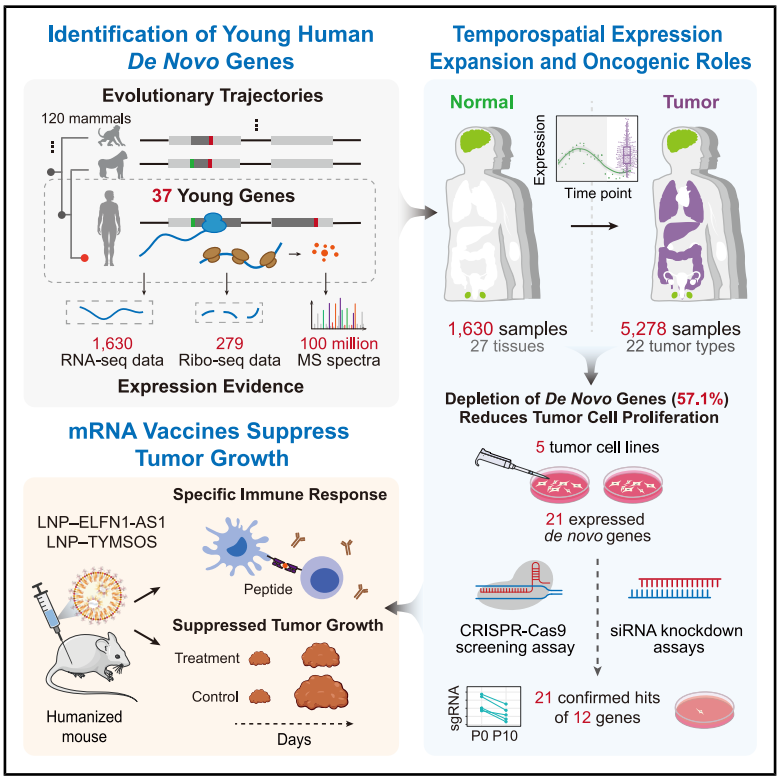
Cell Genomics

Article

Oncogenic roles of young human *de novo* genes and their potential as neoantigens in cancer immunotherapy

Graphical abstract



Highlights

- 37 young human *de novo* genes with clear evolutionary trajectories are identified
- These genes show temporospatial expression expansion across tumors
- Depletion of 57.1% of these genes suppresses tumor cell proliferation
- mRNA vaccines expressing two young genes trigger specific immune responses

Authors

Chunfu Xiao, Xiaoge Liu, Peiyu Liu, ..., Qiang Cheng, Ni A. An, Chuan-Yun Li

Correspondence

qiangcheng@pku.edu.cn (Q.C.), annie@genetics.ac.cn (N.A.A.), chuanyunli@genetics.ac.cn (C.-Y.L.)

In brief

Xiao et al. report 37 young *de novo* genes in humans within an updated genomic context. Their expression is temporospatially expanded across tumors. Depleting 57.1% of them reduces tumor cell proliferation. mRNA vaccines expressing two of these young genes trigger specific immune responses and inhibit tumor growth in humanized mice.

Xiao et al., 2025, Cell Genomics *5*, 100928
September 10, 2025
https://doi.org/10.1016/j.xgen.2025.100928

CellPress

Please cite this article in press as: Xiao et al., Oncogenic roles of young human *de novo* genes and their potential as neoantigens in cancer immunotherapy, Cell Genomics (2025), https://doi.org/10.1016/j.xgen.2025.100928

Cell Genomics

CellPress
OPEN ACCESS

Article

Oncogenic roles of young human *de novo* genes and their potential as neoantigens in cancer immunotherapy

Chunfu Xiao,[1,2,3,15] Xiaoge Liu,[1,2,3,15] Peiyu Liu,[4,15] Xinwei Xu,[1,2,3] Chao Yao,[1,2,3] Chunqiong Li,[1,2,3] Qi Xiao,[5,6] Tiannan Guo,[5,6] Li Zhang,[7] Yongjun Qian,[8] Chao Wang,[9] Yiting Dong,[10] Yingxuan Wang,[11] Zhi Peng,[12] Chuanhui Han,[13] Qiang Cheng,[4,*] Ni A. An,[2,3,*] and Chuan-Yun Li[2,3,14,16,*]

[1]State Key Laboratory of Gene Function and Modulation Research, Institute of Molecular Medicine, College of Future Technology, Peking University, Beijing 100871, China
[2]Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China
[4]Department of Biomedical Engineering, College of Future Technology, Peking University, Beijing 100871, China
[5]Westlake Center for Intelligent Proteomics, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang Province 310030, China
[6]School of Medicine, School of Life Sciences, Westlake University, Hangzhou, Zhejiang Province 310030, China
[7]Chinese Institute for Brain Research, Beijing 102206, China
[8]IDG/McGovern Institute for Brain Research, Peking-Tsinghua Center for Life Sciences, College of Future Technology, Peking University, Beijing 100871, China
[9]Department of Gastrointestinal and Colorectal Surgery, China-Japan Union Hospital, Jilin University, Changchun 130033, China
[10]State Key Laboratory of Molecular Oncology, Department of Medical Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences Peking Union Medical College, Beijing 100021, China
[11]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Department of Gastrointestinal Oncology, Peking University Cancer Hospital & Institute, Beijing 100142, China
[12]State Key Laboratory of Holistic Integrative Management of Gastrointestinal Cancers, Department of Gastrointestinal Oncology, Peking University Cancer Hospital & Institute, Beijing 100142, China
[13]School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China
[14]Southwest United Graduate School, Kunming 650092, China
[15]These authors contributed equally
[16]Lead contact
*Correspondence: qiangcheng@pku.edu.cn (Q.C.), annie@genetics.ac.cn (N.A.A.), chuanyunli@genetics.ac.cn (C.-Y.L.)
https://doi.org/10.1016/j.xgen.2025.100928

SUMMARY

Young human *de novo* genes, recently emerging from non-coding regions, are expected to contribute to human-specific traits and diseases. However, systematic explorations of this connection have been lacking. Here, we report 37 recently originated *de novo* genes in humans, with their evolution and characteristics defined within an updated genomic context. The expression of these genes is significantly upregulated and temporospatially expanded in tumors, partially associated with extrachromosomal DNA amplification. Depletion of 57.1% of these genes suppresses tumor cell proliferation, underscoring their roles in tumorigenesis. As a proof of concept, we developed mRNA vaccines expressing *ELFN1-AS1* and *TYMSOS*—young genes specifically expressed during early development but reactivated exclusively in tumors. In humanized mice, these vaccines triggered specific T cell activation and inhibited tumor growth. The antigens derived from these genes are immunogenic and capable of eliciting antigen-specific T cell activation in colorectal cancer patients. These findings underscore young human *de novo* genes as neoantigens in cancer immunotherapy.

INTRODUCTION

De novo genes arise recently from non-coding genomic regions, lacking pre-existing ''mother'' genes to serve as templates for their protein sequences, structures, and functions.[1–3] Recent studies have identified substantial numbers of such genes across a wide range of species.[4–15] In humans, the ''motherless'' origin and limited evolutionary time for functional refinement suggest the potential roles of young *de novo* genes in shaping human-specific traits[5,16–18] and disease susceptibility.[19–23] However, this connection has not been systematically explored due to the intrinsic characteristics of these genes, which complicate gene annotation. Additionally, the absence of mother genes limits our understanding of their evolutionary trajectories, functions, and disease relevance.

