Long-read sequencing reveals widespread novel splicing and neojunction-derived neoantigens in nasopharyngeal carcinoma

Yi Shuai, Hualiang Yao, Bo Wang, et al.

Genome Res. published online July 8, 2026
Access the most recent version at doi:10.1101/gr.281467.125

Long-read sequencing reveals widespread novel splicing and neojunction-derived neoantigens in nasopharyngeal carcinoma

Yi Shuai[1,2,4], Hualiang Yao[1,4], Bo Wang[2,3], Grace TY Chung[2,3], Xiangeng Wang[1,4,6], Ming Zhong[1,4], Zhongxu Zhu[1], Cheuk Shuen Li[5], Chi Man Tsang[2,3], Kwok-Wai LO[2,3*], Xin Wang[1,4,6*]

[1]Department of Surgery, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China.

[2]Department of Anatomical and Cellular Pathology, The Chinese University of Hong Kong, Hong Kong SAR, China.

[3]State Key Laboratory of Translational Oncology, Sir YK Pao Centre for Cancer, The Chinese University of Hong Kong, Hong Kong SAR, China

[4]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China

[5]School of Biomedical Sciences, University of Hong Kong, Pokfulam, Hong Kong SAR, China

[6]Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen 518000, China

* Correspondence: xinwang@cuhk.edu.hk; kwlo@cuhk.edu.hk

Abstract

The widespread transcriptomic diversity driven by alternative splicing (AS) contributes to all hallmarks of cancer and represents a critical source of neoantigens for personalized immunotherapy. However, unlike other major malignancies, the full repertoire of alternative splicing in nasopharyngeal carcinoma (NPC) remains underexplored. Here, we employ long-read sequencing (LR-seq) to generate a high-resolution, isoform-level transcriptomic atlas from a cohort of 14 NPC tumor samples and four immortalized nasopharyngeal epithelial cell lines. We identify a substantial number of full-length novel transcripts (22,687; ~44.38%), which reveal diverse splicing patterns and previously unannotated splicing events. By integrating short-read RNA-seq data to quantify isoform expression, we discover a subset of novel transcripts that are differentially expressed between tumor samples and

immortalized nasopharyngeal epithelial cell lines. Furthermore, LR-seq enables precise identification of chimeric read-through fusion transcripts, such as *CLDN15-FIS1* and *FOXRED2-TXN2*. Finally, we develop a computational framework-**T**umor-**s**pecific **s**plicing **n**eo**a**ntigen **d**etection (TS-SNAD) to predict neoantigens originating from novel exon-exon junctions (neojunctions) in tumor-specific novel transcripts. Using this framework, we identify neojunction-derived neoantigens and experimentally validate the immunogenicity of selected HLA-B*40:01-restricted neoantigens. These neojunction-derived peptides constitute a new class of non-canonical neoantigens with significant potential for developing personalized cancer vaccines for NPC.

Keywords: nasopharyngeal carcinoma; long-read sequencing; short-read sequencing; alternative splicing; isoform diversity; TS-SNAD; neojunction-derived neoantigens; cancer immunotherapy.

Introduction

Nasopharyngeal carcinoma (NPC), which originates from the epithelium of the nasopharynx, is endemic in Southeast Asia and Southern China. Distinct from other head and neck cancers, NPC is characterized by its consistent association with Epstein-Barr virus (EBV) infection, unique genomic landscapes and specific histopathological features (Tsang et al. 2020). As early-stage NPC is often asymptomatic, approximately 80% of patients present with locoregionally advanced or metastatic disease at diagnosis. For these patients, chemoradiotherapy is the standard of care, but treatment outcomes remain unsatisfactory for those with locoregionally advanced or metastatic disease. Consequently, novel therapeutic strategies, such as targeted therapies and immunotherapies, are urgently needed (Siak et al. 2023).

Alternative splicing (AS) is a key driver of oncogenesis, promoting cancer progression through isoform switches that directly regulate core cancer hallmarks. Most tumors exhibit widespread splicing dysregulation, altering exon-intron architecture, promoter usage, and isoform expression, ultimately disrupting the function of critical tumor suppressors and oncogenes (Huang et al. 2021; Bradley and Anczukow 2023; Maurin et al. 2023; Joglekar et al. 2024; Naro et al. 2024). A comprehensive characterization of transcriptomic diversity is therefore essential for understanding cancer

development (Bhattacharya et al. 2023). Conventionally, the detection and quantification of AS events in cancers have relied on short-read RNA sequencing (RNA-seq) (Hong et al. 2020), which is inherently limited by its dependence on algorithmic reconstruction of transcripts from short fragments aligned to a reference genome. As a result, this strategy often yields an incomplete view of the splicing repertoire, as short reads are frequently insufficient to unambiguously resolve complex splice junctions or capture full-length transcripts (Hook and Timp 2023). Recent advances in long-read sequencing (LR-seq) technologies, such as Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies (ONT) nanopore sequencing, are poised to overcome these challenges (Monzo et al. 2025; Somalraju et al. 2025). By generating reads that span multiple kilobases, LR-seq enables the direct, accurate identification of full-length alternatively spliced transcripts, allowing for a more comprehensive characterization of the AS landscape in cancer.

Several studies have reported that tumor-specific or -associated non-synonymous mutations, gene fusions, transposable elements, and post-translational modifications could generate neoantigens (Burbage et al. 2023; Xie et al. 2023; Hu et al. 2024; Kumar et al. 2024). In practice, neoantigens derived from non-synonymous single-nucleotide variants (SNVs) and gene fusions have been the primary focus of personalized cancer vaccine development in melanoma (Lauss et al. 2020; Ott et al. 2020; Borgers et al. 2025), glioma (Platten et al. 2021; Dunn et al. 2022; Zelba et al. 2025), and pancreatic cancer (Sethna et al. 2025). More recently, alternative RNA splicing has emerged as a validated source of neoantigens, including immunogenic cryptic peptides arising from unannotated splice junctions, thereby expanding the antigen repertoire and eliciting antitumor immune responses (Smart et al. 2018; Lu et al. 2021; Merlotti et al. 2023). Despite these novel insights, the non-canonical peptide landscape remains incompletely defined. Previous investigations and computational pipelines for splicing-derived neoantigen discovery have largely relied on short-read RNA-seq (Pan et al. 2023) and reference-based annotations, which require transcript assembly from fragmented reads and may compromise the accurate reconstruction of full-length transcripts and their open reading frames (ORFs).

In this study, we employ LR-seq to provide the first systematic view of transcriptomic diversity in NPC, leading to the discovery and characterization of full-length transcripts, including numerous previously undescribed transcripts. By directly resolving full-length transcript structures, LR-seq also reveals widespread alternatively spliced isoforms and read-through transcripts (RTTs). To further validate and quantify the LR-seq-derived transcriptome, we integrate short-read RNA-seq data and identify a tumor-specific subset. Most importantly, we develop **T**umor-**s**pecific **s**plicing **n**eo**a**ntigen **d**etection (TS-SNAD), a new computational framework that leverages LR-seq to identify tumor-specific, novel exon-exon junctions (neojunctions) from novel transcripts by directly resolving transcript structures. Our approach minimizes errors introduced by transcript assembly in short-read RNA-seq analyses and enables accurate prediction of neojunction-derived neoantigens. We further employ TS-SNAD to predict neojunction-derived neoantigens and experimentally validate the immunogenicity of selected candidates predicted to bind the high-frequency NPC allele HLA-B*40:01. Our study provides novel insights into the NPC transcriptome and a valuable resource for developing off-the-shelf cancer immunotherapies for NPC patients.

Results

Long-read sequencing unveils a vast repertoire of novel transcripts and extensive splicing diversity in the NPC transcriptome

To comprehensively profile the AS landscape in NPC, we performed long-read RNA sequencing on a cohort of 18 samples, including both non-malignant and malignant samples. This cohort included two NPC cell lines, ten patient-derived xenografts (PDXs), two NPC tumor tissues, and four immortalized nasopharyngeal epithelial (NP) cell lines. All samples were analyzed using SMRT circular consensus sequencing (CCS) on the PacBio Sequel II platform. On average, ~783,000 CCS reads were obtained per library, yielding ~350,000 clustered isoforms after downstream processing **(Table 1 and Supplemental Table S1).** For each sample, the numbers of genes and full-length transcripts identified in LR-seq after collapsing were reported (**Supplemental Fig. S1B)**. The observed differences across samples were expected and likely reflected variation in RNA quality and biological heterogeneity.

After merging transcripts across samples and collapsing redundant isoforms, we identified 51,115 high-quality full-length transcripts with a median transcript length of approximately 3.0 kb **(Fig. 1A and Fig. 1C)**, constructing a comprehensive LR-seq-derived transcriptome for NPC.

Transcripts were categorized using SQANTI3 based on their structural alignment to the GENCODE v38 reference transcriptome. This classification revealed 24.10% (12,319) as full splice match (FSM), perfectly matching a known reference transcript; 26.10% (13,341) as incomplete splice match (ISM), which align to a known transcript but lack complete 5' or 3' ends; 22.65% (11,578) as novel in catalog (NIC), containing known splice sites in novel combinations; 21.73% (11,109) as novel not in catalog (NNC), featuring at least one previously unannotated splice site. The remaining 2,768 transcripts were classified as antisense, genic, fusion, or intergenic transcripts (**Fig. 1A**). Novel transcripts (NIC/NNC) constituted a substantial portion of the transcriptome, representing between 17% and 36% of all transcripts per sample (**Fig. 1B**), underscoring the substantial unannotated transcriptional complexity in NPC. Collectively, these transcripts were mapped to 11,807 annotated genes and 2,259 novel genomic loci. Notably, 56.66% (6,690) of the genes in the LR-seq-derived transcriptome expressed two or more transcripts, indicating widespread alternative splicing (**Fig. 1D**).

To assess the reliability of the LR-seq-derived transcriptome, we aligned short-read RNA-seq data from matched samples to our LR-seq-derived transcriptome. This validation revealed a clear hierarchy of support [TPM (transcripts per million) $\geq$1]: FSM transcripts showed the highest validation rate (~68-80%); NIC showed intermediate validation rates (~40-60%); NNC and ISM transcripts showed lower support (~30-40%); Antisense, genic, fusion, or intergenic transcripts exhibited the lowest validation rates (~10-25%) (**Fig. 1E**). We next evaluated the distribution of isoforms across samples. We quantified the number of samples supporting each isoform using two criteria: long-read presence/absence calls (binary 0/1 per sample) and short-read RNA-seq expression (counting samples with TPM above the multiple thresholds). FSM transcripts exhibited low heterogeneity, with many isoforms detected across multiple samples. Conversely, novel transcripts exhibited high sample-to-sample variability, with the majority detected in only a few samples, highlighting their greater heterogeneity relative to well-annotated FSM isoforms **(Supplemental Fig. S2A-B).**

Comprehensive characterization of novel transcripts with quality evaluation reveals their structural diversity, functional relevance, and potential roles in NPC pathogenesis

Next, we performed quality assessments of the identified transcripts, focusing on coding potential, susceptibility to nonsense-mediated mRNA decay (NMD), and splice site characteristics. We observed that the vast majority of FSM and novel transcripts were predicted to be protein-coding (94-97%), whereas antisense, genic, fusion, or intergenic transcripts largely lacked coding potential. NMD prediction revealed that NNC, antisense, genic, fusion, or intergenic transcripts were more likely to contain premature termination codons (PTCs) and non-canonical splice sites **(Fig. 2A)**, consistent with increased susceptibility to RNA surveillance and degradation (Supek et al. 2021)**.** Investigating the mechanisms triggering NMD indicated that the use of novel splice sites and intron retention were the primary causes, often generating PTCs that mark transcripts for degradation **(Supplemental Fig. S2C)**. For example, two novel *ANO9* transcripts (TCONS_00009816 and TCONS_00009828) were predicted to undergo NMD due to a novel splice site and intron retention, respectively, both introducing PTCs (**Supplemental Fig. S2D**). Furthermore, we observed a positive correlation between the number of exons per gene and the number of novel transcripts produced by that gene, suggesting that genes with higher exon complexity have a greater potential for generating novel splice variants (**Supplemental Fig. S2E**).

Comparison with reference transcripts suggested that NNC isoforms often arose from unannotated transcriptional start sites (TSSs) and termination sites (TTSs) located >100 bp from known transcript boundaries **(Supplemental Fig. S2F).** In addition to RNA-seq, we integrated multiple orthogonal datasets to validate novel isoforms identified by LR-seq, including cap analysis of gene expression (CAGE) and 3'-seq peaks from the poly(A) site database. The 5' ends of many novel isoforms overlapped with FANTOM5 CAGE-defined TSSs, and their 3' ends were validated by bona fide TTSs mapped by 3'-seq poly(A) databases **(Supplemental Fig. S2G)**. Comparative analysis revealed distinct structural features distinguishing FSM transcripts from novel transcripts. NNC isoforms were typically shorter in total length than FSM and NIC transcripts, yet novel transcripts often possessed longer coding sequences (CDSs) than FSM isoforms. Furthermore, splice junctions in FSM transcripts

showed higher read coverage, while novel isoforms, particularly those with novel splice sites, had lower read support **(Fig. 2B)**.

To evaluate the biological relevance of novel transcripts, we performed pathway enrichment analysis. Genes associated with novel transcripts were significantly enriched in pathways related to RNA processing, nuclear export, and intracellular transport, based on a matched background gene set **(Fig. 2C)**. Notably, several established oncogenes and tumor suppressors, including *FGFR1*, which can be activated by LMP1 to promote glycolysis and invasion in NPC (Lo et al. 2021), *DDX5*, implicated in NAT10-driven immune suppression (Xie et al. 2025), and *MST1R*, which is linked to early-onset NPC and innate immunity (Dai et al. 2016; Cazes et al. 2022) exhibited substantial numbers of novel isoforms, suggesting a crucial role in NPC pathogenesis **(Fig. 2D)**. Most novel transcripts (67.8%) shared the same translation start sites as known protein isoforms, indicating that the associated splicing events mainly affected internal exons. In contrast, 32.2% potentially used alternative translation start sites, possibly associated with alternative promoter usage or first exon selection, which may alter the 5′ UTR or the exon harboring the canonical start codon **(Fig. 2E)**.

