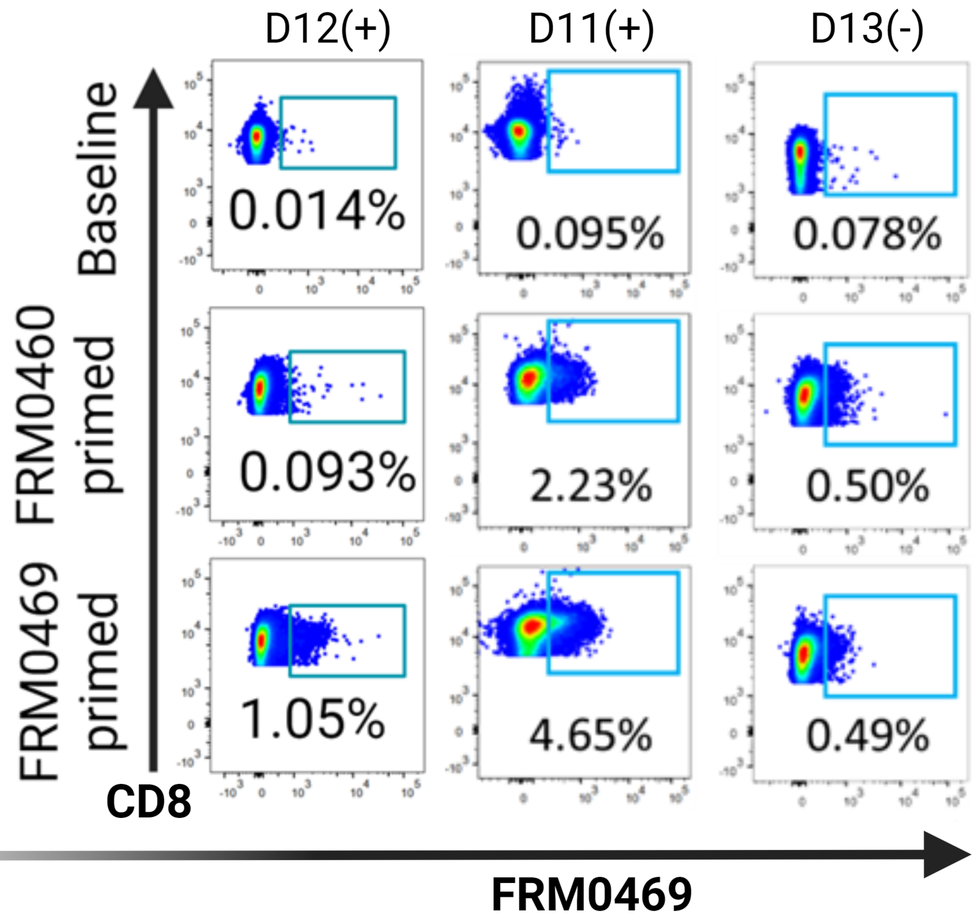
Peptide VTKAGVQW (FRM0469)