Please cite this article in press as: Xiao et al., Oncogenic roles of young human *de novo* genes and their potential as neoantigens in cancer immunotherapy, Cell Genomics (2025), https://doi.org/10.1016/j.xgen.2025.100928

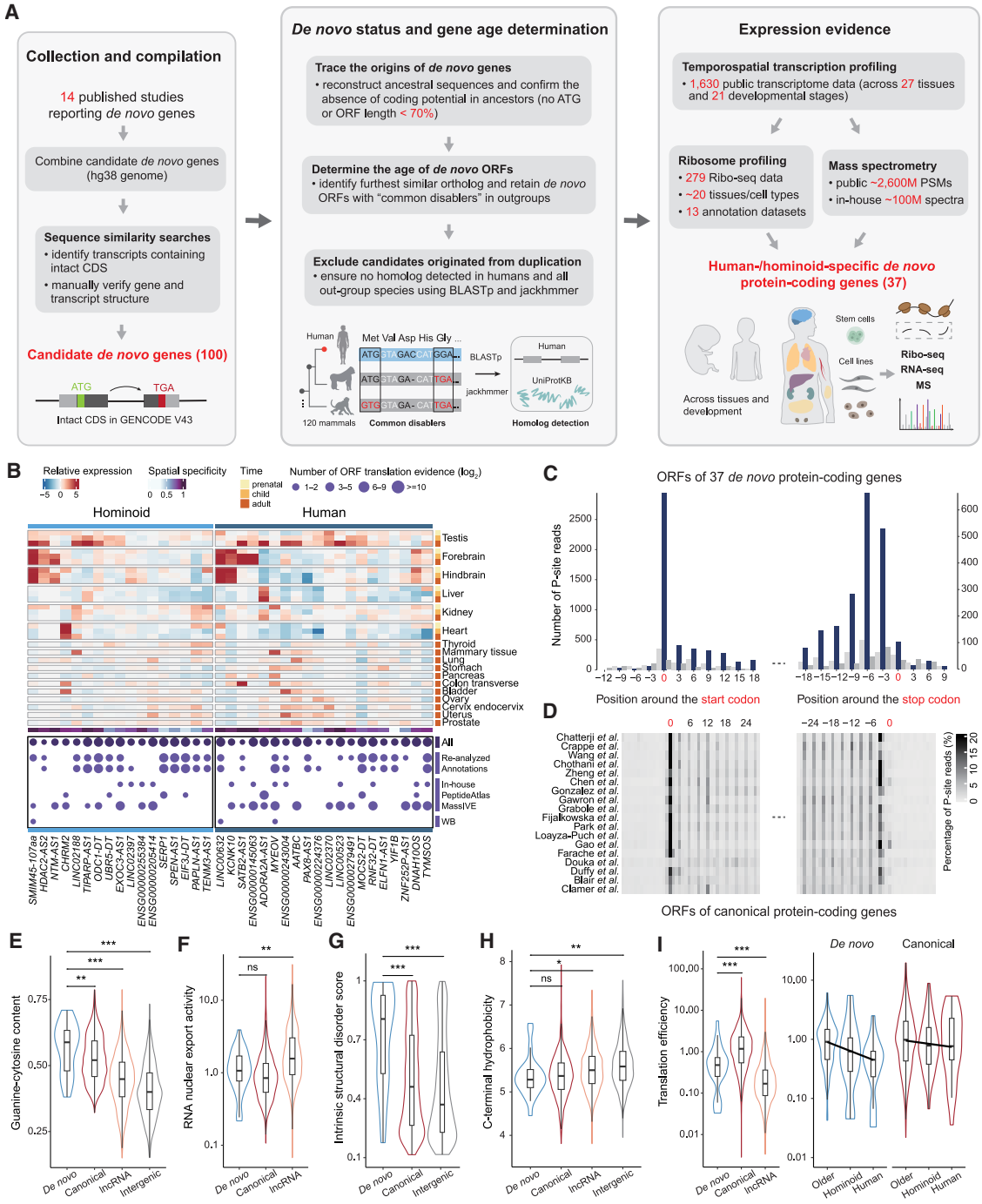


Figure 1. Newly originated *de novo* genes in humans and their evolutionary trajectories

(A) Workflow for identifying *bona fide* young *de novo* genes encoding human- or hominoid-specific proteins (see STAR Methods). The pipeline involves three steps: (1) compilation of candidate *de novo* genes with verified transcript structures, (2) phylogenetic reconstruction of evolutionary trajectories and precise gene age estimation, and (3) rigorous validation of expression. CDS, coding sequences; PSMs, peptide-spectrum matches.

(B) Expression profiles for 17 hominoid-specific and 20 human-specific *de novo* genes. Top: heatmap showing relative expression levels across tissues and developmental stages, with spatial specificity of expression scores shown in the bottom row. Bottom: dot heatmap summarizing translational evidence from multiple sources: summarized (All), re-analyzed Ribo-seq data (Re-analyzed), publicly available Ribo-seq annotations (Annotations), in-house-generated MS data (In-house), public MS database annotations (PeptideAtlas and MassIVE), and western blot validations (WB). Methodological details are provided in STAR Methods. *SMIM45-107aa* represents the 107-amino-acid young protein encoded by *SMIM45*.

(legend continued on next page)

First, these genes exhibit intrinsic characteristics such as shorter open reading frames (ORFs),[24,25] restricted and lower expression,[6,16,24,26] frequent occurrence in repetitive genomic regions,[9,10] and limited cross-species conservation, all of which complicate their definition and annotation. As a result, current computational pipelines struggle to consistently annotate these genes, reliably trace their evolutionary trajectories, and provide convincing evidence of their *in vivo* expression and functions.[27,28] Recent advances in reference genomes,[29–31] including telomere-to-telomere coverage,[32] and functional genomics across a wider range of primate species and tissues[33–36] offer new opportunities to identify these genes, clarify their evolutionary trajectories, and better understand their roles in this updated genomic context. Additionally, the evolving definition of *de novo* genes, aided by ribosome profiling (Ribo-seq) data[33,37,38] that complement traditional mass spectrometry (MS), has been discussed recently.[27] However, the systematic integration of these genomic data and the validation of their applications remain areas for further exploration.

Second, understanding the functions of these genes, particularly their disease implications, lags behind due to the scarcity of functional assays to establish causal relationships.[2] This challenge is exacerbated by their lack of sequence homology to known proteins, which would otherwise provide functional clues. Pilot studies have linked some of these *de novo* genes to brain development and spermatogenesis through mechanisms such as neural stem cell amplification and cell fate determination,[16,17] as well as anti-apoptotic effects in spermatogenesis.[39–41] These tumor-like features suggest their potential roles as oncogenes in tumorigenesis. Consistently, case studies have implicated some of these *de novo* genes in tumor development and prognosis,[19–22,42–44] and speculative genomic studies have associated these genes with tumorigenesis.[45] However, a systematic investigation of this connection and the underlying mechanisms remains unexplored. If confirmed, this link could position these young genes as universal neoantigens—similar to those encoded by noncanonical ORFs[46]—providing new strategies to address key challenges in cancer immunotherapy.