In addition to novel host transcripts identified in NPC samples, we also detected previously unannotated EBV transcripts, including transcripts with alternative first exon in *LMP2* and exon skipping of exons 2 and 3 in *RPMS1* **(Supplemental Fig. S2H).** To assess the differences between sequencing approaches, we compared the transcriptome derived from LR-seq with the transcriptome assembled from short-read data using StringTie. The StringTie-assembled transcriptome comprised 239,952 transcripts, far exceeding the number identified by LR-seq. However, only 14,198 transcripts were commonly identified by both methods, most of which were FSMs. This discrepancy underscored the substantial variation in transcript identification between short-read and long-read sequencing technologies, highlighting the limitations of short-read-based assembly in reconstructing full-length isoforms **(Supplemental Fig. S2I).**

Moreover, a key advantage of LR-seq lies in its ability to accurately identify RTTs by capturing full-length transcript structures. This enabled precise delineation of fusion transcripts spanning adjacent genes. These RTTs and their encoded products have emerged as potential diagnostic biomarkers and

therapeutic targets in various cancers (Alpert et al. 2020; Hu et al. 2022). In our study, we identified and experimentally validated four such RTTs, including *SFT2D2-TBX19*, *FOXRED2-TXN2*, *ELAC1-SMAD4* and *CLDN15-FIS1*, using RT-PCR with specific primers targeting the fusion junctions **(Supplemental Fig. S3A-B).**

Next, we annotated seven types of alternative splicing events (ASEs) using SUPPA2, including skipping exons (SE), alternative 3' splice sites (A3), alternative 5' splice sites (A5), alternative first exon (AF), alternative last exon (AL), intron retention (RI), and mutually exclusive exons (MX) categories **(Fig. 3A)**, in GENCODE v38 transcriptome and its merged transcriptome incorporating novel transcripts identified by LR-seq. Notably, incorporating novel isoforms increased the number of detected ASEs, particularly AF and SE, while the counts of AL and MX remained largely unchanged **(Fig. 3B).** Differential ASE analysis between NPC tumor samples and immortalized NP cell lines identified 102 differentially spliced events ($P < 0.05$, $|\Delta PSI| > 0.1$; **Supplemental Table S2**), with AF (43, 42.16%) and SE (25, 24.51%) being the most prevalent **(Fig. 3C)**. A prime example was *TLN1*, which exhibited a tumor-specific switch involving inclusion of a cancer-associated cassette exon (17b). This isoform, which enhances vinculin binding and cell motility (Gallego-Paez et al. 2023), was significantly upregulated in NPC samples (**Fig. 3C-D**). We validated this isoform switch by RT-qPCR, confirming that the novel isoform (TCONS_00049374) accounted for a significantly larger fraction of total *TLN1* expression in tumors than the canonical isoforms (ENST00000314888.10) (**Fig. 3E**).

Long-read transcriptomics identifies candidate transmembrane therapeutic targets and tumor-specific transcripts

To identify transcripts enriched in NPC tumors, we quantified expression by mapping matched short-read RNA-seq data to our LR-seq-derived transcriptome. Analysis of transcript expression density curves and distribution patterns revealed that most FSMs exhibited relatively higher expression levels compared with novel transcripts **(Fig. 4A-B).** Using differential transcript expression (DTE) analysis between NPC tumor samples and immortalized NP cell lines, we identified 1,812 significantly dysregulated transcripts ($|\log_2 FC| > 2$, BH-adjusted $P < 0.05$). This included 460 upregulated and 325 downregulated FSM transcripts, alongside 241 upregulated and 117 downregulated NIC transcripts,

and 211 upregulated and 86 downregulated NNC transcripts (**Supplemental Table S3**). Highly expressed alternatively spliced isoforms in NPC included known isoforms of *CSF2RB*, *LUM*, and *COL1A2*, as well as novel isoforms of *PRRX1*, *LGALS9*, and *CD74* (**Fig. 4C**). Notably, overexpression of *CSF2RB* and *PRRX1* contributes to promoting oncogenic activity and maintaining epithelial-mesenchymal transition (EMT) in human cancers (Du et al. 2021; Charlet et al. 2022). High expression of *LGALS9* (Klibi et al. 2009; Kam et al. 2025) and *CD74* (Chow et al. 2022) has been reported to play important roles in promoting immune evasion and remodeling the tumor microenvironment in NPC. While the well-known transcripts were used in previous functional studies, the identification of alternative and novel transcripts, especially the novel *LGALS9* and *CD74* transcripts, in NPC indicated the potential for new oncogenic activities of these transcripts in NPC tumorigenesis. Because of the potential functional difference between novel and known transcripts, further studies on the NPC-specific transcripts are needed to clarify their key oncogenic activities in tumorigenesis.

We next prioritized transcripts predicted to encode plasma membrane transmembrane proteins because these proteins are stably expressed at the cell surface, present extracellularly accessible domains, and therefore represent attractive targets for therapeutic antibodies (Hu et al. 2021; Bardia et al. 2024; Neri et al. 2024; Zaidi et al. 2024; Boixareu et al. 2025). Based on predictions of transmembrane topology and plasma membrane localization, 1,333 known transcripts from 859 genes were predicted to encode cell-surface transmembrane proteins. Additionally, 2,159 novel transcripts from 595 genes were predicted to encode cell-surface proteins containing transmembrane domains **(Supplemental Fig. S4A and Supplemental Table S4**). Notably, we identified 294 novel transcripts from 95 genes that were predicted to gain cell-surface localization with transmembrane helices, even though these genes were absent from the TCSA reference surfaceome set (Hu et al. 2021). In contrast, we identified 1,378 novel transcripts from 360 genes in the curated TCSA reference surfaceome whose encoded protein products were predicted to lose cell-surface localization relative to the annotated protein products of the same genes **(Supplemental Table S5)**. DTE analysis revealed significant upregulation of several cell membrane-associated transcripts in NPC tumor samples (e.g., *IGSF9*, *CXADR*, *TMPRSS2*, *INSR*;

238 **Supplemental Fig. S4B**, **Supplemental Table S3**). Survival analysis of an NPC public dataset
239 (GSE102349) indicated that high expression of these cell membrane-associated transcripts correlated
240 with poorer progression-free survival (PFS), highlighting their potential clinical relevance
241 (**Supplemental Fig. S4C**).

242 Among the predicted plasma membrane transmembrane isoforms, predominant expression of the
243 NTRK2-T1 protein isoform was reported in our recent publication (Wang et al. 2025). The expression
244 of the *NTRK2-T1* transcript was also confirmed by short-read RNA-sequencing, and the corresponding
245 protein expression was confirmed by western blotting in NPC. By IHC, we demonstrated the
246 membrane staining of NTRK2-T1 protein isoform in the NPC xenografts and primary tumors
247 (**Supplemental Fig. S5A-C)**. To further validate the newly identified transmembrane CSF2RB protein
248 isoform TCONS_00033496, we examined its protein expression and subcellular localization in vitro.
249 Western blot analysis confirmed expression of FLAG-tagged TCONS_00033496 in both HEK293T
250 and NP69 cells after transfection. Immunofluorescence staining showed that the FLAG signal was
251 predominantly localized at the cell periphery and colocalized with the plasma membrane marker WGA
252 in both cell lines, whereas no specific anti-FLAG signal was detected in the corresponding negative
253 control cells. These findings indicated that TCONS_00033496 encoded a novel CSF2RB protein
254 isoform that was properly translated and targeted to the plasma membrane **(Supplemental Fig. S5D).**

255 We next assessed the tumor specificity of all AS variants. While the majority of transcripts (34,416,
256 67.33%) were expressed in both NPC tumors and immortalized NP cell lines, and 6,487 (12.69%)
257 were lowly expressed in all samples, a significant fraction (9,349 transcripts, 18.29%) was exclusively
258 detected in NPC tumor samples **(Supplemental Table S6)**. In contrast, only 1.69% (863 transcripts)
259 were specific to immortalized NP cell lines **(Fig. 4D)**. Tumor-specific novel transcripts constituted a
260 substantial proportion of this group, suggesting they may play a functional role in tumor biology and
261 represent a reservoir of novel biomarkers.

262 To validate these findings, we selected a subset of highly expressed, tumor-specific novel isoforms
263 that showed TPM $\geq$ 10 in at least one tumor sample and harbored isoform-specific novel junctions
264 with short-read splice junction coverage $\geq$ 10. A heatmap of their expression patterns in our RNA-seq

cohort confirmed their tumor-specific expression, though substantial inter-tumor heterogeneity was observed **(Fig. 4E).** This tumor specificity was further validated by RT-qPCR **(Fig. 4F),** and the isoform structures were depicted in **Supplemental Fig. S6**. We further validated these results by integrating and analyzing three public NPC RNA-seq datasets (GSE118719, GSE134886, GSE68799), comprising 52 NPC tumors and 11 non-tumor controls after batch-effect correction **(Supplemental Fig. S7A)**. The expression patterns of the novel transcripts in these independent cohorts were consistent with our in-house data, confirming their specific and elevated expression in NPC tumors **(Supplemental Fig. S7B).**

A computational pipeline identifies potential tumor-specific neojunction-derived neoantigens for NPC immunotherapy development

To expand the repertoire of immunotherapeutic targets, we investigated neojunctions derived from tumor-specific novel transcripts. These neojunctions represented a source of non-canonical tumor-specific antigens, significantly broadening the landscape of potential targets for cancer therapy. Among all detected splice junctions, the majority (94.1%) were known, while 38,193 (5.90%) were novel **(Fig. 5A).** Most transcripts identified with novel splice junctions contained only a single novel junction, with progressively fewer transcripts carrying multiple novel junctions. Notably, most novel junctions were supported by canonical splice sites (85.3%), while only a minority involved non-canonical splice sites (14.7%), indicating that novel splicing largely conformed to established splice site rules **(Fig. 5B).** Moreover, most novel junctions were derived from novel transcripts, with few originating from intergenic, genic, fusion, or antisense regions **(Fig. 5C)**.

To systematically identify tumor-specific neoantigens derived from these neojunctions, we developed a customized computational workflow, TS-SNAD. This pipeline identified novel splice junctions from tumor-specific novel transcripts by comparison with reference genome annotations, as well as excluding any splice junctions present in normal tissues from GTEx short-read RNA-seq and long-read RNA-seq data from 88 samples from GTEx tissues (Glinos et al. 2022), thereby reducing the likelihood of selecting peptides with potential expression in vital normal organs and minimizing off-tumor toxicity risk. Tumor specificity was further supported by significantly higher read coverage of

these novel splice junctions in NPC tumor samples compared to four immortalized nasopharyngeal epithelial cell lines in our cohort. The pipeline subsequently generated neojunction-spanning 9-mer peptides as candidate neoantigens. The binding affinity of these candidate neoantigens to HLA-I molecules was predicted using NetMHCpan, with an IC_{50} < 500 nM and a percentile rank < 0.5 set as selection criteria (**Fig. 5D**). To assess the binding potential of neojunction-derived neoantigens, we compiled HLA-I genotypes from 70 NPC patients from our in-house cohort with whole-genome sequencing data (Bruce et al. 2022) and predicted HLA-I types from 113 Chinese patients with RNA-seq data from the GSE102349 dataset. Despite the high polymorphism of HLA-I, both cohorts showed a consistent presence of high-frequency alleles, including HLA-A*02:07, HLA-A*11:01, HLA-B*46:01, HLA-B*40:01, HLA-C*01:02, and HLA-C*07:02 **(Supplemental Table S7**), likely reflecting the population-specific genetic background.

Using TS-SNAD, we identified 944 tumor-specific novel transcripts harboring novel splice junctions with read coverage enriched in tumor samples. From these, we generated a total of 3,326 neojunction-derived 9-mer neoantigens, of which 242 neojunction-derived 9-mer neoantigens matched peptide sequences encoded by other coding regions in the UniProtKB/Swiss-Prot human proteome. Of the 3,326 neojunction-derived 9-mer candidates, 533 candidate neoantigens were predicted to bind patient-specific HLA-I alleles with high affinity, suggesting a high likelihood of being presented (**Fig. 5E).** Notably, most of these neoantigens were derived from unique tumor-specific novel transcripts, while 70 were produced by multiple isoforms **(Fig. 5F).** Analyzing the diversity of HLA-I interactions, we found that most neoantigens could bind multiple HLA-I alleles **(Fig. 5G),** suggesting their potential to be presented across a broader patient population with diverse immune contexts, thereby enhancing their therapeutic relevance.

Neojunction-derived neoantigens effectively elicit immunogenic responses

To evaluate the potential of neojunction-derived neoantigens as targets for broadly applicable immunotherapies in NPC, we selected 34 neojunction-derived peptides with high predicted binding affinity for HLA-B*40:01, a high-frequency allele present in approximately 29.20% of NPC patients, and assessed their immunogenicity. Using peripheral blood mononuclear cells (PBMCs) from three

HLA-B*40:01-positive healthy donors, we measured interferon-γ (IFN-γ) responses following stimulation with peptide-MHC (pMHC) complexes. Six of the 34 predicted neojunction-derived neoantigens induced robust and reproducible IFN-γ responses in all three donors **(Fig. 6A, Supplemental Fig. S8A and Supplemental Table S8)**. The absence of response in DMSO controls and the robust activation induced by a CD3-specific antibody confirmed the specificity and reliability of the assay. The most immunogenic peptide, REPFVQAKL, was derived from a novel *TYK2* transcript (TCONS_00025200) generated through skipping of exons 7-9. We observed considerable inter-donor variability in the magnitude of IFN-γ responses, with donor 3 showing consistently stronger responses across all peptides, likely reflecting differences in individual immunological history.

Of note, each immunogenic neoantigen originated from a distinct tumor-specific novel transcript in our cohort**,** with the majority of underlying novel splice junctions resulting from exon skipping or cassette exon inclusion **(Fig. 6B and Supplemental Fig. S8B(a))**. The novel transcripts giving rise to these immunogenic neoantigens exhibited multiple features supporting their translational potential, including canonical splice sites, methionine-initiated ORFs, and no predicted susceptibility to nonsense-mediated mRNA decay. Analysis of public NPC datasets confirmed that these novel isoforms exhibited high tumor-specific expression (**Supplemental Fig. S8B(b)**), minimizing potential risks of off-target reactivity against control tissues. Neojunction-derived neoantigens largely recapitulated the canonical HLA-B*40:01 binding motif, with a dominant glutamate at P2 and hydrophobic residues at the C terminus **(Supplemental Fig. S8C).**

As each neoantigen was found to be present in only 10-20% of NPC patients, these results support the development of personalized multi-epitope vaccines tailored to individual splicing profiles to maximize immune targeting and overcome immune evasion. Furthermore, we confirmed the presence of neojunction-derived unique peptides in NPC samples via mass spectrometry **(Fig. 7A).** It is

noteworthy that junction-spanning peptides are typically underrepresented in mass spectrometry datasets due to the cleavage specificity of trypsin (Kahles et al. 2018), underscoring the significance of these findings. Additionally, to assess the protein-coding potential of these novel RNA isoforms that generated validated immunogenic neoantigens, we overexpressed their ORFs fused to a FLAG tag in both HEK293T and NP69 cells. Western blot analysis showed that all selected novel ORFs generated detectable proteins in vitro **(Fig. 7B).** Collectively, our results demonstrated that neojunction-derived neoantigens constituted a rich source of immunogenic targets, highlighting their strong potential for NPC immunotherapy development.