Epitopes tested in in-vitro immunogenicity

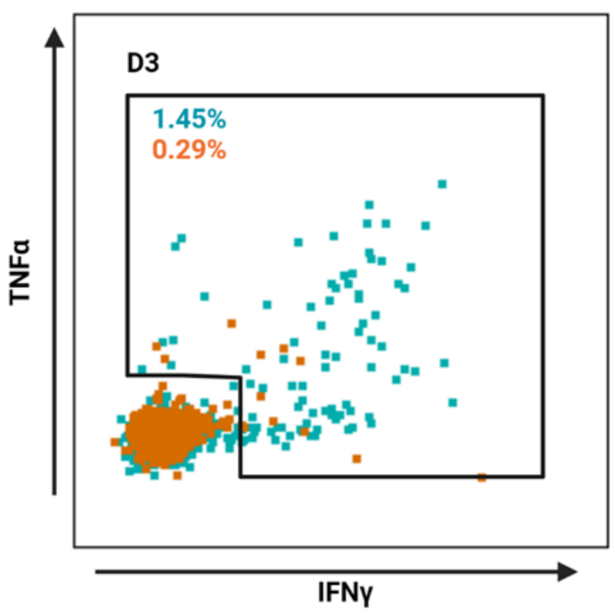


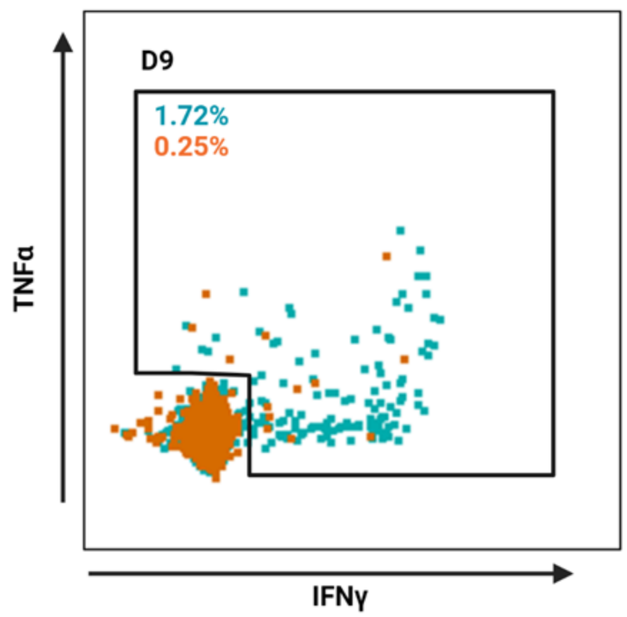