In this study, we identify 37 newly emerged *de novo* genes in humans and clarify their evolutionary trajectories within an updated genomic context. We observe a widespread upregulation and temporospatial expansion of their expression in tumors, with circular extrachromosomal DNA (ecDNA) amplifications possibly playing a role in this process. Furthermore, we experimentally establish the causal relationship between these young *de novo* genes and tumorigenesis. As a proof of concept, we develop two mRNA vaccines targeting these genes and demonstrate their potential as neoantigens in anti-tumor therapies.

RESULTS

Identification of human *de novo* genes with clear evolutionary trajectories

To identify an accurate list of *de novo* protein-coding genes recently emerged in humans, we systematically evaluated their origination, evolutionary trajectories, and expression within an updated genomic context (Figure 1A). We began by compiling a list of 100 candidate *de novo* genes from 14 public studies, followed by manual curation to confirm intact gene structures and ORFs of these candidate genes (STAR Methods; Table S1). Building upon recent methodological advances in *de novo* gene identification, which address false positives arising from rapid sequence divergence, gene loss, and distant homology,[12,47–51] we developed a computational pipeline to rigorously assess the *de novo* origination of candidate genes in the hominoid lineage. Briefly, we reconstructed ancestral genomic sequences for these candidates using whole-genome synteny alignments across 120 mammalian species.[47,52] To establish *de novo* emergence in the hominoid lineage, we required (1) the absence of intact ORFs in orthologous regions of ancestral sequences predating the divergence of Old World monkeys and hominoids and (2) the presence of shared disruptive mutations (“common disablers”) in outgroup lineages that disrupt the ORFs. We further distinguished true *de novo* genes from gene duplications by confirming the absence of sequence homology with these ORFs among the human genome, the human transcriptome, and all annotated proteomes in UniProtKB database (253,206,170 entries, 1,333,558 species; STAR Methods). Notably, a stricter criterion was applied to ensure the complete absence of coding potential in ancestral regions, even if these regions encode entirely different proteins in the ancestral state (STAR Methods).

For the candidate *de novo* genes that recently originated in humans, we next investigated their *in vivo* expression in human tissues. This was based on genomic profiles processed from 1,630 transcriptomes, 279 Ribo-seq datasets, and 100 million in-house-generated MS spectra from various human tissues and cell lines (STAR Methods). To reduce false positives due to pervasive natural antisense transcripts, we defined representative regions unique to each *de novo* transcript for expression quantification (Figure 1B; STAR Methods). Taken together, we identified 37 *de novo*-originated genes in humans—20

(C) Ribosomal peptidyl site (P-site) positioning profiles around the start and stop codons of *de novo* genes, derived from ribosome-protected fragments of 279 Ribo-seq datasets.
(D) Percentage of P-site reads around the start and stop codons of canonical protein-coding genes across all datasets.
(E–H) Comparative analyses of four core gene properties (see STAR Methods): guanine-cytosine content of ORF sequences (E), RNA nuclear export activity (F), degree of intrinsic structural disorder (G), and C-terminal hydrophobicity of proteins (H), for canonical protein-coding genes (Canonical), *de novo* genes (*De novo*), and non-coding genes (lncRNA) or regions (Intergenic).
(I) Violin plots showing translation efficiency distributions for *de novo* genes, canonical protein-coding genes, and non-coding genes (left) or for *de novo* genes and canonical genes grouped by evolutionary age: human-specific (Human), hominoid-specific (Hominoid), or more conserved (Older, right).
Two-tailed Wilcoxon rank-sum test ($n = 37$ [*De novo*] or 4,000 [Canonical, lncRNA, and Intergenic]). ns, not significant; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. See also Figure S1 and Tables S1 and S2.

Please cite this article in press as: Xiao et al., Oncogenic roles of young human *de novo* genes and their potential as neoantigens in cancer immunotherapy, Cell Genomics (2025), https://doi.org/10.1016/j.xgen.2025.100928

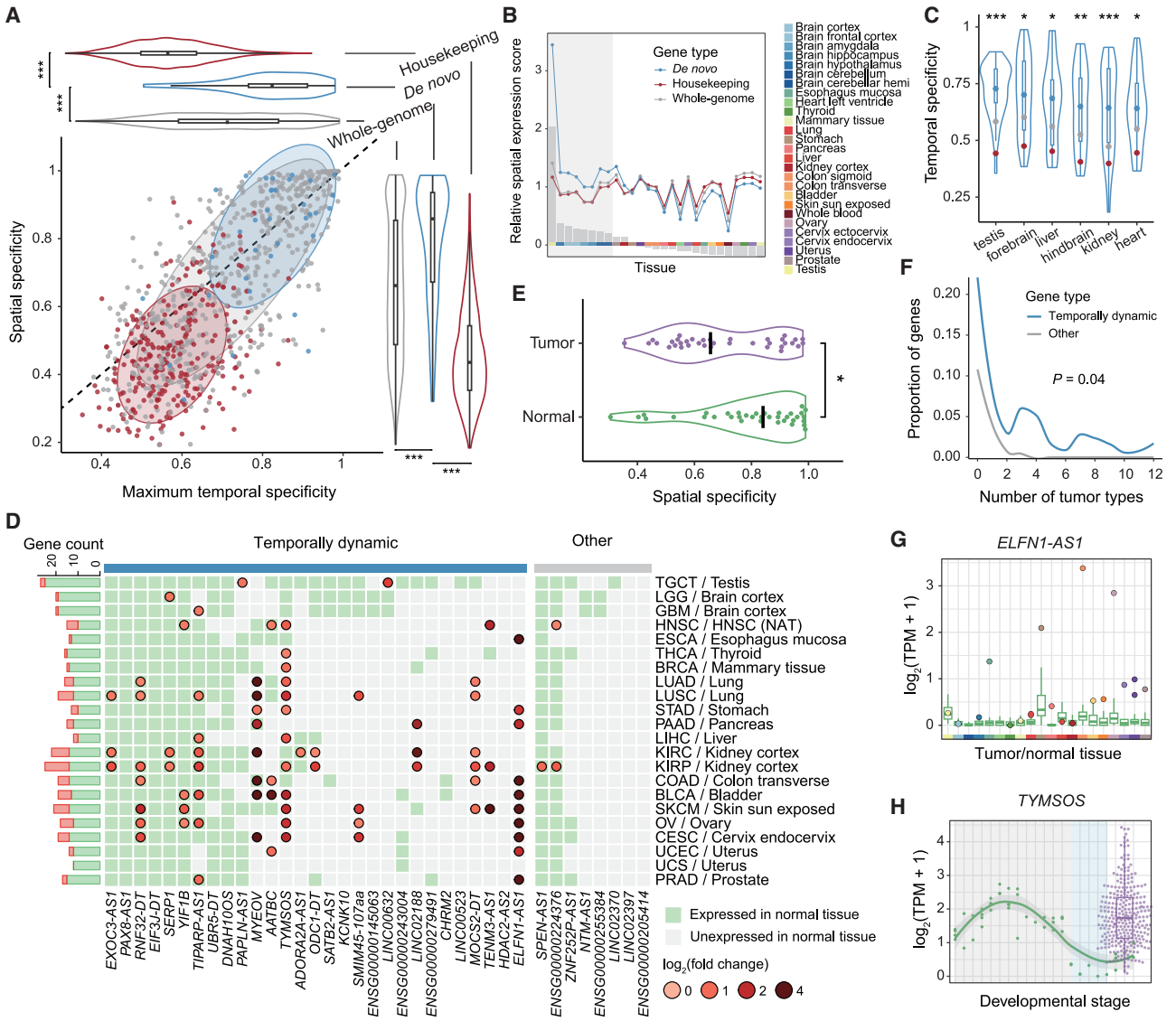


Figure 2. Upregulation and temporospatial expansion of *de novo* gene expression in tumors

(A) Spatial specificity and maximum temporal specificity of gene expression (in six organs) for housekeeping genes (red dots, Housekeeping), *de novo* genes (blue dots, *De novo*), and shuffled genomic background (gray dots, Whole-genome).