Discussion

In this study, we leveraged LR-seq to construct a comprehensive, isoform-level transcriptomic atlas of NPC, revealing widespread transcript diversity and novel splicing events that were largely undetectable by short-read sequencing. We validated the reliability of these full-length transcripts and demonstrated that previously unannotated isoforms, particularly those classified as NIC, NNC, and read-through chimeric transcripts, substantially contributed to the transcriptomic complexity of NPC. Integrated analysis with RNA-seq data revealed that many of these novel isoforms exhibited tumor-specific expression patterns, some of which were strongly associated with clinical outcomes, underscoring their potential as biomarkers and therapeutic targets. We selected a subset of tumor-specific novel transcripts for RT-qPCR validation using isoform-specific primers and observed a relatively high overall validation rate. The inability to validate a small subset of candidates was likely due to two practical factors. First, for genes with multiple isoforms sharing extensive exonic regions, isoform-level abundance estimates inferred from short-read RNA-seq using the LR-seq-derived transcriptome as a reference may be less accurate, which could have affected the reliable assessment of tumor specificity. Second, successful qPCR primer design was not always feasible, because it had to satisfy not only standard primer-design criteria, including appropriate GC content, minimal primer-dimer or hairpin formation, high specificity, and suitable amplicon length, but also the requirement to uniquely distinguish the isoform of interest from other highly similar transcripts. Therefore, failure to validate a small subset of candidates by qPCR did not necessarily indicate that these transcripts were

false positives, but more likely reflected technical limitations in isoform quantification and assay design.

In contrast to conventional neoantigens derived from gene mutations, we identified a rich source of tumor-specific antigens originating from neojunctions within novel transcripts, which were efficiently detected by LR-seq. These splicing alterations often disrupt canonical open reading frames, giving rise to novel protein products. To systematically harness this potential, we developed the TS-SNAD pipeline. TS-SNAD leverages long-read-resolved full-length transcript structures to more accurately identify novel isoforms and their associated novel splice junctions, providing a more reliable assessment of complex splicing events and their coding consequences than de novo assembly based on short-read RNA-seq or reference-guided transcript reconstruction. In addition, TS-SNAD prioritizes tumor-specific novel junctions by integrating group-wise junction coverage analysis and dual filtering against GTEx short-read and long-read normal tissue exon-exon junctions. Finally, it generates peptides that truly span novel junctions and further evaluates their HLA-binding potential, thereby providing an integrated framework for identifying splicing-derived neoantigens. Using TS-SNAD, we initially identified 3,326 tumor-specific neojunction-derived candidate neoantigens, 533 of which were predicted to bind with high affinity to specific NPC HLA-I alleles and were not present in the reference proteome. TS-SNAD thus provides an effective framework for prioritizing immunogenic neojunction-derived neoantigens from tumor-specific novel transcripts, significantly expanding the repertoire of potential targets for immunotherapy.

To advance the development of broadly applicable therapies, we summarized high-frequency HLA-I alleles within the East Asian NPC population, including HLA-A*02:07, HLA-A*11:01, HLA-B*46:01, HLA-B*40:01, HLA-C*01:02, and HLA-C*07:02. This stable HLA landscape provided a foundation for designing personalized yet population-compatible immunotherapeutic strategies. Focusing on HLA-B*40:01, we selected 34 predicted neoantigens for experimental validation. Ex vivo IFN-γ ELISpot assays confirmed that six of these elicited robust immune responses across multiple donors. These results provide further evidence that TS-SNAD can effectively identify immunogenic neojunction-derived neoepitopes.

Despite these encouraging results, several limitations should be acknowledged. First, the control group comprised immortalized nasopharyngeal epithelial cell lines rather than true normal nasopharyngeal epithelial tissues. Though these cell lines provide a practical epithelial control, they may not fully recapitulate the molecular features of native normal epithelium. In addition, publicly available bulk RNA-seq datasets from normal nasopharyngeal biopsies or adjacent non-malignant tissues are suboptimal for validating the differentially spliced events identified here, because these specimens contain very few nasopharyngeal epithelial cells and are dominated by stromal and immune cells. The lack of RNA-seq data from microdissected normal nasopharyngeal epithelial cells therefore limits direct validation of tumor-associated splicing changes in primary normal epithelial counterparts. Future studies using microdissected normal epithelium, epithelial-enriched samples, single-cell or long-read transcriptomic approaches will be important to further confirm the specificity of these splicing events. Second, we partially validated their immunogenicity using ex vivo IFN-γ ELISpot assays. Nevertheless, further in vitro and in vivo studies will be required to confirm antigen presentation and functional relevance in the tumor context. Third, while LR-seq accurately captures full-length isoforms, its lower throughput and sequencing depth relative to short-read RNA-seq may limit the detection of low-abundance transcripts. Finally, as our HLA frequency and immunogenicity data were primarily derived from Chinese cohorts, extrapolation to other ethnicities requires validation in more diverse populations.

In conclusion, our work highlights the critical value of integrating LR-seq into cancer transcriptomics research, not only for discovering novel isoforms and splicing events but also for unveiling a hidden landscape of functionally and clinically relevant neoantigens. The TS-SNAD pipeline establishes a foundation for rational neoantigen selection and paves the way for developing off-the-shelf or personalized cancer vaccines for NPC. Future directions include immunopeptidomics validation of predicted neoantigens, functional assessment of T cell responses, and integration with single-cell transcriptomics and proteogenomics to achieve a more comprehensive immunological understanding.

Methods

Long-read library preparation and sequencing

A total of 18 samples were collected and subjected to long-read sequencing, including four immortalized nasopharyngeal cell lines (NP361, NP460, NP550, and NP69) as controls, and 14 NPC samples comprising two cell lines (NPC43 and NPC53), 10 patient-derived xenografts (X2117, X17, C15, X111, X113, X666, X76, X47, X23, and X32), and two tumor tissues (NPC-T9 and NPC-T64). This sample set enabled transcript discovery across diverse biological sources. Following RNA extraction, full-length cDNA was synthesized from poly(A)-tailed RNA using the SMARTer PCR cDNA Synthesis Kit (Clontech/Takara; Cat. #634926). The cDNA was PCR-amplified to yield 1 to 2 μg of product and purified using AMPure XP magnetic beads (Beckman Coulter; A63881). Size selection was performed using the Sage Science BluePippin system to remove short cDNA fragments and reduce preferential loading and sequencing of shorter molecules. SMRTbell adapters were ligated to cDNA ends, and libraries were purified using magnetic beads with the SMRTbell Template Prep Kit (Pacific Biosciences), followed by sequencing on PacBio Sequel II platform.

RNA sequencing

RNA was extracted using the RNeasy Mini Prep kit (Qiagen; Cat. #74106), quantified on a Qubit 2.0 fluorometer (Thermo Fisher Scientific), and assessed for quality using a 2100 Bioanalyzer (Agilent) with the RNA 6000 Pico kit. cDNA libraries were prepared using Illumina TruSeq Stranded mRNA Library Prep Kit. The final libraries were sequenced on an Illumina NovaSeq 6000 platform to generate 150-bp paired-end reads.

Quantitative PCR with reverse transcription

Reverse transcription was performed with a High-Capacity RT kit (Applied Biosystems, Thermo Fisher Scientific; Cat#4368813). The resulting cDNA was diluted 1:3 using nuclease-free water, and 2 μL was used per quantitative PCR (qPCR) reaction. qPCR with reverse transcription was performed using the POWER SYBR Green Master Mix (Applied Biosystems, Thermo Fisher Scientific; Cat#4367659). All samples were analyzed in technical triplicates using the QuantStudio 5 (Thermo

Fisher Scientific), and all gene expression data were normalized to human *GAPDH*. Primer sequences were listed in **Supplemental Table S9**.

Long-read sequencing data analysis

Raw PacBio subreads were processed using the Iso-Seq3 pipeline (version 3.8.0) (https://github.com/PacificBiosciences/IsoSeq/) to generate high-quality, full-length transcript sequences for downstream analysis. Briefly, raw subreads were processed to generate CCS reads with a minimum predicted accuracy of 0.99. Full-length reads were generated by removing the 5′ and 3′ cDNA primers and performing demultiplexing using the 'lima' module with default parameters. Artificial concatemers and poly(A) tails were removed using 'isoseq3 refine' to produce full-length, non-chimeric (FLNC) reads. These FLNC reads were clustered into high-quality transcripts using 'isoseq3 cluster'. The filtered transcripts were aligned to the human reference genome (GRCh38/hg38) using pbmm2 (version 1.10.0) (https://github.com/PacificBiosciences/pbmm2). Redundant isoforms were collapsed within each sample using the collapse_isoforms_by_sam.py script from the cDNA_Cupcake ToFU toolkit (https://github.com/Magdoll/cDNA_Cupcake). Finally, transcripts from all samples were merged into a unified, non-redundant transcriptome using gffcompare (version 0.11.2) (Pertea and Pertea 2020) for subsequent analysis (see **Supplemental Fig. S1A**).

Isoform annotation and quality control

Full-length isoforms were annotated and subjected to quality control using SQANTI3 (version 5.2) (Pardo-Palacios et al. 2024) with GENCODE v38 as the reference annotation. SQANTI3 was used to annotate isoforms and generate quality-control metrics, including CAGE peaks (Noguchi et al. 2017), poly(A) sites (Moon et al. 2025), and NMD predictions, to help filter potential artifacts and ensure high-quality transcript annotations. Isoform validity was further supported by short-read RNA-seq data. The coverage of isoform splice junctions was calculated using the junction file (commonly referred to as SJ.out.tab) generated from RNA-seq data. ORFs were predicted with ORFfinder (https://github.com/Chokyotager/ORFFinder), retaining only those encoding ≥100 amino acids.

De novo transcriptome assembly

Sorted BAM files (hg38-aligned) were assembled *de novo* using StringTie (version 2.1.1) (Pertea et al. 2015), with parameters: -m 200 -f 0.05 -c 1.5 -p 10. Individual assemblies from the NPC samples were merged with StringTie --merge with default parameters. Both the *de novo* and the LR-seq-derived transcriptomes were compared to GENCODE v38 using gffcompare -G.

Short-read RNA sequencing analysis

The quality control of short-read RNA-seq data was performed using FastQC (version 0.11.9). The low-quality reads and adapters were filtered and trimmed by using cutadapt (version 4.4) (Kechin et al. 2017). Clean reads were mapped to the human reference genome hg38 using STAR (version 2.7.10b) (Dobin et al. 2013). Specifically, a two-pass strategy was used to extract splice junctions from each sample. The TPM abundances of all isoforms were calculated using RSEM (version 1.3.1) (Li and Dewey 2011) using the LR-seq-derived transcriptome as reference. Public RNA-seq data from different cohorts were integrated, followed by batch effect removal using ComBat from the R sva package.

Alternative splicing event analysis

We used SUPPA2 (version 2.3) (Trincado et al. 2018) to calculate seven types of alternative splicing events including A3, A5, AF, AL, SE, RI, and MX. To incorporate novel transcripts, a unified GTF file (merged from GENCODE v38 and long-read-derived novel isoforms) was used with the generateEvents command (-f ioe). Differential splicing analysis was performed with SUPPA2 diffSplice, with significance thresholds set at $|\Delta PSI| > 0.1$ and $P < 0.05$. To improve accuracy, splice sites were held to stringent boundaries, while transcription start/end sites (AF/AL events) allowed flexible boundaries (±48 nt).

Pathway enrichment analysis

Pathways significantly enriched (false discovery rate [FDR] < 0.05) among genes with novel isoforms detected by LR-seq were identified using the MSigDB C2, C5, and Hallmark collections (Liberzon et al. 2015). Enrichment analyses were performed using the R package clusterProfiler (version 3.16.1)

(Wu et al. 2021). A matched background gene set was used for all enrichment tests and was defined as the intersection of (i) genes detectable in our long-read data and (ii) genes that could be mapped to the selected MSigDB collections. This background ensured that enrichment was evaluated relative to genes testable in our experiment and reduced potential bias.

To control for potential bias arising from gene expression levels, we performed an expression-matched permutation analysis. For each gene, expression level was defined as the median TPM across samples within the expressed gene universe. Genes were stratified into quantile-based expression bins (10 bins by default). For the observed gene set generating novel isoforms, we recorded the number of genes in each expression bin and generated matched random gene sets with the same expression-level distribution and repeated this procedure across 200 permutations. Pathway enrichment analysis was then performed for each randomized gene set. For each pathway significantly enriched in the genes with novel isoforms, an empirical p value (p_{emp}) was calculated as the fraction of expression-matched permuted gene sets whose p value (p_{perm}) was less than or equal to the observed p value (p_{obs}) obtained for the gene set with novel isoforms:

$$p_{\text{emp}} = \frac{1 + \sum_{b=1}^{B} 1\left(p_{\text{perm}}^{(b)} \leq p_{\text{obs}}\right)}{B + 1}$$

where B denotes the number of permutations. This analysis provided a direct test of whether the pathway enrichments observed in genes generating novel isoforms exceeded what would be expected based solely on their expression-level distribution. Pathways with empirical $p_{\text{emp}} < 0.05$ were considered significant.

Differential transcript expression analysis

RSEM expected read counts were used as input for downstream differential expression analysis. Differential transcript expression between NPC tumor samples and immortalized nasopharyngeal cell lines was assessed in DESeq2 (v1.38.3) (Love et al. 2014) using the Wald test under a negative binomial generalized linear model, with Benjamini-Hochberg correction for multiple testing.

Transcripts with $|\log_2 FC| > 2$ and adjusted $P < 0.05$ were considered significantly differentially expressed.