HLA-B*57:01 Dextramer

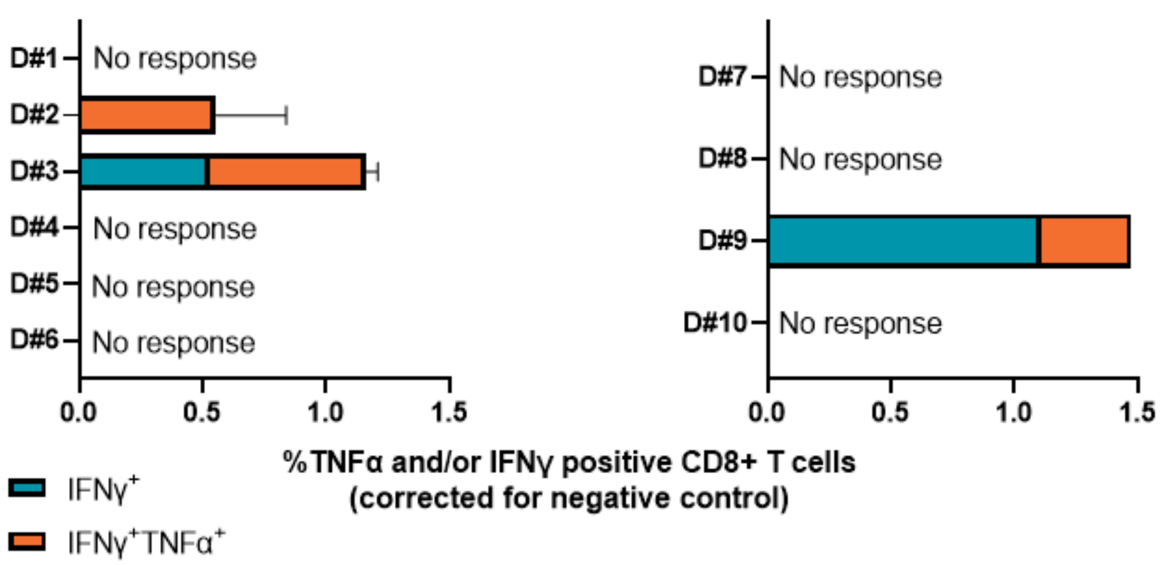
B



C







D

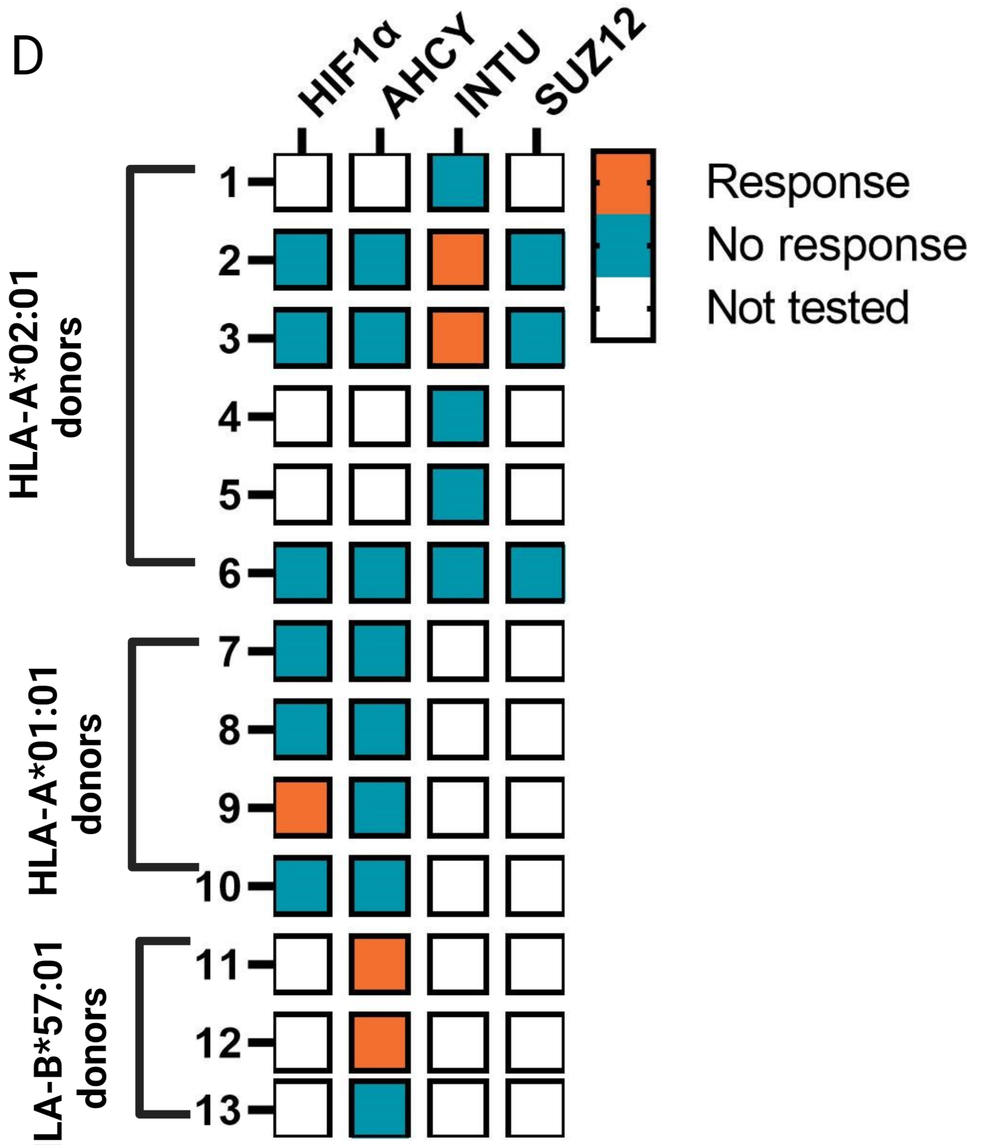