(B) Comparative tissue-specific expression profiles across the three gene groups. Heatmap showing relative expression levels (median normalized across all tissues) for each gene group among 27 human tissues, with brain and testis highlighted (shaded block). Bar plots quantify the differences between *De novo* genes and Whole-genome background across tissue types.

(C) Temporal specificity distributions for *De novo* genes in six organs, with median specificity values for Housekeeping (red dots), *De novo* (blue dots), and Whole-genome (gray dots) highlighted. Statistical significance is shown for differences between *De novo* genes and Whole-genome background. Two-tailed Wilcoxon rank-sum test.

(D) *De novo* gene expression across tumor types in TCGA. Genes with expression in normal tissues (green boxes, median transcripts per million [TPM] values ≥0.5, estimated by GTEx RNA-seq data) and those significantly upregulated in tumors at varying degrees (red dots) are highlighted. Fold change was calculated using upper quantile TPM values of tumor versus normal tissues. For each tissue type, the number of *de novo* genes in the two categories is shown as green and red bars on the left side. NAT, normal adjacent tissue. Tumor abbreviations follow TCGA study abbreviations.

(E) Spatial specificity of *de novo* gene expression in paired tumor and normal tissues.

(F) Proportions of *de novo* genes significantly upregulated across varying numbers of tumor types, categorized by temporally dynamic and non-dynamic *de novo* genes.

(G) Spatial expansion of *ELFN1-AS1* expression in tumors. Boxplots represent expression in normal tissue samples; dots show median expression in tumor samples of the corresponding tissue type.

(legend continued on next page)

Please cite this article in press as: Xiao et al., Oncogenic roles of young human *de novo* genes and their potential as neoantigens in cancer immunotherapy, Cell Genomics (2025), https://doi.org/10.1016/j.xgen.2025.100928

human-specific and 17 hominoid-specific genes shared by humans and apes—supported by robust transcriptional and translational evidence from at least three independent datasets or studies (Figures 1A and 1B; Table S2). Ribosome footprints corresponding to these young ORFs exhibited distinct three-nucleotide periodicity, as determined by ribosome-protected fragment analysis of Ribo-seq data (Figure 1C; STAR Methods), a signal indicative of *in vivo* translation, aligning with patterns observed for canonical protein-coding genes (Figures 1D and S1A–S1C).

Using the list of 37 *bona fide de novo* genes, we then investigated their evolutionary trajectories in terms of key gene properties.[16,25,53] We found that the evolution of core properties for ORF formation, mRNA translation opportunity, and protein stability follows a preadaptive process. Specifically, young genes display comparable, or even enhanced, gene-like characteristics. We quantified four core properties for these young *de novo* genes, canonical genes, and non-coding regions: (1) guanine-cytosine content of the ORF sequence, where higher content is linked to easier exon origination[54,55] and the formation of longer coding sequences[56] (Figure 1E); (2) nuclear export activity of transcribed RNA, where higher activity indicates a greater likelihood of RNA molecules being transported to the cytosol for translation (Figure 1F); and (3) intrinsic structural disorder degree of protein linked to the aggregation prevention[57,58] (Figure 1G). These disordered regions also serve as flexible linkers between structural domains, with a higher degree of disorder associated with increased protein interaction potential, enhanced functional diversity,[58,59] and accelerated evolutionary rates[25,60]; and (4) C-terminal hydrophobicity of protein, where lower hydrophobicity promotes protein stability by reducing proteasomal degradation or membrane targeting[61] (Figure 1H). We analyzed RNA sequencing (RNA-seq) data from the nucleus and cytoplasm to evaluate RNA nuclear export, alongside computational approaches to assess three other gene properties (STAR Methods). Strikingly, young *de novo* genes exhibited levels of these properties comparable to—or even exceeding—those of canonical genes and non-coding regions (Figures 1E–1H), contrary to the intermediate characteristics expected by a gradual transition model between non-coding and coding states. These gene-like properties may reflect either evolutionary selection for adaptive functions[14,16] or preferential emergence within genomic regions harboring preadaptive, gene-like architectures—so-called "hopeful monsters"[16,25]—that facilitate *de novo* gene origination. In contrast, for properties fine-tuning gene functions, such as translation efficiency estimated from RNA-seq and Ribo-seq data, young *de novo* genes showed intermediate levels between canonical protein-coding genes and non-coding genes (Figures 1I, S1B, and S1C; STAR Methods). Translation efficiency appears to be further optimized with age, as older genes demonstrate higher translation efficiency (Figure 1I), supporting a continuum evolution model.[53] Together, these findings indicate that young *de novo* genes likely originate from pre-existing, gene-like "precursors" shaped by selective forces for adaptive functions, with subsequent optimization of certain features to form efficient proteins in humans.

Temporospatially expanded expression of *de novo* genes in tumors

Gene expression profiles are often indicative of their biological functions, as suggested by the omnigenic model.[62] To investigate the expression dynamics of young *de novo* genes, we first analyzed their transcriptomic profiles using RNA-seq data spanning 27 human tissues and multiple developmental stages from the Genotype-Tissue Expression (GTEx) project and Cardoso-Moreira et al.[34,63] and systematically quantified their temporal and spatial specificity of expression (STAR Methods). A significant correlation was observed between temporal and spatial specificity across different genes (Figure S2; Pearson correlation coefficient $r = 0.73$, $p = 2.2 \times 10^{-16}$), with young *de novo* genes showing significantly higher temporal and spatial specificity compared to housekeeping genes and randomly shuffled genes (Figure 2A), consistent with previous findings.[6,24,63] These young *de novo* genes showed predominant expression in testes and brain tissues compared to other gene classes (Figures 1B and 2B). Notably, these genes also exhibited restricted temporal expression patterns, suggesting tightly regulated developmental windows of activity (Figure 2C).