Identification of transcripts encoding putative cell-surface transmembrane proteins

An isoform was classified as encoding a putative cell-surface transmembrane protein if its ORF contained at least one transmembrane helix predicted by DeepTMHMM (Hallgren et al. 2022) and was predicted by DeepLoc (Odum et al. 2024) to localize to the cell membrane. For external validation, we used the TCSA Cancer Surfaceome Atlas as a gene-level reference set of genes encoding cell-surface proteins. TCSA integrates nine independent resources to define a curated cancer surfaceome. Because currently available databases generally lack isoform-resolved subcellular localization annotations, we could not reliably assign cell-surface localization to a single canonical isoform for every gene. Therefore, the gain or loss of predicted cell-surface localization was assessed at the isoform level relative to the annotated protein products of the corresponding gene.

Identification of splicing-derived neoantigens by TS-SNAD

Tumor specificity was assessed by integrating long-read and short-read RNA-seq data, as the sequencing depth of our long-read dataset was insufficient for robust transcript-level abundance estimation. Long-read RNA-seq was used primarily to construct a LR-seq-derived transcriptome by resolving full-length transcript structures and identifying both annotated and previously unannotated isoforms, whereas short-read RNA-seq was used to quantify transcript expression based on the LR-seq-derived transcriptome. Tumor-specific transcripts were defined as transcripts with TPM ≥ 1 in at least one tumor sample and not detected in any of four immortalized NP cell lines in our cohort. Splice junction coverage was quantified using SJ.out.tab files generated by STAR alignment of RNA-seq data to the hg38 genome, with the LR-seq-derived transcriptome annotation provided as a reference. For neojunction prioritization, all novel splice junctions derived from tumor-specific novel transcripts were subjected to further filtering. Candidate neojunctions were required to meet the following criteria: (i) the average junction read coverage in tumor samples was higher than that in immortalized non-malignant nasopharyngeal epithelial cell lines; (ii) <5 reads supporting the novel splice junctions in each immortalized NP cell line; and (iii) <5 reads supporting the novel splice junctions in each of 11

non-malignant nasopharyngeal tissue samples from three public NPC RNA-seq datasets (GSE118719, GSE134886, and GSE68799); (iv) the novel splice junctions were required to be absent from all GTEx datasets, including short-read RNA-seq from normal tissues and Oxford Nanopore long-read data from 88 GTEx tissues and cell lines.

For neoantigen prediction, 9-mer peptides spanning neojunctions were extracted. NetMHCpan (version 4.0) (Jurtz et al. 2017) was then used with default parameters to predict HLA binding affinity, using neojunction-derived 9-mer peptides and patient-specific HLA genotypes as input. Only candidate neoantigen peptides with strong predicted HLA-I binding were retained, defined as those with a binding affinity percentile rank < 0.5 and predicted IC_{50} < 500 nM.

HLA-I haplotype prediction

HLA-I genotypes for 70 patients from our local cohort were determined from our previous whole-genome sequencing data and subsequently used for neoantigen binding affinity prediction. For public datasets, HLA-I genotypes were predicted from RNA-seq FASTQ files using OptiType (version 1.3.3) (Szolek et al. 2014). We selected OptiType based on its demonstrated high accuracy (over 99%) in predicting HLA-I alleles from RNA-seq data.

IFN-γ enzyme-linked immunospot (ELISpot) assay

The immunogenicity of each neoantigen peptide was evaluated using an IFN-γ ELISpot assay. Briefly, PBMCs from donors carrying the HLA-B*40:01 allele were isolated from whole blood using Ficoll-Paque (Cytiva, 10294426) density gradient centrifugation. The collected PBMCs were then resuspended in RPMI-1640 (Gibco, 25367098) complete medium supplemented with 10% fetal bovine serum (Gibco, 24341305) and incubated at 37°C in a humidified incubator with 5% CO□ to allow cell recovery and stabilization. Subsequently, 3 × 10□ PBMCs in 100 μL were seeded into each well of a 96-well ELISpot plate (Mabtech, 3420-4AST-2), followed by the addition of 2 μL of peptide stimulator (final concentration 20 μg/mL). Anti-CD3 monoclonal antibody was used as a positive control, and medium containing an equivalent concentration of DMSO served as a negative control. Plates were then incubated at 37°C in a humidified 5% CO□ incubator for at least 20 hours without disturbance. Measures were taken to prevent evaporation during incubation. After incubation, cells

were removed, and the plates were washed five times with PBS. A biotinylated anti-IFN-γ detection antibody (1:1000 dilution) was added to each well and incubated for 2 hours at 37°C. After another round of washing, streptavidin-ALP (1:1000) was added and incubated for 1 hour at room temperature. Finally, BCIP/NBT substrate solution was applied to develop colorimetric spots representing IFN-γ-secreting cells. Spot-forming units (SFUs) were quantified using an automated ELISpot reader (AID iSpot Spectrum). Responses were considered positive if spot counts exceeded background by at least threefold.

Cell cultures, plasmids and transfection

NP69 cells were cultured in Keratinocyte SFM (1X) Kit (Gibco, 17005042), supplemented with 1% penicillin-streptomycin (10,000 U/mL, Gibco, 15140122). HEK293T cells were maintained in DMEM (high glucose, Gibco, 11965118) supplemented with 10% FBS (Sigma-Aldrich, F2442-500ML) and 1% penicillin-streptomycin. All cell lines were cultured at 37 °C with 5% CO_2 and were routinely tested and confirmed to be free of mycoplasma contamination.

ORFs of the novel transcripts, followed by a 3×FLAG tag, were cloned into the pcDNA3.1(+) vector, and all plasmid constructs were verified by sequencing. HEK293T cells were transiently transfected using Lipofectamine 3000 reagent (Invitrogen, L3000001) according to the manufacturer's instructions. NP69 cells were transfected with FuGENE HD (Promega, E2311) according to the manufacturer's instructions.

Reagents and antibodies for immunoblotting

The following antibodies were obtained from the indicated sources and used for immunoblotting: anti-FLAG M2 mouse mAb (Sigma-Aldrich, F1804) and anti-β-actin rabbit antibody (ABclonal AC038). Secondary antibodies were obtained from the following sources: HRP-conjugated goat anti-mouse IgG (H+L) (Invitrogen, 31430) and HRP-conjugated goat anti-rabbit IgG (H+L) (ABclonal, AS014).

Immunofluorescence

Cells grown on coverslips were fixed with 4% paraformaldehyde for 10 min at room temperature and washed with PBS. Fixed, non-permeabilized cells were then incubated with wheat germ agglutinin-

AF488 (MCE, HY-NP163A) in PBS for 10 min at room temperature to label the plasma membrane. Following WGA staining, cells were washed with PBS and subsequently permeabilized with 0.2% Triton X-100 in PBS for 5 min. Cells were then blocked with 5% BSA in PBS for 60 min at room temperature. Cells were incubated with anti-FLAG M2 mouse monoclonal antibody (Sigma-Aldrich, F1804) overnight, followed by incubation with goat anti-mouse IgG (H+L) secondary antibody (Invitrogen, A11004) for 60 min. Nuclei were stained with DAPI. Coverslips were mounted using antifade mounting medium, and images were captured using a Leica STELLARIS 8 FALCON confocal microscope.

Proteomic profiling and peptide database search

Public NPC mass spectrometry data were obtained from the iProX partner repository (Accession No. IPX0001265000). Peptide identification was performed using MS-GF+ (version 2018.10.15) (Kim and Pevzner 2014) against a custom sequence database comprising a total of 42,613 merged ORFs, including 22,017 ORFs derived from long-read novel transcripts and 20,596 reviewed human protein sequences from UniProtKB/Swiss-Prot database. The following parameters were set for database searching: Carbamidomethyl (C), iTRAQ4plex (N-term), and iTRAQ4plex (K) were specified as fixed modifications; Oxidation (M), Deamidated (NQ), Acetyl (K), and Methyl (K) were specified as variable modifications. The precursor mass tolerance was set to 20 ppm, and the product ion tolerance for MS/MS was 0.05 Da. Trypsin was specified as the digestion enzyme, allowing up to two missed cleavages. mzID profiles identified from the search engine were then pooled using the R/Bioconductor package MSnbase, and peptide-to-spectrum matches (PSMs) satisfying both spectra and peptide false discovery rate cutoffs <1% were retained for further analysis. To validate neojunction-derived neoantigens by mass spectrometry, only unique 9-mer peptides with unambiguous spectral matches were considered.

Survival analysis

Survival analysis was performed to evaluate the prognostic associations of transcript expression levels in NPC patients. Patients were stratified into high- and low-expression groups based on the median

transcript levels. Kaplan-Meier survival curves were generated and compared using the log-rank test with the survival (version 3.3) and survminer (version 0.4.9) R packages.

Data Sets

Public RNA-seq datasets from NPC cohorts were obtained from the Gene Expression Omnibus (GEO) with accession numbers GSE118719, GSE134886, GSE68799 and GSE102349. RNA-seq splice junction data from normal human tissues were retrieved from the GTEx project (https://www.gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression). Long-read RNA-seq data for 88 GTEx tissue and cell lines, based on the Oxford Nanopore Technologies platform, were also retrieved from the GTEx project (https://www.gtexportal.org/home/downloads/adult-gtex/long_read_data). Public mass spectrometry proteomic data of NPC were downloaded from the iProX partner repository (number IPX0001265000).

Ethics Declarations

For the two NPC tumor samples, written patient consents were obtained from the patients according to institutional clinical research approval. The study protocol was approved by the Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee at the Chinese University of Hong Kong, Hong Kong SAR. The collection and use of PBMCs from healthy donors were approved under the protocol “Collection of peripheral blood mononuclear cells for new drug research and development” (protocol No. LP202210). Written informed consent was obtained from all PBMC donors before sample collection.

Data Access

Long-read RNA sequencing data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA1333213.

Code availability

All code required to reproduce the data processing and downstream analyses in this study has been provided as Supplemental Code files and archived in a publicly accessible GitHub repository. Within

this repository, the analysis scripts are available at https://github.com/CityUHK-CompBio/NPC-LR-Seq/tree/main/script, whereas the TS-SNAD pipeline is available at https://github.com/CityUHK-CompBio/NPC-LR-Seq/tree/main/TS-SNAD. The TS-SNAD directory includes the full pipeline implementation, example input and output files, parameter descriptions, and a step-by-step usage guide.

Competing interest statement

The authors have no conflicts of interest to declare.

Acknowledgments

This work was funded by grants from Shenzhen Medical Research Funds (C2303002, X.W.), a startup grant (4937084, X.W.), a direct grant (2024.175, X.W.), grants by the Faculty Postdoctoral Fellowship Scheme (FPFS/24-25/053, FPFS/23-24/061C, FPFS/23-24/060, X.W.), Research Committee - Group Research Scheme 2022-23 (WW/rc/grs2223/0560/23en, X.W.), and Faculty Innovation Award (FIA2020/A/01, C.M.T.) from the Chinese University of Hong Kong, and grants from the Research Grants Council (AoE/M-401/20 to X.W. & K.W.L.; R4007-23, C4024-22GF, 14104223, 11103921, and 14111522 to X.W.; 14101721 to K.W.L.; 14124925, 14116124, 14113620, 14114523, and 24114922 to C.M.T.), The Innovation and Technology Fund (MRP/036/21X, K.W.L.), and Health and Medical Research Fund (08192166 to X.W.; 08191046 to K.W.L; 09203176 to C.M.T.). This work was also partially sponsored by Shenzhen Bay Scholars Program and Jiangxi Overseas High-Level Talent Project (20232BCJ25029) awarded to X.W. The authors gratefully acknowledge the valuable contributions of all members of Prof. Xin Wang's and Prof. K.W. Lo’s laboratory, who provided insightful feedback and shared their expertise in conducting the experiments and performing the bioinformatics computations.

Author contributions

X.W., Y.S., and K.W.L. conceived the project and designed the research. Y.S. conducted the bioinformatics analyses. H.L.Y., B.W., G.T.Y.C., M.Z. and Y.S. performed the experimental

validation. Y.S. prepared the figures and tables and drafted the manuscript. X.W., Y.S., K.W.L., Z.X.Z., C.S.L., and X.G.W. interpreted the data and discussed the results. X.W. and K.W.L. revised the manuscript. X.W., K.W.L., and C.M.T. provided funding and material support. X.W. and K.W.L. supervised the study. All authors have read and approved the final version of the manuscript.

References

Alpert T, Straube K, Oesterreich FC, Herzel L, Neugebauer KM. 2020. Widespread Transcriptional Readthrough Caused by Nab2 Depletion Leads to Chimeric Transcripts with Retained Introns. *Cell Rep* **33**: 108496.

Bardia A, Rugo HS, Tolaney SM, Loirat D, Punie K, Oliveira M, Brufsky A, Kalinsky K, Cortes J, Shaughnessy JO et al. 2024. Final Results From the Randomized Phase III ASCENT Clinical Trial in Metastatic Triple-Negative Breast Cancer and Association of Outcomes by Human Epidermal Growth Factor Receptor 2 and Trophoblast Cell Surface Antigen 2 Expression. *J Clin Oncol* **42**: 1738-1744.

Bhattacharya A, Vo DD, Jops C, Kim M, Wen C, Hervoso JL, Pasaniuc B, Gandal MJ. 2023. Isoform-level transcriptome-wide association uncovers genetic risk mechanisms for neuropsychiatric disorders in the human brain. *Nat Genet* **55**: 2117-2128.

Boixareu C, Taha T, Venkadakrishnan VB, de Bono J, Beltran H. 2025. Targeting the tumour cell surface in advanced prostate cancer. *Nat Rev Urol* doi:10.1038/s41585-025-01014-w.

Borgers JSW, Lenkala D, Kohler V, Jackson EK, Linssen MD, Hymson S, McCarthy B, O'Reilly Cosgrove E, Balogh KN, Esaulova E et al. 2025. Personalized, autologous neoantigen-specific T cell therapy in metastatic melanoma: a phase 1 trial. *Nat Med* **31**: 881-893.

Bradley RK, Anczukow O. 2023. RNA splicing dysregulation and the hallmarks of cancer. *Nat Rev Cancer* **23**: 135-155.

Bruce JP, To KF, Lui VWY, Chung GTY, Chan YY, Tsang CM, Yip KY, Ma BBY, Woo JKS, Hui EP et al. 2022. Author Correction: Whole-genome profiling of nasopharyngeal carcinoma reveals viral-host co-operation in inflammatory NF-kappaB activation and immune escape. *Nat Commun* **13**: 4353.