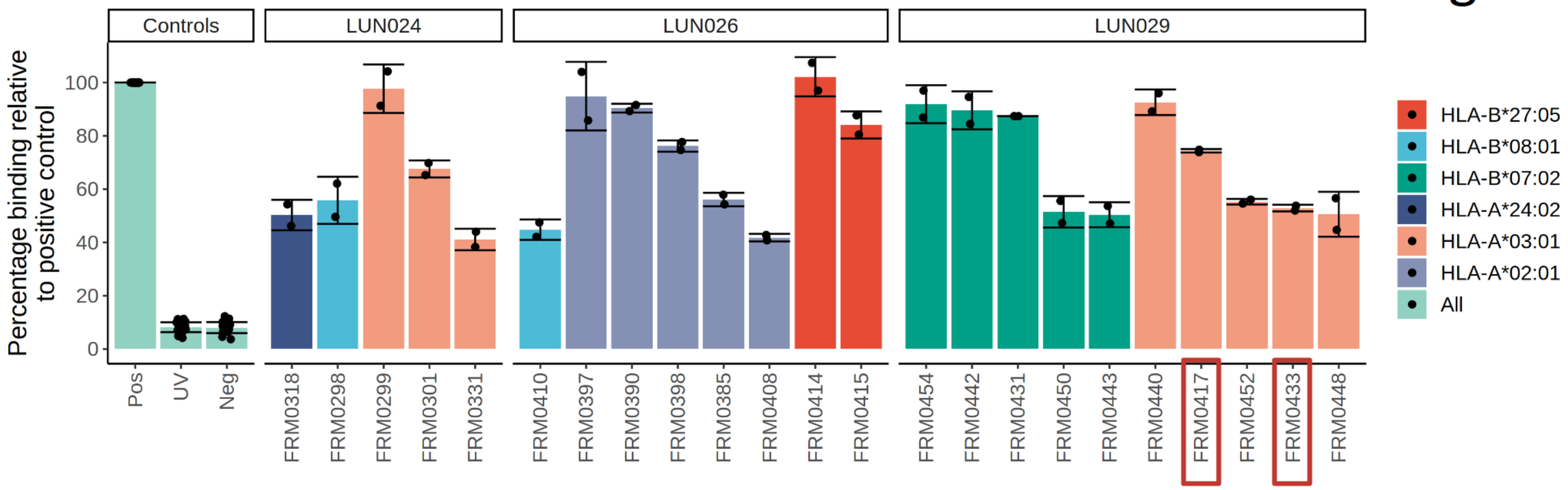
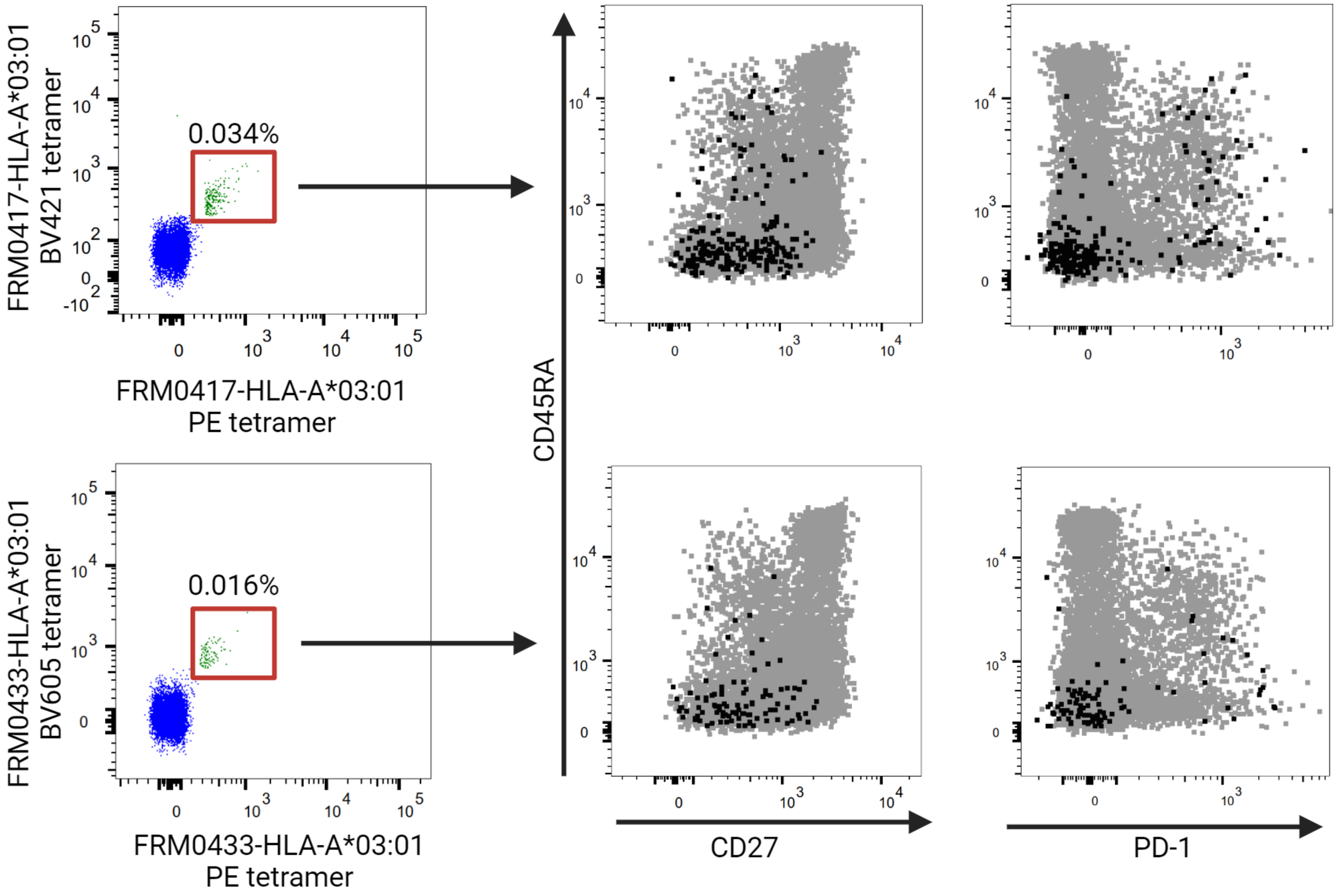


Figure 6

A



B



C