We then identified temporally dynamic *de novo* genes with significant expression changes across developmental stages and classified them into two groups—early-expressed and late-expressed genes—based on their expression profiles showing predominant expression at early or late developmental stages, respectively (Figures S3 and S4A; STAR Methods). Notably, most temporally dynamic *de novo* genes in the testes were late expressed, recapitulating a pattern observed in *Drosophila*,[64] while many in the hindbrain and liver were early expressed (Figure S4B). The temporospatially restricted expression of young human *de novo* genes recapitulates their roles in repressing apoptosis in spermatocytes[39,41] and maintaining the neural stem cell pool during early brain development.[16,17]

Given their roles in stem cell maintenance and anti-apoptosis, we explored the implications of *de novo* genes in tumorigenesis.[2,26,40] While exhibiting low expression in most normal tissues (with exceptions in brain and testes), these genes showed significant upregulation in corresponding tumor types, as evidenced by transcriptomes of 5,278 samples across 22 tumor types from The Cancer Genome Atlas (TCGA) (Figure 2D). Strikingly, 13 genes (68.4% of the tumor-upregulated gene set) demonstrated expanded expression profiles, with no detectable expression in matched normal tissues (Figures 2D and S7A). This expression expansion was significantly more prevalent among *de novo* genes and occurred across a broader spectrum of tumor types compared to background genes (Figure S7B; Monte

(H) Temporal expansion of *TYMSOS* expression in LIHC. The green dots represent the median *TYMSOS* expression in livers across different developmental stages, with a fitting curve shown; purple boxplots and dots depict its expression distribution in LIHC tumor samples.
(A, C, and E) *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. (A and C) Two-tailed Wilcoxon rank-sum test ($n = 37$ [*De novo*], 1,456 [Housekeeping], or 3,000 [Whole-genome]). (E) Two-tailed Wilcoxon signed-rank test. (F) One-tailed Kolmogorov-Smirnov test. See also Figures S3 and S7.

Please cite this article in press as: Xiao et al., Oncogenic roles of young human *de novo* genes and their potential as neoantigens in cancer immunotherapy, Cell Genomics (2025), https://doi.org/10.1016/j.xgen.2025.100928

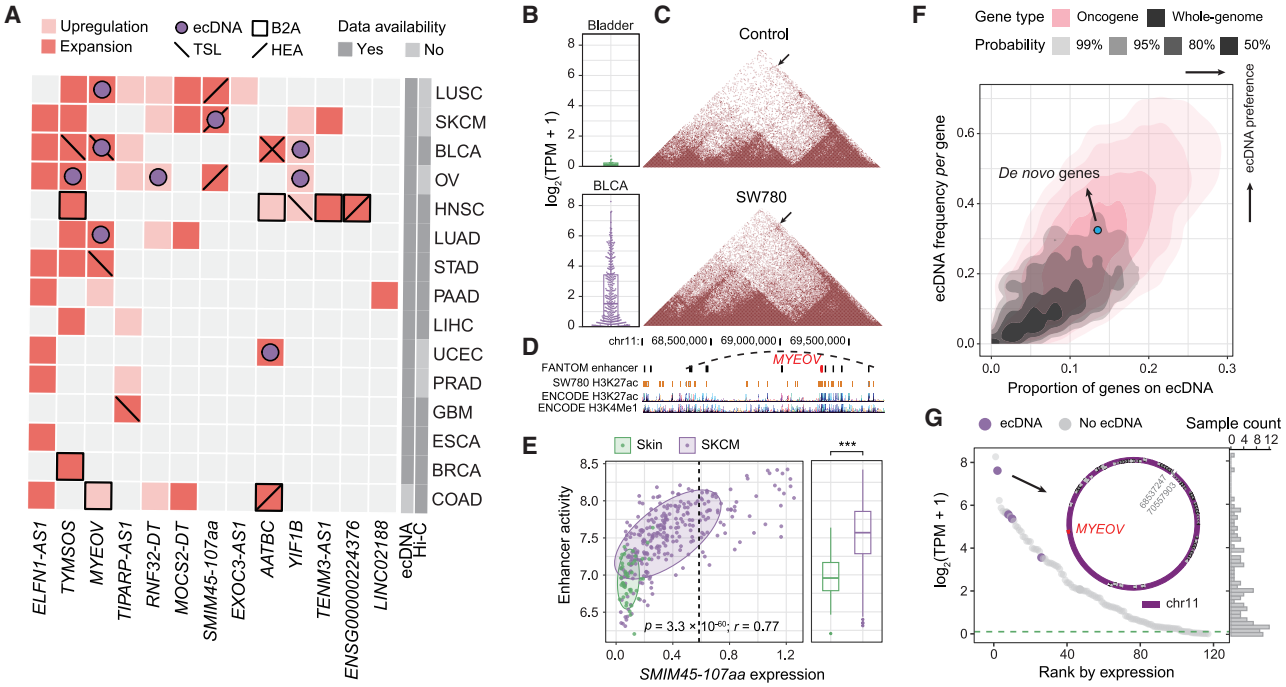


Figure 3. Association between genome reorganization and altered expression profiles of *de novo* genes in tumors

(A) Events of upregulation or expansion of *de novo* gene expression across various tumor types, with four types of genome reorganization events indicated.

(B–D) Tumor-specific expansion of *MYEOV* expression in BLCA, compared to normal bladder tissue samples (B). This expansion is potentially driven by the formation of a tumor-specific enhancer-promoter interaction mediated by a chromatin loop, indicated by arrows in the normalized Hi-C contact maps for SW780 (bladder cancer) and Control (human epithelial cell) (C). For regions surrounding this enhancer-promoter interaction, active enhancer positions, as annotated by the Functional Annotation of the Mammalian Genome (FANTOM) project, and associated epigenetic marks indicating enhancers are shown (D).

(E) Correlation between *SMIM45-107aa* expression and the activity of a nearby enhancer (chr22:41949353–41949930) across normal skin (Skin) and skin cutaneous melanoma (SKCM, left). Spearman correlation coefficient $r = 0.77$. Enhancer activity in SKCM and Skin samples was compared (right). Two-tailed Wilcoxon rank-sum test ($n = 50$ [Skin] or 295 [SKCM]). ***$p < 0.001$.

(F) Proportion of genes detected on ecDNA amplifications across three gene categories: *de novo* genes (indicated by blue dot with arrow), background protein-coding genes, and known oncogenes. Data represent 114 BLCA samples, with y axis showing the occurrence frequency of ecDNA amplifications per gene. The probability distributions for background and oncogenes were estimated using Monte Carlo simulations (STAR Methods).

(G) *MYEOV* expression in 114 BLCA tumor samples. The median *MYEOV* expression in normal bladder samples is marked with a green dotted line, and samples with predicted, *MYEOV*-carrying circular ecDNA are highlighted as purple dots. The structure of one *MYEOV*-carrying circular ecDNA from a highlighted sample (arrow) is shown.

See also Figures S8–S12.

Carlo simulation, $p < 0.0001$; STAR Methods). Consequently, *de novo* genes exhibited significantly reduced spatial specificity in tumor versus matched normal tissues (Figure 2E; Wilcoxon signed-rank test, two tailed, $p = 0.02$), indicative of relaxed expression constraints during oncogenesis. Notably, temporally dynamic genes showed particularly pronounced upregulation across multiple tumor types (Figures 2D and 2F; Kolmogorov-Smirnov test, one tailed, $p = 0.04$). Two representative cases illustrate this phenomenon: *ELFN1-AS1*, which is virtually silent across all normal tissues, displayed marked expression in colon adenocarcinoma (COAD) and ovarian cancers (Figure 2G), while *TYMSOS*, normally restricted to early liver development, showed substantial reactivation in liver hepatocellular carcinoma (LIHC) (Figure 2H). Additional examples of tumor-associated expression expansion of *de novo* genes are presented in Figures S7A and S7C.