706 Burbage M, Rocanin-Arjo A, Baudon B, Arribas YA, Merlotti A, Rookhuizen DC, Heurtebise-
707 Chretien S, Ye M, Houy A, Burgdorf N et al. 2023. Epigenetically controlled tumor antigens
708 derived from splice junctions between exons and transposable elements. *Sci Immunol* **8**:
709 eabm6360.
710 Cazes A, Childers BG, Esparza E, Lowy AM. 2022. The MST1R/RON Tyrosine Kinase in Cancer:
711 Oncogenic Functions and Therapeutic Strategies. *Cancers (Basel)* **14**.
712 Charlet A, Kappenstein M, Keye P, Klasener K, Endres C, Poggio T, Gorantla SP, Kreutmair S,
713 Sanger J, Illert AL et al. 2022. The IL-3, IL-5, and GM-CSF common receptor beta chain
714 mediates oncogenic activity of FLT3-ITD-positive AML. *Leukemia* **36**: 701-711.
715 Chow LK-Y, Chung DL-S, Tao L, Chan KF, Tung SY, Ngan RKC, Ng WT, Lee AW-M, Yau CC,
716 Kwong DL-W. 2022. Epigenomic landscape study reveals molecular subtypes and EBV-
717 associated regulatory epigenome reprogramming in nasopharyngeal carcinoma. *EBioMedicine*
718 **86**.
719 Dai W, Zheng H, Cheung AK, Tang CS, Ko JM, Wong BW, Leong MM, Sham PC, Cheung F,
720 Kwong DL et al. 2016. Whole-exome sequencing identifies MST1R as a genetic susceptibility
721 gene in nasopharyngeal carcinoma. *Proc Natl Acad Sci U S A* **113**: 3317-3322.
722 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR.
723 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.
724 Du W, Liu X, Yang M, Wang W, Sun J. 2021. The Regulatory Role of PRRX1 in Cancer Epithelial-
725 Mesenchymal Transition. *Onco Targets Ther* **14**: 4223-4229.
726 Dunn GP, Sherpa N, Manyanga J, Johanns TM. 2022. Considerations for personalized neoantigen
727 vaccination in Malignant glioma. *Adv Drug Deliv Rev* **186**: 114312.
728 Gallego-Paez LM, Edwards WJS, Chanduri M, Guo Y, Koorman T, Lee CY, Grexa N, Derksen P,
729 Yan J, Schwartz MA et al. 2023. TLN1 contains a cancer-associated cassette exon that alters
730 talin-1 mechanosensitivity. *J Cell Biol* **222**.
731 Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, Dai X, Aguet F, Brown KL,
732 Garimella K et al. 2022. Transcriptome variation in human tissues revealed by long-read
733 sequencing. *Nature* **608**: 353-359.

734 Hallgren J, Tsirigos KD, Pedersen MD, Almagro Armenteros JJ, Marcatili P, Nielsen H, Krogh A,
735 Winther O. 2022. DeepTMHMM predicts alpha and beta transmembrane proteins using deep
736 neural networks. *bioRxiv* doi:10.1101/2022.04.08.487609: 2022.2004.2008.487609.
737 Hong M, Tao S, Zhang L, Diao LT, Huang X, Huang S, Xie SJ, Xiao ZD, Zhang H. 2020. RNA
738 sequencing: new technologies and applications in cancer research. *J Hematol Oncol* **13**: 166.
739 Hook PW, Timp W. 2023. Beyond assembly: the increasing flexibility of single-molecule sequencing
740 technology. *Nat Rev Genet* **24**: 627-641.
741 Hu W, Wu Y, Shi Q, Wu J, Kong D, Wu X, He X, Liu T, Li S. 2022. Systematic characterization of
742 cancer transcriptome at transcript resolution. *Nat Commun* **13**: 6803.
743 Hu Z, Guo X, Li Z, Meng Z, Huang S. 2024. The neoantigens derived from transposable elements - A
744 hidden treasure for cancer immunotherapy. *Biochim Biophys Acta Rev Cancer* **1879**: 189126.
745 Hu Z, Yuan J, Long M, Jiang J, Zhang Y, Zhang T, Xu M, Fan Y, Tanyi JL, Montone KT et al. 2021.
746 The Cancer Surfaceome Atlas integrates genomic, functional and drug response data to
747 identify actionable targets. *Nat Cancer* **2**: 1406-1422.
748 Huang KK, Huang J, Wu JKL, Lee M, Tay ST, Kumar V, Ramnarayanan K, Padmanabhan N, Xu C,
749 Tan ALK et al. 2021. Long-read transcriptome sequencing reveals abundant promoter
750 diversity in distinct molecular subtypes of gastric cancer. *Genome Biol* **22**: 44.
751 Joglekar A, Hu W, Zhang B, Narykov O, Diekhans M, Marrocco J, Balacco J, Ndhlovu LC, Milner
752 TA, Fedrigo O et al. 2024. Single-cell long-read sequencing-based mapping reveals
753 specialized splicing patterns in developing and adult mouse and human brain. *Nat Neurosci* **27**:
754 1051-1063.
755 Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. 2017. NetMHCpan-4.0: Improved
756 Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding
757 Affinity Data. *J Immunol* **199**: 3360-3368.
758 Kahles A, Lehmann KV, Toussaint NC, Huser M, Stark SG, Sachsenberg T, Stegle O, Kohlbacher O,
759 Sander C, Cancer Genome Atlas Research N et al. 2018. Comprehensive Analysis of
760 Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell* **34**: 211-224 e216.

761 Kam NW, Lau CY, Lau JYH, Dai X, Liang Y, Lai SPH, Chung MKY, Yu VZ, Qiu W, Yang M et al.
762 2025. Cell-associated galectin 9 interacts with cytotoxic T cells confers resistance to tumor
763 killing in nasopharyngeal carcinoma through autophagy activation. *Cell Mol Immunol* **22**:
764 260-281.
765 Kechin A, Boyarskikh U, Kel A, Filipenko M. 2017. cutPrimers: A New Tool for Accurate Cutting of
766 Primers from Reads of Targeted Next Generation Sequencing. *J Comput Biol* **24**: 1138-1143.
767 Kim S, Pevzner PA. 2014. MS-GF+ makes progress towards a universal database search tool for
768 proteomics. *Nat Commun* **5**: 5277.
769 Klibi J, Niki T, Riedel A, Pioche-Durieu C, Souquere S, Rubinstein E, Le Moulec S, Guigay J,
770 Hirashima M, Guemira F et al. 2009. Blood diffusion and Th1-suppressive effects of galectin-
771 9-containing exosomes released by Epstein-Barr virus-infected nasopharyngeal carcinoma
772 cells. *Blood* **113**: 1957-1966.
773 Kumar H, Luo R, Wen J, Yang C, Zhou X, Kim P. 2024. FusionNeoAntigen: a resource of fusion
774 gene-specific neoantigens. *Nucleic Acids Res* **52**: D1276-D1288.
775 Lauss M, Donia M, Harbst K, Andersen R, Mitra S, Rosengren F, Salim M, Vallon-Christersson J,
776 Torngren T, Kvist A et al. 2020. Author Correction: Mutational and putative neoantigen load
777 predict clinical benefit of adoptive T cell therapy in melanoma. *Nat Commun* **11**: 1714.
778 Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without
779 a reference genome. *BMC Bioinformatics* **12**: 323.
780 Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular
781 Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**: 417-425.
782 Lo AK, Dawson CW, Lung HL, Wong KL, Young LS. 2021. The Role of EBV-Encoded LMP1 in the
783 NPC Tumor Microenvironment: From Function to Therapy. *Front Oncol* **11**: 640207.
784 Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq
785 data with DESeq2. *Genome Biol* **15**: 550.
786 Lu SX, De Neef E, Thomas JD, Sabio E, Rousseau B, Gigoux M, Knorr DA, Greenbaum B, Elhanati
787 Y, Hogg SJ et al. 2021. Pharmacologic modulation of RNA splicing enhances anti-tumor
788 immunity. *Cell* **184**: 4032-4047 e4031.

789 Maurin M, Ranjouri M, Megino-Luque C, Newberg JY, Du D, Martin K, Miner RE, 3rd, Prater MS,
790 Wee DKB, Centeno B et al. 2023. RBFOX2 deregulation promotes pancreatic cancer
791 progression and metastasis through alternative splicing. *Nat Commun* **14**: 8444.
792 Merlotti A, Sadacca B, Arribas YA, Ngoma M, Burbage M, Goudot C, Houy A, Rocanin-Arjo A,
793 Lalanne A, Seguin-Givelet A et al. 2023. Noncanonical splicing junctions between exons and
794 transposable elements represent a source of immunogenic recurrent neo-antigens in patients
795 with lung cancer. *Sci Immunol* **8**: eabm6359.
796 Monzo C, Liu T, Conesa A. 2025. Transcriptomics in the era of long-read sequencing. *Nat Rev Genet*
797 doi:10.1038/s41576-025-00828-z.
798 Moon Y, Herrmann CJ, Mironov A, Zavolan M. 2025. PolyASite v3.0: a multi-species atlas of
799 polyadenylation sites inferred from single-cell RNA-sequencing data. *Nucleic Acids Res* **53**:
800 D197-D204.
801 Naro C, Antonioni A, Medici V, Caggiano C, Jolly A, de la Grange P, Bielli P, Paronetto MP, Sette C.
802 2024. Splicing targeting drugs highlight intron retention as an actionable vulnerability in
803 advanced prostate cancer. *J Exp Clin Cancer Res* **43**: 58.
804 Neri P, Leblay N, Lee H, Gulla A, Bahlis NJ, Anderson KC. 2024. Just scratching the surface: novel
805 treatment approaches for multiple myeloma targeting cell membrane proteins. *Nat Rev Clin*
806 *Oncol* **21**: 590-609.
807 Noguchi S, Arakawa T, Fukuda S, Furuno M, Hasegawa A, Hori F, Ishikawa-Kato S, Kaida K, Kaiho
808 A, Kanamori-Katayama M. 2017. FANTOM5 CAGE profiles of human and mouse samples.
809 *Scientific data* **4**: 170112.
810 Odum MT, Teufel F, Thumuluri V, Almagro Armenteros JJ, Johansen AR, Winther O, Nielsen H.
811 2024. DeepLoc 2.1: multi-label membrane protein type prediction using protein language
812 models. *Nucleic Acids Res* **52**: W215-W220.
813 Ott PA, Hu-Lieskovan S, Chmielowski B, Govindan R, Naing A, Bhardwaj N, Margolin K, Awad
814 MM, Hellmann MD, Lin JJ et al. 2020. A Phase Ib Trial of Personalized Neoantigen Therapy
815 Plus Anti-PD-1 in Patients with Advanced Melanoma, Non-small Cell Lung Cancer, or
816 Bladder Cancer. *Cell* **183**: 347-362 e324.

817 Pan Y, Phillips JW, Zhang BD, Noguchi M, Kutschera E, McLaughlin J, Nesterenko PA, Mao Z,
818 Bangayan NJ, Wang R et al. 2023. IRIS: Discovery of cancer immunotherapy targets arising
819 from pre-mRNA alternative splicing. *Proc Natl Acad Sci U S A* **120**: e2221116120.
820 Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomas J, Amorin R,
821 Estevan-Morio E, Liu T, Nanni A, McIntyre L et al. 2024. SQANTI3: curation of long-read
822 transcriptomes for accurate identification of known and novel isoforms. *Nat Methods* **21**: 793-
823 797.
824 Pertea G, Pertea M. 2020. GFF Utilities: GffRead and GffCompare. *F1000Res* **9**.
825 Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables
826 improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290-295.
827 Platten M, Bunse L, Wick A, Bunse T, Le Cornet L, Harting I, Sahm F, Sanghvi K, Tan CL, Poschke I
828 et al. 2021. A vaccine targeting mutant IDH1 in newly diagnosed glioma. *Nature* **592**: 463-
829 468.
830 Sethna Z, Guasp P, Reiche C, Milighetti M, Ceglia N, Patterson E, Lihm J, Payne G, Lyudovyk O,
831 Rojas LA et al. 2025. RNA neoantigen vaccines prime long-lived CD8(+) T cells in pancreatic
832 cancer. *Nature* **639**: 1042-1051.
833 Siak PY, Heng WS, Teoh SSH, Lwin YY, Cheah SC. 2023. Precision medicine in nasopharyngeal
834 carcinoma: comprehensive review of past, present, and future prospect. *J Transl Med* **21**: 786.
835 Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, Fugmann T, Wong KK, Van
836 Allen EM. 2018. Intron retention is a source of neoepitopes in cancer. *Nat Biotechnol* **36**:
837 1056-1058.
838 Somalraju S, Salem DH, Janga SC. 2025. Investigating RNA dynamics from single molecule
839 transcriptomes. *Trends Genet* doi:10.1016/j.tig.2025.05.001.
840 Supek F, Lehner B, Lindeboom RGH. 2021. To NMD or Not To NMD: Nonsense-Mediated mRNA
841 Decay in Cancer and Other Genetic Diseases. *Trends Genet* **37**: 657-668.
842 Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. 2014. OptiType: precision HLA
843 typing from next-generation sequencing data. *Bioinformatics* **30**: 3310-3316.

844 Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, Eyras E. 2018. SUPPA2: fast,
845 accurate, and uncertainty-aware differential splicing analysis across multiple conditions.
846 *Genome Biol* **19**: 40.
847 Tsang CM, Lui VWY, Bruce JP, Pugh TJ, Lo KW. 2020. Translational genomics of nasopharyngeal
848 cancer. *Semin Cancer Biol* **61**: 84-100.
849 Wang X, Yu L, Zhou X, Chung GT, Liu AM, Chan YY, Wu M, Chau KY, Lo KW, Wu AR. 2025.
850 Characterizing resistant cellular states in nasopharyngeal carcinoma during EBV lytic
851 induction. *Oncogene* **44**: 1805-1819.
852 Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L et al. 2021.
853 clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)*
854 **2**: 100141.
855 Xie H, Zhang K, Yin H, Zhang S, Pan S, Wu R, Han Y, Xu Y, Jiang W, You B. 2025.
856 Acetyltransferase NAT10 inhibits T-cell immunity and promotes nasopharyngeal carcinoma
857 progression through DDX5/HMGB1 axis. *J Immunother Cancer* **13**.
858 Xie N, Shen G, Gao W, Huang Z, Huang C, Fu L. 2023. Neoantigens: promising targets for cancer
859 therapy. *Signal Transduct Target Ther* **8**: 9.
860 Zaidi S, Park J, Chan JM, Roudier MP, Zhao JL, Gopalan A, Wadosky KM, Patel RA, Sayar E,
861 Karthaus WR et al. 2024. Single-cell analysis of treatment-resistant prostate cancer:
862 Implications of cell state changes for cell surface antigen-targeted therapies. *Proc Natl Acad*
863 *Sci U S A* **121**: e2322203121.
864 Zelba H, Shao B, Rabsteyn A, Reinhardt A, Greve C, Oenning L, Kayser S, Kyzirakos C, Latzer P,
865 Riedlinger T et al. 2025. In-depth characterization of vaccine-induced neoantigen-specific T
866 cells in patients with IDH1-mutant glioma undergoing personalized peptide vaccination. *J*
867 *Immunother Cancer* **13**.