It has been reported that the mechanism of ecDNA can promote oncogene expression through gene amplification and enhanced chromatin accessibility.[65–68] Additionally, dysregulation of topologically associating domains and enhanced chromatin interactions can drive oncogene overexpression during tumorigenesis.[69,70] Motivated by these findings, we next investigated whether the rewired genome architectures are associated with the observed temporospatial expansion of *de novo* gene expression in tumors. We integrated and re-analyzed 58 public Hi-C sequencing datasets from 10 tumor types and corresponding normal tissues. This revealed tumor-specific compartment transitions from B to A (B2A), tumor-specific chromatin loops with enhancer-promoter interactions (TSLs), and tumor-specific higher enhancer activity (HEA) (STAR Methods; Figure S8), all contributing to upregulation and temporospatial expansion of gene expression in tumors (Figure S9). Two cases of *de novo* gene with expanded expressions in tumors, potentially through these mechanisms, are shown (Figures 3B–3E). We also integrated annotated ecDNAs from the eccDNA Atlas database[71] (STAR Methods). Based on these genome reorganization events,

Please cite this article in press as: Xiao et al., Oncogenic roles of young human *de novo* genes and their potential as neoantigens in cancer immunotherapy, Cell Genomics (2025), https://doi.org/10.1016/j.xgen.2025.100928

CellPress
OPEN ACCESS

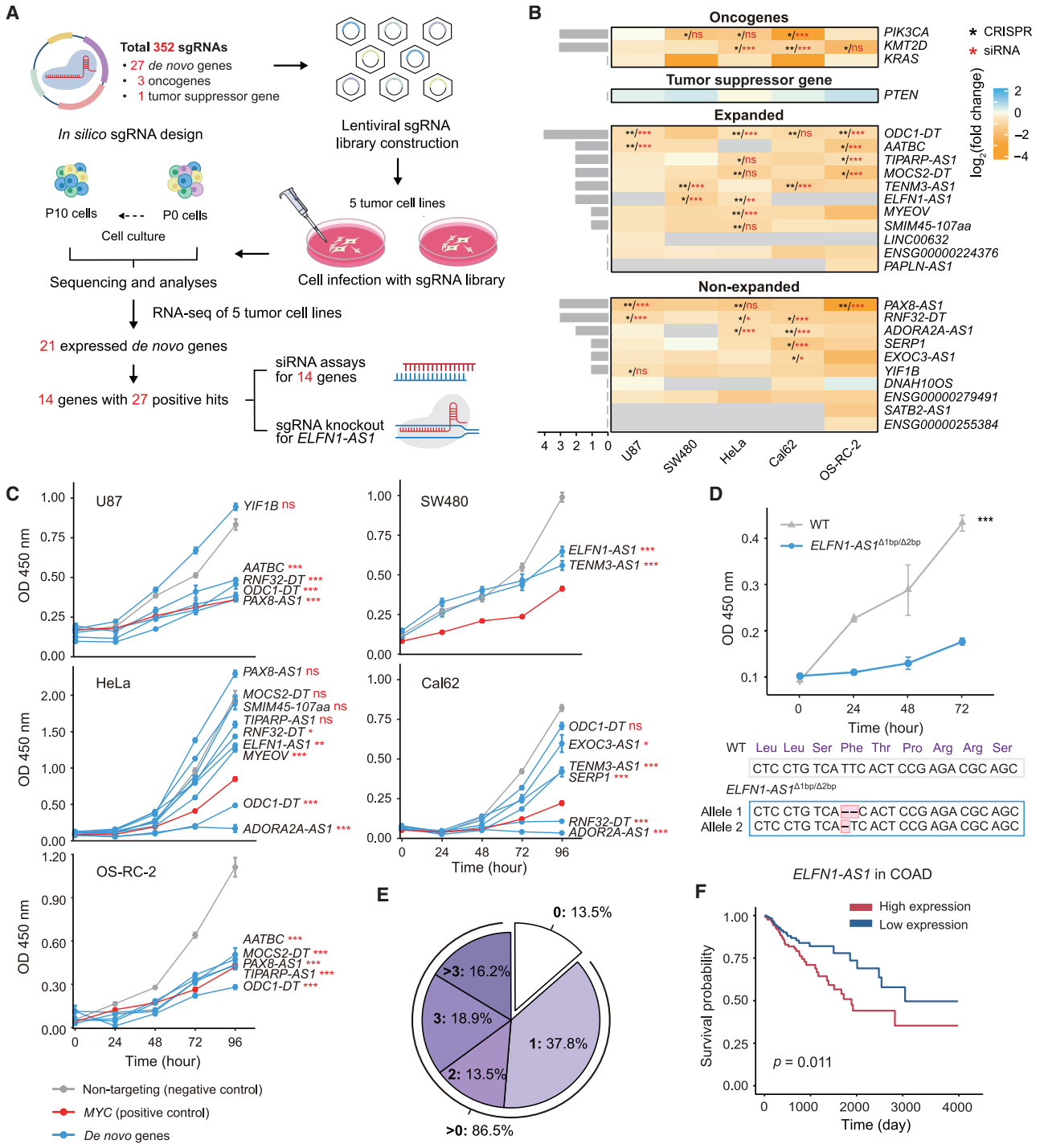


Figure 4. Functional roles of young *de novo* genes in tumor growth

(A) Schematic representation of the CRISPR-Cas9 library design, lentiviral infection of five tumor cell lines (U87, SW480, HeLa, Cal62, and OS-RC-2), targeted sequencing, assessment of sgRNA abundance in samples immediately after infection (P0) and after 10 passages (P10), followed by validations through siRNA knockdown assays and CRISPR-Cas9 knockout experiments.

(B) Heatmap depicting the $\log_2$ (fold change) of sgRNA abundance between P0 and P10 across five tumor cell lines. The targets include 3 oncogenes, 1 tumor-suppressor gene, and 21 *de novo* genes classified as Expanded (with expression expansion in tumors) or Non-expanded. *p*-values from the initial CRISPR-Cas9 screening are indicated by black asterisks with paired t test, while those from the siRNA knockdown assays are indicated by red asterisks, with two-way ANOVA test. *De novo* genes unexpressed in specific cell lines are masked (gray blocks; STAR Methods).

(legend continued on next page)

we found that 45.9% of tumor-specific expansion in *de novo* gene expression was associated with at least one of these mechanisms, with the ecDNA amplification being the most prevalent, accounting for 16.2% of expansion events and 16.4% of tumor-specific upregulation of *de novo* genes (Figure 3A). Notably, a larger proportion of upregulated *de novo* genes in tumors were found amplified by ecDNAs compared to randomly selected background genes ($p = 0.035$, Fisher's exact test), a portion comparable to that observed for known oncogenes ($p = 0.18$, Fisher's exact test; Figure S10A). For the other three mechanisms—B2A, TSLs, and HEAs—no significant differences were detected (Figures S10B–S10D).

To confirm the contribution of ecDNAs, we further identified the genomic amplification events by re-analyzing whole-genome sequencing data of 114 bladder urothelial carcinoma (BLCA) samples and their corresponding transcriptomes from TCGA (STAR Methods). Our analysis revealed 12 distinct circular ecDNA amplification events encompassing five *de novo* genes (Figure S11A). Strikingly, these events occurred at significantly higher frequencies for *de novo* genes compared to background genes, with prevalence rates approaching those of established oncogenes in these samples (Figure 3F; Monte Carlo simulation, $p = 0.0013$). Consistent with the known transcriptional enhancement mediated by ecDNA amplification,[66,67] these *de novo* genes associated with ecDNA are generally upregulated (Figure S11B). An example is *MYEOV*, which, when amplified on circular ecDNAs, showed markedly elevated expression in BLCA tumors (Figure 3G). Similar patterns were observed for other *de novo* genes (Figures S12A–S12C).