868

Figure legends

Fig. 1. Profiling of full-length transcripts in nasopharyngeal carcinoma by long-read sequencing. (A) Pie chart illustrating the distribution of full-length transcripts across different subcategories. A total of 51,115 full-length transcripts were identified across 18 long-read sequencing samples, with the numbers and proportions of different transcript subcategories highlighted in different colors. (B) Bar plots showing LR-seq isoforms detected in individual NPC tumor samples and immortalized NP cell lines, colored by the subcategories defined in (A) and grouped by sample origin. (C) Histograms comparing length distributions of long-read transcripts between different subcategories. (D) Bar plot showing the number of transcripts identified per gene by LR-seq. (E) Boxplot comparing the proportions of long-read transcripts supported by short-read Illumina RNA-seq among different subcategories in matched samples.

Fig. 2. Characterization of full-length transcripts identified in LR-seq. (A) Barplot comparing the coding potential, predicted NMD status, and splice site canonicality between transcript subcategories. (B) Boxplots comparing isoform length, CDS length, and junction coverage across FSM, NIC and NNC transcript categories. (C) Dot plot showing pathways significantly enriched among genes with novel transcripts detected by LR-seq (FDR < 0.05), based on the MSigDB C2, C5, and Hallmark collections. Enrichment remained significant after controlling for gene expression using expression-matched permutation analysis. (D) Dot plot illustrating the numbers of novel LR-seq transcripts compared with annotated GENCODE isoforms detected in selected oncogenes (left) and tumor suppressors (right). (E) Pie chart showing the proportion of novel transcripts sharing the same translation start site as annotated protein isoforms. NS, not significant; *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001.

Fig. 3. Novel transcripts contribute to isoform-level diversity and differential alternative splicing events. (A) Schematic illustration of the seven types of AS events, including A5: Alternative 5’ splice site, A3: Alternative 3’ splice site, AF: Alternative first exon, AL: Alternative last exon, MX: Mutually exclusive exons, RI: Retained intron, and SE: Skipping exon. (B) Bar plot comparing alternative splicing events before and after incorporating novel transcripts into GENCODE v38

transcriptome. (C) Dot plot showing differential alternative splicing events between NPC tumor samples and immortalized NP cell lines, identified by SUPPA2 with $P < 0.05$ and $|\Delta PSI| > 0.1$. (D) Schematic illustrations of an example of a differential alternative splicing event involving the inclusion of a cassette exon (17b) in *TLN1* and the resulting transcripts. (E) Bar plot comparing differential isoform usage of *TLN1* (left), with and without exon 17b, between NPC samples and immortalized NP cell lines, validated by RT-qPCR (right). Each RT-qPCR experiment was repeated three times with independent RT preparations, with the standard deviation of each triplicate indicated by the error bars.

Fig. 4. Identification and validation of tumor-specific transcripts. (A) Density curves comparing transcript expression across different transcript subcategories. (B) Bar plot showing the numbers of FSM, NIC and NNC transcripts across differential median expression thresholds. (C) Volcano plot illustrating differential transcript expression analysis between NPC tumor samples and immortalized NP cell lines (BH-adjusted $P < 0.05$ and $|\log_2 FC| > 2$). (D) Ring charts showing the classification of transcripts according to their expression specificity. (E) Heatmap showing the expression of a subset of predicted tumor-specific novel transcripts identified in LR-seq. (F) Bar plots depicting RT-qPCR validation results of transcripts in (E). RT-qPCR assays were carried out in three independent RT replicates, with error bars representing the standard deviation.

Fig. 5. Identification of potential neojunction-derived neoantigens from tumor-specific novel transcripts using TS-SNAD. (A) Pie chart showing the distribution of splice junction categories detected by LR-seq, with known splice junctions (blue) and novel splice junctions (yellow). (B) Bar plot showing the number of transcripts with different numbers of novel splice junctions. Inset pie chart depicting the proportion of canonical and non-canonical splice sites. (C) Bar plot showing the distribution of transcripts containing novel splice junctions, categorized as NNC, NIC, intergenic, genic, fusion, or antisense. (D) Schematic overview of the TS-SNAD computational workflow for identifying neojunction-derived neoantigens from tumor-specific novel transcripts. (E) Ring plot showing that a total of 3,326 neojunction-derived neoantigens were identified, of which 533 were predicted to bind strongly to at least one NPC HLA class I allele. (F) Bar plot showing the distribution of 533 strong-binding neoantigens by the number of tumor-specific novel transcripts from which they

924 were derived. (G) Bar plot showing the number of distinct HLA class I alleles bound by the 533
925 neoantigens.

926 **Fig. 6. Neojunction-derived neoantigens effectively elicit immunogenic responses.** (A) ELISpot
927 assay results demonstrating that six out of the 34 splicing-derived neoantigens elicited immunogenic
928 responses. PBMCs from three independent HLA-B*40:01 donors were stimulated with peptides (20
929 μg/mL), and IFN-γ secretion was measured by ELISpot (n = 3). DMSO served as the negative control,
930 and anti-CD3 mAb as the positive control. Data are presented as mean ± SD. (B) Schematic
931 illustrations of the exon-intron structures of the novel transcripts from which the six immunogenic
932 neoantigens were derived.

933 **Fig. 7. Proteomic and protein expression validation of ORFs encoding immunogenic**
934 **neojunction-derived neoantigens in NPC.** (A) Mass spectrometry provided protein-level evidence
935 supporting the existence of neojunction-derived neoantigens. (B) Western blot analysis showing the
936 expression of FLAG-tagged ORFs encoding the validated immunogenic neojunction-derived
937 neoantigens in HEK293T and NP69 cells after transfection.

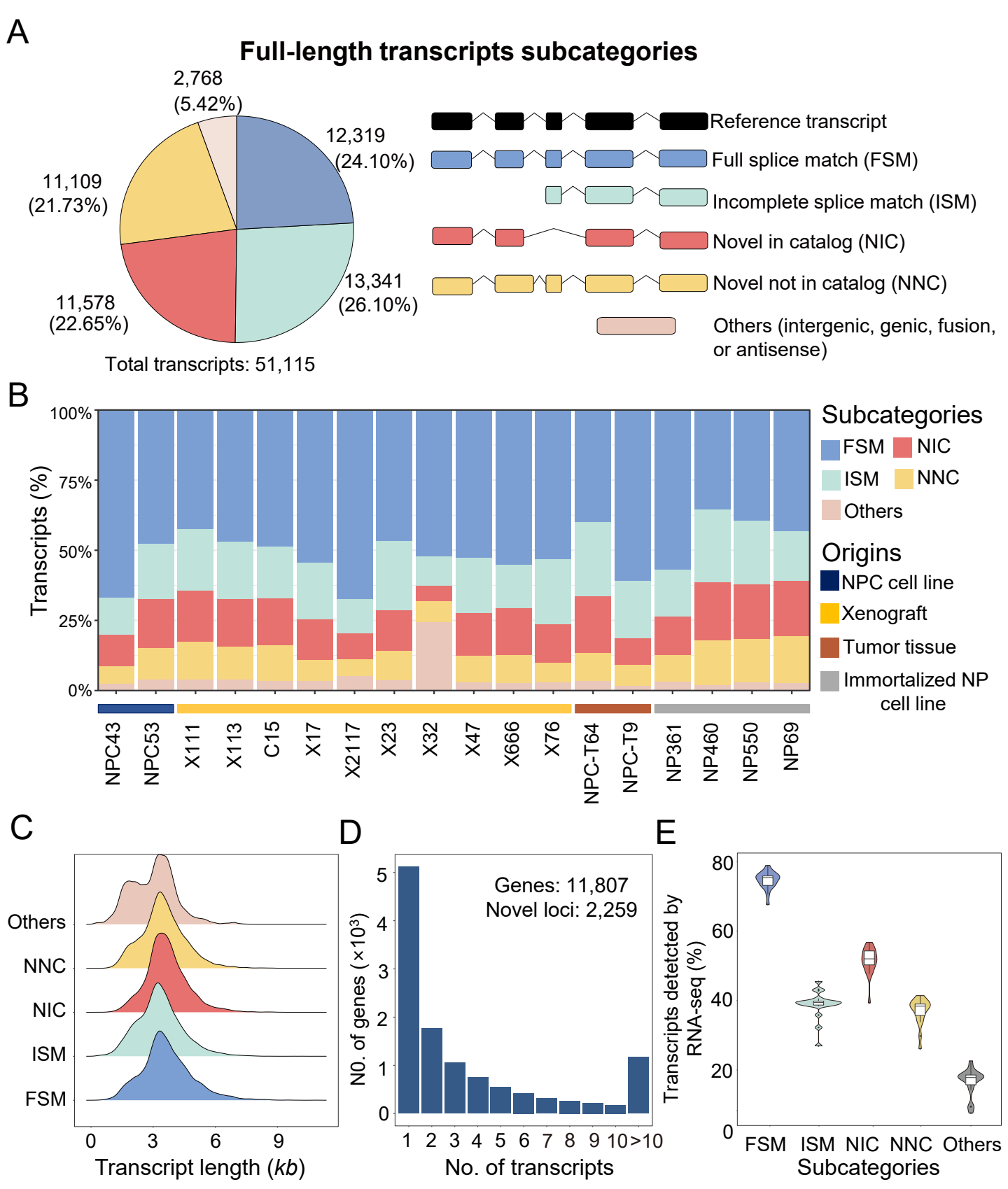
938

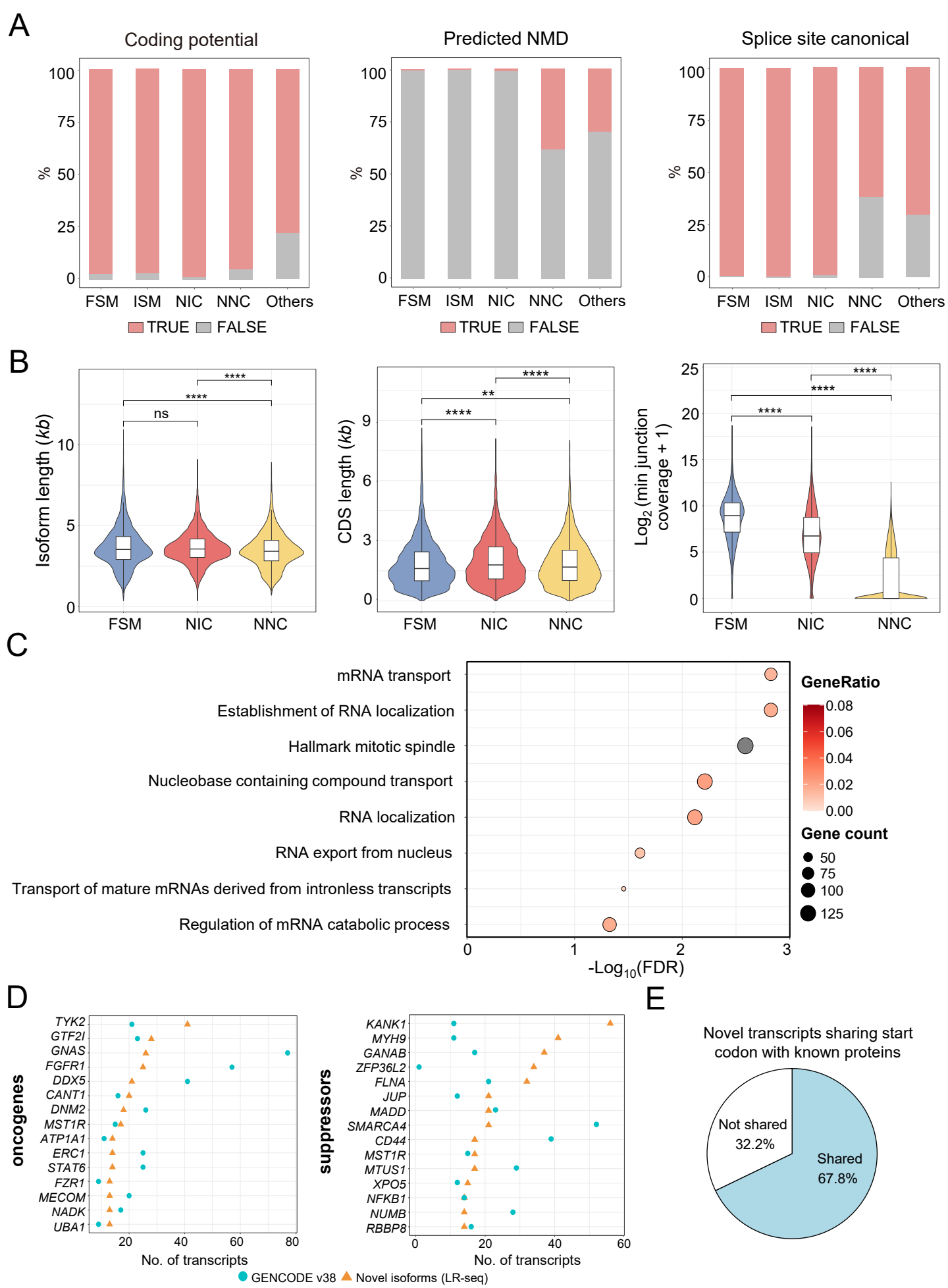
939 # Tables
940

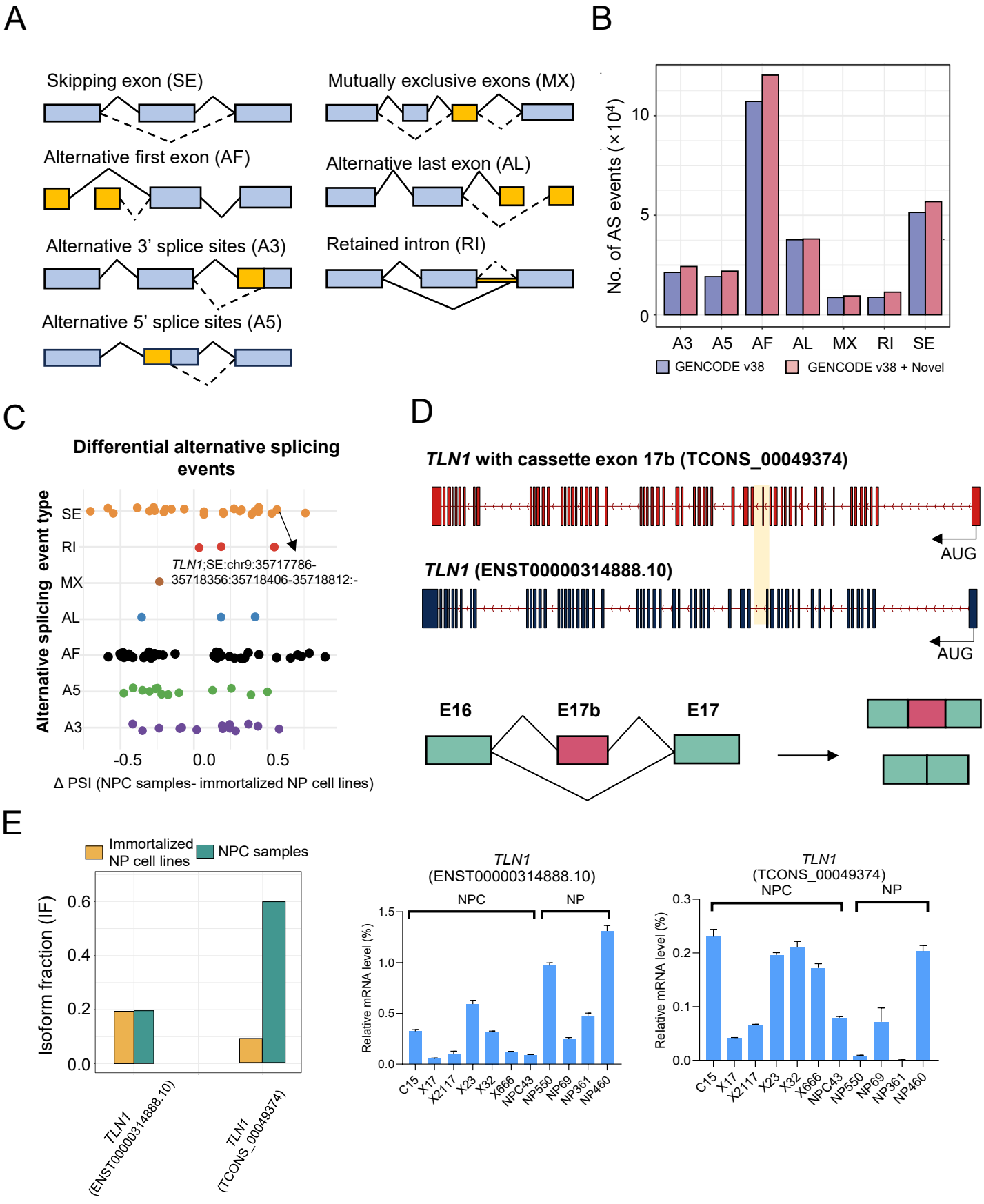
Table 1. NPC/immortalized NP cell lines for long-read sequencing

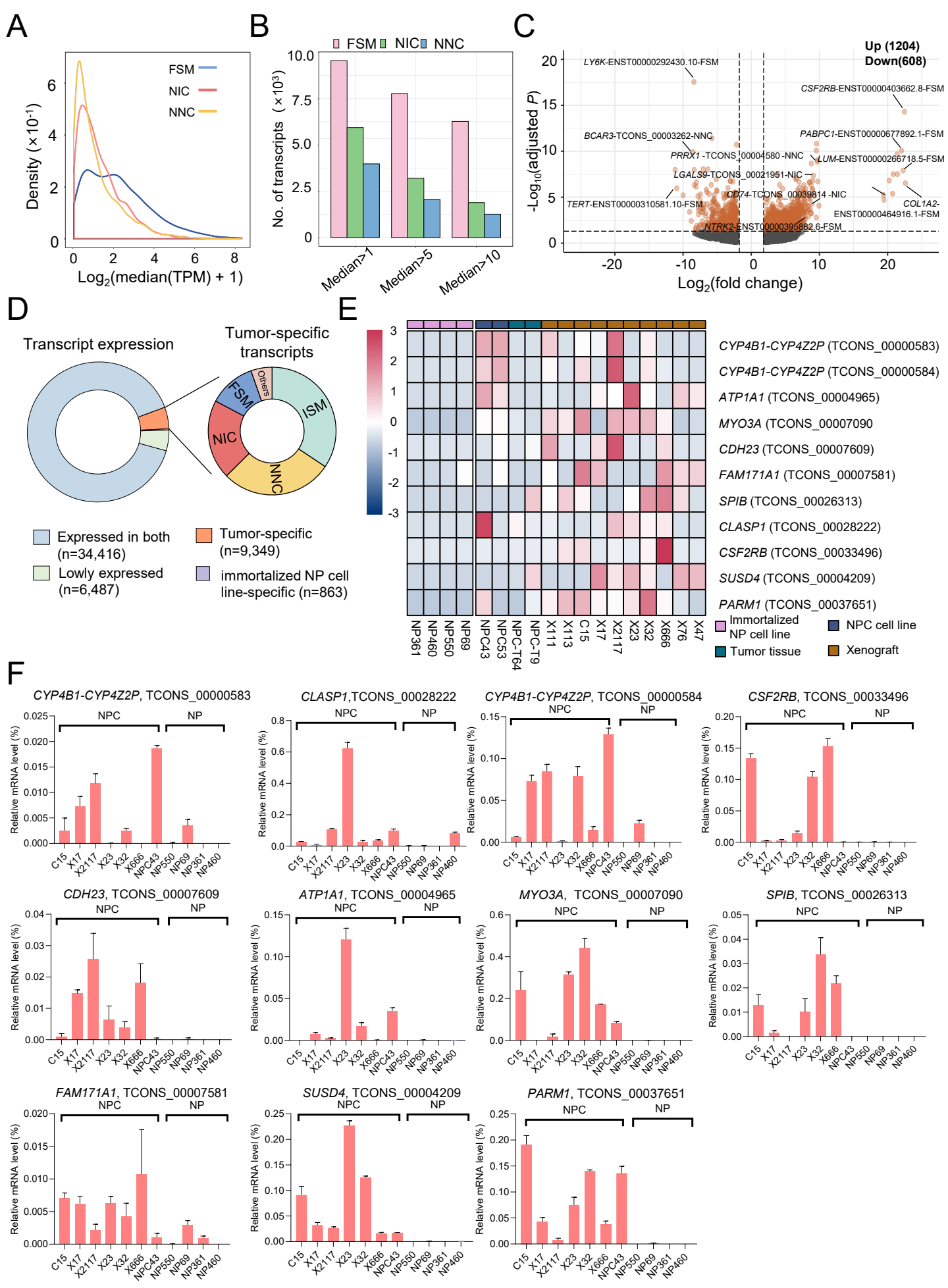
Sample Name	Sample Type	EBV Infection
NPC43	NPC cell line	Yes
NPC53	NPC cell line	No
NP361	Immortalized nasopharyngeal epithelial (NP) cell lines	No
NP460	Immortalized nasopharyngeal epithelial (NP) cell lines	No
NP550	Immortalized nasopharyngeal epithelial (NP) cell lines	No
NP69	Immortalized nasopharyngeal epithelial (NP) cell lines	No
X111	Patient-derived tumor xenograft model	Yes
X113	Patient-derived tumor xenograft model	Yes
C15	Patient-derived tumor xenograft model	Yes
X17	Patient-derived tumor xenograft model	Yes
X2117	Patient-derived tumor xenograft model	Yes
X23	Patient-derived tumor xenograft model	Yes
X32	Patient-derived tumor xenograft model	Yes
X47	Patient-derived tumor xenograft model	Yes
X666	Patient-derived tumor xenograft model	Yes
X76	Patient-derived tumor xenograft model	Yes
NPC-T64	Tumor tissue	Yes
NPC-T9	Tumor tissue	Yes