In summary, the upregulation and temporospatial expansion of *de novo* gene expression in tumors implicate their roles in tumorigenesis, possibly through the reprogramming of proliferative malignant cells and the maintenance of cellular plasticity, similar to carcinoembryonic genes widely used in early cancer screening.[72]

Functions of young *de novo* genes in tumor cell proliferation

Inspired by the observed upregulation and temporospatial expansion of *de novo* gene expression in tumors, we investigated whether these young genes directly contribute to tumorigenesis. To assess this, we developed a CRISPR-Cas9 screening assay targeting 27 young *de novo* genes along with three well-established oncogenes (*PIK3CA*, *KMT2D*, and *KRAS*) and a tumor-suppressor gene (*PTEN*) as positive controls. The library included 352 single-guide RNAs (sgRNAs), with 4–5 sgRNAs per gene, and 20 non-targeting sgRNAs as negative controls. We evaluated their effects on the growth of multiple cancer cell lines representing diverse tumor types, including U87 (glioblastoma), SW480 (COAD), HeLa (cervical carcinoma), Cal62 (thyroid anaplastic carcinoma), and OS-RC-2 (renal carcinoma) (Figure 4A). The relative abundance of each sgRNA was quantitatively assessed following viral infection and subsequently after 10 cell passages to detect changes in sgRNA representation over time (Figure 4A; STAR Methods). A significant reduction in sgRNAs targeting known oncogenes validated the assay's efficacy (Figure 4B). We then performed RNA-seq for these five cell lines to quantify the expression of these *de novo* genes and identified 21 genes expressed in at least one of the cell lines (Figure 4B; STAR Methods). Due to the limited availability of public ribosome profiling data for these specific cell lines, direct evidence of translation of these genes remains inconclusive. Nonetheless, since these genes were confirmed to be translated in our initial identification (Figures 1B and 1C), it is likely that they are also translated, given the observed transcription levels in these cell lines. Notably, we observed a significant reduction in the abundance of sgRNAs targeting 14 out of the 21 expressed *de novo* genes in at least one tumor cell line (Figure 4B; STAR Methods), indicating that depletion of these genes directly impairs tumor cell proliferation. Specifically, for all eight *de novo* genes exhibiting expanded expression profiles in tumors, the depletion of each of them significantly reduced cell proliferation, emphasizing their functional relevance in tumor cell proliferation (Figure 4B).

In the CRISPR-Cas9 screening library, we included non-targeting sgRNAs as negative controls to account for any potential confounding effects of lentiviral infection. To further validate the results and eliminate such effects, we performed small interfering RNA (siRNA) knockdown assays to confirm all of the positive hits observed in the CRISPR-Cas9 screening. Notably, among the 27 positive hits (14 *de novo* genes across five cell lines), 21 were confirmed, showing that transient silencing of these young *de novo* genes with specific siRNAs directly inhibited tumor cell proliferation (Figures 4B and 4C).

For one of these genes, *ELFN1-AS1* (Figure S13), we further performed a CRISPR-Cas9 assay to knock it out in the SW480 cell line. We isolated a mutant cell clone where both alleles carried frameshift mutations, effectively disrupting the ORF (Figure 4D). Consistent with the findings from both the CRISPR-Cas9 screening and the siRNA knockdown assays, the proliferation of *ELFN1-AS1* mutant cells was significantly reduced (Figure 4D; two-way ANOVA test, $p = 2.0 \times 10^{-7}$). In total, the depletion of 57.1% (12 of 21) of these young *de novo* genes resulted in a significant suppression of tumor cell proliferation, underscoring their involvement in tumorigenesis.

In line with these findings, these young *de novo* genes may serve as clinical prognostic biomarkers. For 86.5% of these genes (32 out of 37), patients with higher expression levels

(C) CCK-8 cell proliferation curves showing the effect of *de novo* gene knockdown on tumor cell proliferation. $n = 6$ biologically independent samples per group. Two-way ANOVA test.

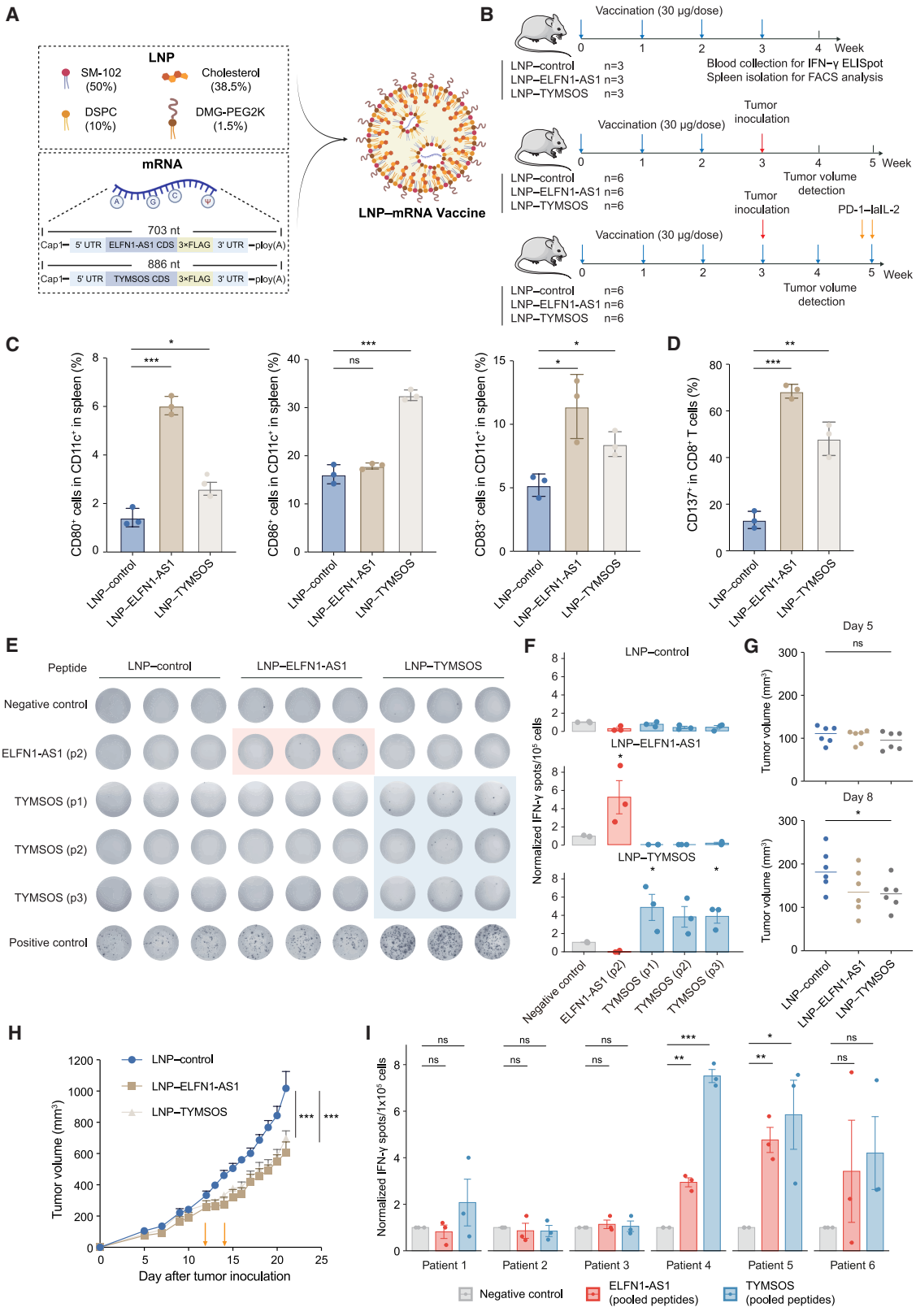
(D) CCK-8 cell proliferation curves comparing wild type and *ELFN1-AS1*$^{\Delta 1bp/\Delta 2bp}$ SW480 cells. $n = 6$ biologically independent samples per group. Two-way ANOVA test.