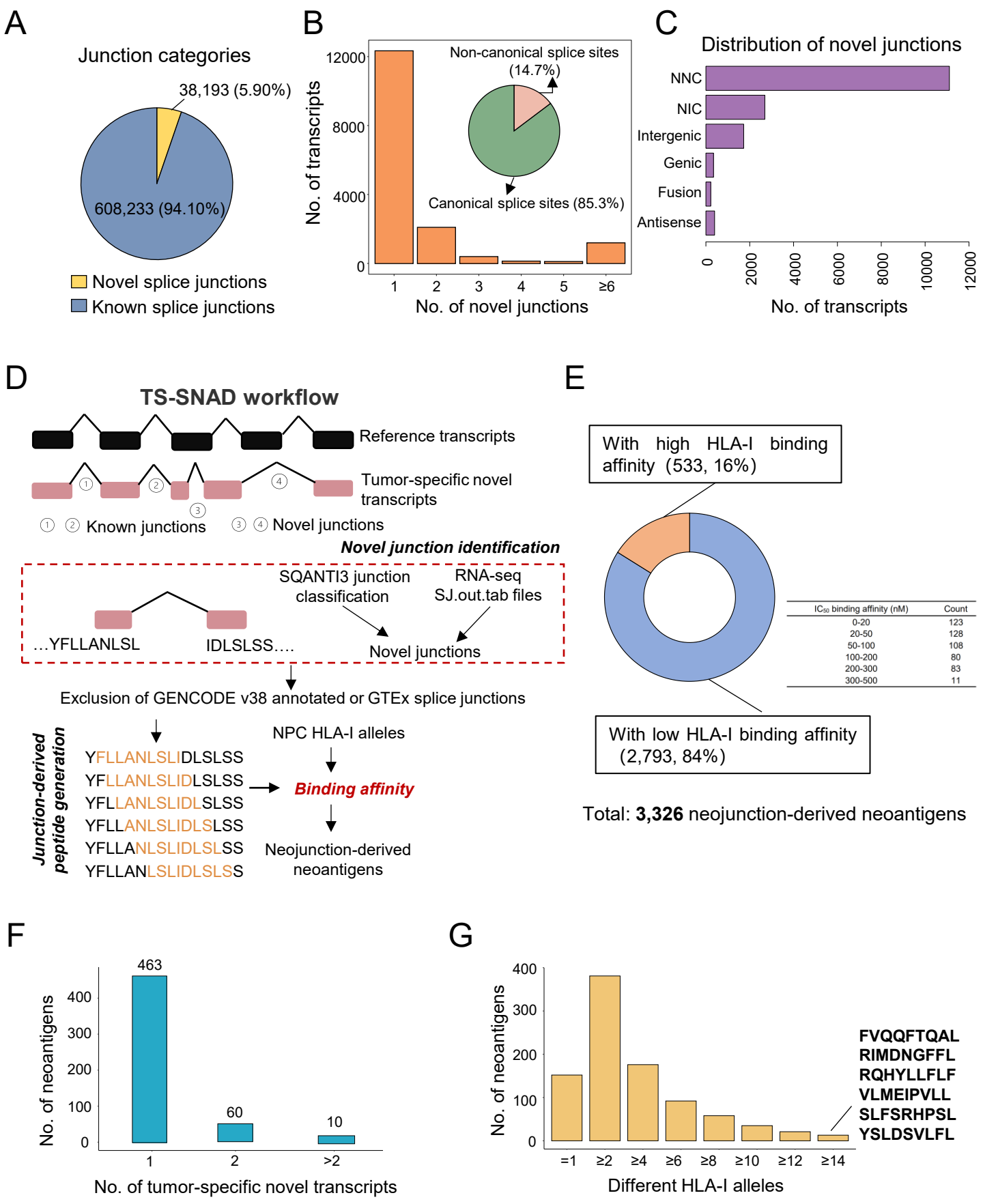
941

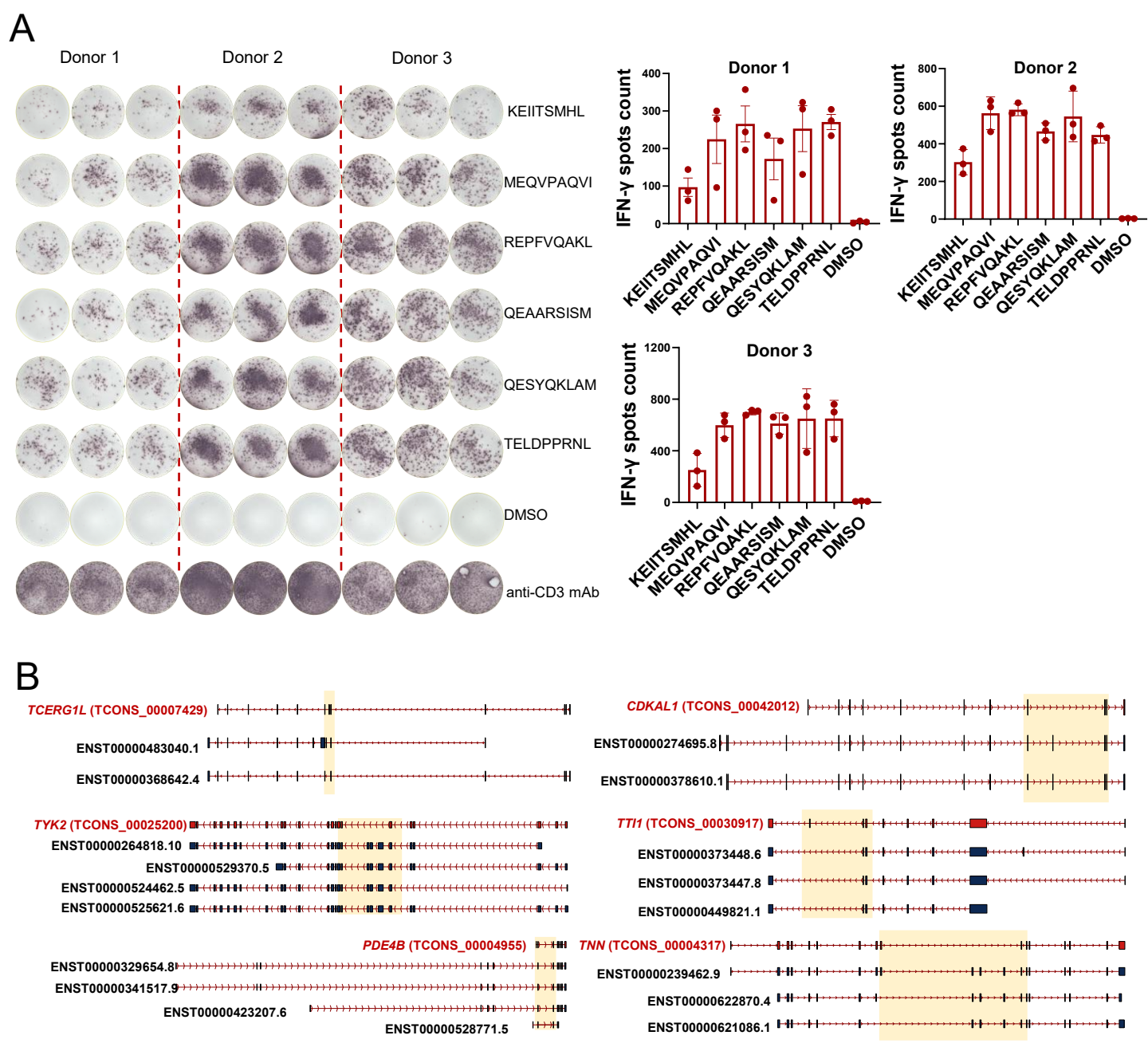

A
Full-length transcripts subcategories
2,768
(5.42%)
12,319
(24.10%)
11,109
(21.73%)
13,341
(26.10%)
11,578
(22.65%)
Total transcripts: 51,115
Reference transcript
Full splice match (FSM)
Incomplete splice match (ISM)
Novel in catalog (NIC)
Novel not in catalog (NNC)
Others (intergenic, genic, fusion, or antisense)
B
Transcripts (%)
100%
75%
50%
25%
0%
Subcategories
FSM
NIC
ISM
NNC
Others
Origins
NPC cell line
Xenograft
Tumor tissue
Immortalized NP cell line
NPC43
NPC53
X111
X113
C15
X17
X2117
X23
X32
X47
X666
X76
NPC-T64
NPC-T9
NP361
NP460
NP550
NP69
C
Others
NNC
NIC
ISM
FSM
0
3
6
9
Transcript length (kb)
D
N0. of genes (×10³)
Genes: 11,807
Novel loci: 2,259
1 2 3 4 5 6 7 8 9 10 >10
No. of transcripts
E
Transcripts detetcted by RNA-seq (%)
80
60
40
20
0
FSM ISM NIC NNC Others
Subcategories


A
Coding potential
Predicted NMD
Splice site canonical
%
FSM ISM NIC NNC Others
TRUE FALSE
B
Isoform length (kb)
CDS length (kb)
Log2 (min junction coverage + 1)
ns
****
**
FSM NIC NNC
C
mRNA transport
Establishment of RNA localization
Hallmark mitotic spindle
Nucleobase containing compound transport
RNA localization
RNA export from nucleus
Transport of mature mRNAs derived from intronless transcripts
Regulation of mRNA catabolic process
-Log10(FDR)
GeneRatio
Gene count
D
oncogenes
TYK2 GTF2I GNAS FGFR1 DDX5 CANT1 DNM2 MST1R ATP1A1 ERC1 STAT6 FZR1 MECOM NADK UBA1
suppressors
KANK1 MYH9 GANAB ZFP36L2 FLNA JUP MADD SMARCA4 CD44 MST1R MTUS1 XPO5 NFKB1 NUMB RBBP8
No. of transcripts
GENCODE v38
Novel isoforms (LR-seq)
E
Novel transcripts sharing start codon with known proteins
Not shared 32.2%
Shared 67.8%

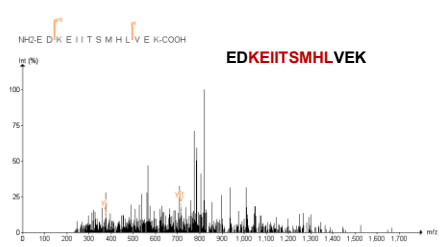

A
Skipping exon (SE)
Mutually exclusive exons (MX)
Alternative first exon (AF)
Alternative last exon (AL)
Alternative 3' splice sites (A3)
Retained intron (RI)
Alternative 5' splice sites (A5)
B
No. of AS events (×10⁴)
10
5
0
A3
A5
AF
AL
MX
RI
SE
GENCODE v38
GENCODE v38 + Novel
C
Differential alternative splicing events
Alternative splicing event type
SE
RI
MX
AL
AF
A5
A3
TLN1;SE:chr9:35717786-35718356:35718406-35718812:-
-0.5
0.0
0.5
Δ PSI (NPC samples- immortalized NP cell lines)
D
TLN1 with cassette exon 17b (TCONS_00049374)
AUG
TLN1 (ENST00000314888.10)
AUG
E16
E17b
E17
E
Immortalized NP cell lines
NPC samples
Isoform fraction (IF)
0.6
0.4
0.2
0.0
TLN1 (ENST00000314888.10)
TLN1 (TCONS_00049374)
TLN1 (ENST00000314888.10)
NPC
NP
Relative mRNA level (%)
1.5
1.0
0.5
0.0
C15
X17
X2117
X23
X32
X666
NPC43
NP550
NP69
NP361
NP460
TLN1 (TCONS_00049374)
NPC
NP
Relative mRNA level (%)
0.3
0.2
0.1
0.0
C15
X17
X2117
X23
X32
X666
NPC43
NP550
NP69
NP361
NP460


A
FSM
NIC
NNC
Density (×10-1)
Log2(median(TPM) + 1)
B
FSM NIC NNC
No. of transcripts (×10³)
Median>1
Median>5
Median>10
C
Up (1204)
Down(608)
LY6K-ENST00000292430.10-FSM
CSF2RB-ENST00000403662.8-FSM
BCAR3-TCONS_00003262-NNC
PABPC1-ENST00000677892.1-FSM
PRRX1 -TCONS_00004580 -NNC
LUM-ENST00000266718.5-FSM
LGALS9-TCONS_00021951-NIC
CD74-TCONS_00039814 -NIC
TERT-ENST00000310581.10-FSM
COL1A2-ENST00000464916.1-FSM
NTRK2-ENST00000395882.6-FSM
-Log10(adjusted P)
Log2(fold change)
D
Transcript expression
Tumor-specific transcripts
FSM
Others
ISM
NIC
NNC
Expressed in both (n=34,416)
Tumor-specific (n=9,349)
Lowly expressed (n=6,487)
immortalized NP cell line-specific (n=863)
E
CYP4B1-CYP4Z2P (TCONS_00000583)
CYP4B1-CYP4Z2P (TCONS_00000584)
ATP1A1 (TCONS_00004965)
MYO3A (TCONS_00007090)
CDH23 (TCONS_00007609)
FAM171A1 (TCONS_00007581)
SPIB (TCONS_00026313)
CLASP1 (TCONS_00028222)
CSF2RB (TCONS_00033496)
SUSD4 (TCONS_00004209)
PARM1 (TCONS_00037651)
NP361 NP460 NP550 NP69 NPC43 NPC53 NPC-T64 NPC-T9 X111 X113 C15 X17 X2117 X23 X32 X666 X76 X47
Immortalized NP cell line
NPC cell line
Tumor tissue
Xenograft
F
CYP4B1-CYP4Z2P, TCONS_00000583
CLASP1,TCONS_00028222
CYP4B1-CYP4Z2P, TCONS_00000584
CSF2RB, TCONS_00033496
CDH23, TCONS_00007609
ATP1A1, TCONS_00004965
MYO3A, TCONS_00007090
SPIB, TCONS_00026313
FAM171A1, TCONS_00007581
SUSD4, TCONS_00004209
PARM1, TCONS_00037651
NPC
NP
Relative mRNA level (%)
C15 X17 X2117 X23 X32 X666 NPC43 NP550 NP69 NP361 NP460

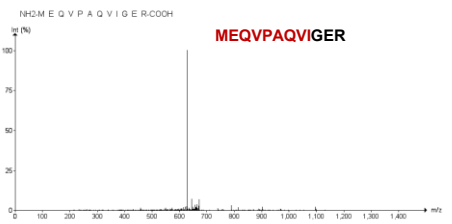

A
Junction categories
38,193 (5.90%)
608,233 (94.10%)
Novel splice junctions
Known splice junctions
B
No. of transcripts
12000
8000
4000
0
Non-canonical splice sites (14.7%)
Canonical splice sites (85.3%)
1
2
3
4
5
≥6
No. of novel junctions
C
Distribution of novel junctions
NNC
NIC
Intergenic
Genic
Fusion
Antisense
0
2000
4000
6000
8000
10000
12000
No. of transcripts
D
TS-SNAD workflow
Reference transcripts
Tumor-specific novel transcripts
① ② Known junctions
③ ④ Novel junctions
Novel junction identification
SQANTI3 junction classification
RNA-seq SJ.out.tab files
…YFLLANLSL
IDLSLSS….
Novel junctions
Exclusion of GENCODE v38 annotated or GTEx splice junctions
Junction-derived peptide generation
YFLLANLSLIDLSLSS
YFLLANLSLIDLSLSS
YFLLANLSLIDLSLSS
YFLLANLSLIDLSLSS
YFLLANLSLIDLSLSS
YFLLANLSLIDLSLSS
NPC HLA-I alleles
Binding affinity
Neojunction-derived neoantigens
E
With high HLA-I binding affinity (533, 16%)
With low HLA-I binding affinity (2,793, 84%)
IC50 binding affinity (nM) | Count
0-20 | 123
20-50 | 128
50-100 | 108
100-200 | 80
200-300 | 83
300-500 | 11
Total: 3,326 neojunction-derived neoantigens
F
No. of neoantigens
463
60
10
1
2
>2
No. of tumor-specific novel transcripts
G
No. of neoantigens
FVQQFTQAL
RIMDNGFFL
RQHYLLFLF
VLMEIPVLL
SLFSRHPSL
YSLDSVLFL
=1
≥2
≥4
≥6
≥8
≥10
≥12
≥14
Different HLA-I alleles


A
Donor 1
Donor 2
Donor 3
KEIITSMHL
MEQVPAQVI
REPFVQAKL
QEAARSISM
QESYQKLAM
TELDPPRNL
DMSO
anti-CD3 mAb
IFN-γ spots count
Donor 1
Donor 2
Donor 3
KEIITSMHL
MEQVPAQVI
REPFVQAKL
QEAARSISM
QESYQKLAM
TELDPPRNL
DMSO
B
TCERG1L (TCONS_00007429)
ENST00000483040.1
ENST00000368642.4
CDKAL1 (TCONS_00042012)
ENST00000274695.8
ENST00000378610.1
TYK2 (TCONS_00025200)
ENST00000264818.10
ENST00000529370.5
ENST00000524462.5
ENST00000525621.6
TTI1 (TCONS_00030917)
ENST00000373448.6
ENST00000373447.8
ENST00000449821.1
PDE4B (TCONS_00004955)
ENST00000329654.8
ENST00000341517.9
ENST00000423207.6
ENST00000528771.5
TNN (TCONS_00004317)
ENST00000239462.9
ENST00000622870.4
ENST00000621086.1

A

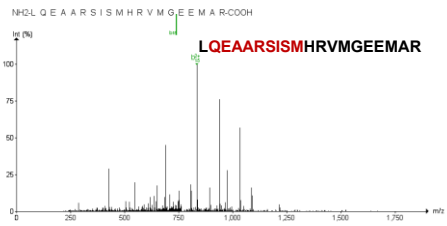


TCERG1L (TCONS_00007429)



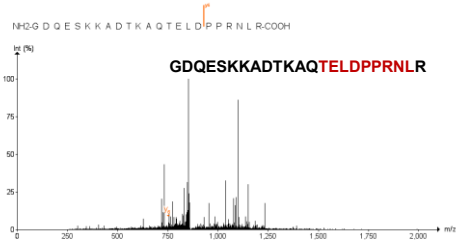
CDKAL1 (TCONS_00042012)

TYK2 (TCONS_ 00025200)



TTI1 (TCONS_00030917)

PDE4B (TCONS_00004955)

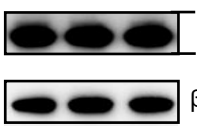


TNN (TCONS_00004317)

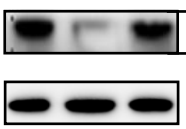
B

HEK293T

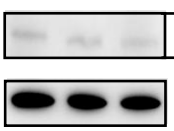
TCERG1L (TCONS_00007429)



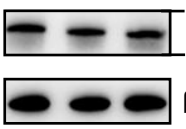
CDKAL1 (TCONS_00042012)



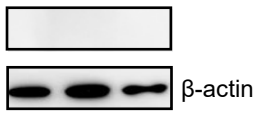
TYK2 (TCONS_00025200)



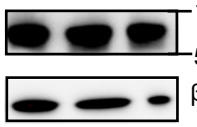
TTI1 (TCONS_00030917)



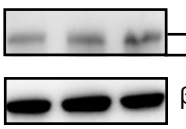
NC



PDE4B (TCONS_00004955)

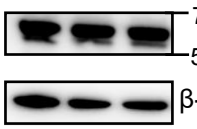


TNN (TCONS_00004317)

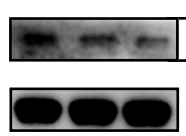


NP69

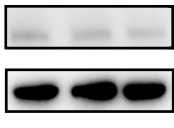
TCERG1L (TCONS_00007429)



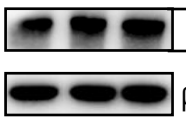
CDKAL1 (TCONS_00042012)



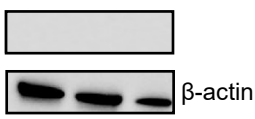
TYK2 (TCONS_00025200)



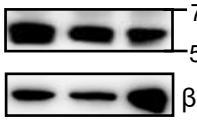
TTI1 (TCONS_00030917)



NC



PDE4B (TCONS_00004955)



TNN (TCONS_00004317)