(E) Proportion of young *de novo* genes whose higher expression was significantly associated with poor prognosis across varying numbers (one, two, three, or more) of tumor types.

(F) Kaplan-Meier survival analysis for *ELFN1-AS1* expression in COAD patients. Prognostic significance was evaluated using the Cox proportional hazards model. Data are presented as the mean ± SEM. ns, not significant; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. See also Figure S13.

Please cite this article in press as: Xiao et al., Oncogenic roles of young human *de novo* genes and their potential as neoantigens in cancer immunotherapy, Cell Genomics (2025), https://doi.org/10.1016/j.xgen.2025.100928

CellPress
OPEN ACCESS



(legend on next page)

Please cite this article in press as: Xiao et al., Oncogenic roles of young human *de novo* genes and their potential as neoantigens in cancer immunotherapy, Cell Genomics (2025), https://doi.org/10.1016/j.xgen.2025.100928

show significantly poorer prognosis in at least one cancer type, as determined from clinical and transcriptomic data of 5,278 patients in TCGA (Figure 4E; STAR Methods). For instance, elevated expression of *ELFN1-AS1* was significantly associated with reduced survival time in COAD patients (p = 0.011; Figure 4F). Collectively, these findings highlight the potential of these young human *de novo* genes as promising candidates for anti-tumor drug targets.

Young *de novo* genes as neoantigens in cancer immunotherapy

While mRNA vaccines show promise in cancer immunotherapy, their application is limited by a lack of efficient and specific tumor neoantigens. This challenge is particularly acute for developing broadly applicable preventive vaccines targeting multiple tumor types across diverse populations. Encouraged by the unique expression patterns of some young *de novo* genes—active during early development, silenced in adult tissues, and reactivated exclusively in tumors—we explored the potential of these genes as neoantigens in anti-tumor mRNA vaccines.

As a proof of concept, we designed mRNA vaccines targeting two of these young *de novo* genes—*ELFN1-AS1* and *TYMSOS*—which are specifically expressed during early development but are reactivated exclusively in tumors (Figures 5A, S3, S13, and S14). The mRNA molecules were modified with pseudouridine and cap1 structures to reduce immunogenicity while improving stability and translational efficiency (Figure 5A; STAR Methods). mRNA lipid nanoparticles (LNP–control, LNP–ELFN1-AS1, and LNP–TYMSOS) were generated and subjected to quality control based on LNP formulations used in several FDA-approved mRNA vaccines (Figure S15A; STAR Methods).[73,74] *In vivo* translation of both mRNA vaccines was verified by western blot analyses (Figure S15B).

To evaluate the immunogenicity and therapeutic effects of these vaccines, we then developed a humanized mouse model with transplanted human immune systems by engrafting human $CD34^+$ hematopoietic stem cells into *C-NKG* mice (Figure S16A). This approach could better mimic the complexity of the human immune response compared to conventional peripheral blood mononuclear cell (PBMC)-based humanized mice (STAR Methods). At 10 weeks post engraftment, flow cytometric analyses confirmed the presence of various human myeloid cell populations, including dendritic cells (DCs), in spleens, and humanized T cells accounted for 30%–60% of PBMCs, thereby validating the successful establishment of the humanized mouse model (Figures S16A and S16B). After weekly vaccinations with LNP–control, LNP–ELFN1-AS1, or LNP–TYMSOS for four doses (Figure 5B), we assessed the immunogenicity of each group. Mice treated with LNP–ELFN1-AS1 or LNP–TYMSOS showed efficient DC activation, as indicated by increased expression of the maturation marker CD83 and co-stimulatory molecules CD80 and CD86 in $CD11c^+$ cells from spleens, compared to those in the LNP–control group (Figures 5C, S17A, and S17B). Additionally, the proliferation of activated $CD8^+$ T cells was also observed in mice treated with the mRNA vaccines, as quantified by the activation marker CD137 (Figures 5D and S17C).

To further validate antigen-specific immune responses, we predicted antigenic epitopes of *ELFN1-AS1* and *TYMSOS* with high binding affinity to human leukocyte antigens (HLAs; STAR Methods; Table S16) and synthesized three specific epitope peptides for each target. Based on T2 cell binding assays, we selected one peptide for *ELFN1-AS1* and three for *TYMSOS* that exhibited strong and specific binding to HLA-A*02:01 (Figure S17). Using these epitope peptides, we assessed T cell responses through IFN-γ ELISpot assays (STAR Methods; Figures S18A and S18B). PBMCs from humanized mice vaccinated with the mRNA vaccines and stimulated with the selected peptides exhibited significantly higher numbers of IFN-γ-producing T cells compared to those from mice treated with LNP–control (Figures 5E and 5F), demonstrating that the epitope peptides successfully induced antigen-specific T cell responses.

Figure 5. Two mRNA vaccines targeting *de novo* genes stimulate immunity and suppress tumor growth in humanized mice

(A) Schematic of the LNP–mRNA vaccine. LNPs containing SM-102, cholesterol, 1,2-distearoyl-*sn*-glycero-3-phosphocholine (DSPC), and DMG-PEG2K deliver mRNAs encoding *ELFN1-AS1* or *TYMSOS*. The mRNA sequence includes 5′ and 3′ UTRs, a coding sequence with a 3×FLAG tag, and a poly(A) tail. The lengths of both mRNAs are indicated. mRNAs were modified with Cap1 at the 5′ end, and uridine residues were replaced with pseudouridine (Ψ).

(B) The vaccination in humanized mice comprised three sequential experiments: (1) immunogenicity assessment through four weekly vaccinations followed by terminal blood and spleen collection for IFN-γ ELISpot analyses and fluorescence-activated cell sorting; (2) prophylactic efficacy evaluation involving three weekly vaccine doses prior to tumor inoculation at week 4, with subsequent tumor volume measurements; and (3) therapeutic efficacy testing with continuous weekly vaccinations combined with PD-1–laIL-2 treatment initiated at week 5, followed by longitudinal tumor volume monitoring.

(C and D) Proportions of activated DCs ($CD80^+$, $CD86^+$, and $CD83^+$ in $CD11c^+$ cells in C and $CD137^+$ $CD8^+$ T cells in D) in spleens of humanized mice, assessed by flow cytometry analyses after treatment with LNP–control, LNP–ELFN1-AS1, or LNP–TYMSOS. n = 3 biological replicates per group; unpaired two-tailed t test.

(E and F) IFN-γ ELISpot assays showing IFN-γ production by PBMCs after epitope peptide stimulation post LNP–mRNA therapy. Plate images (E) and quantification of IFN-γ spots normalized by both positive and negative controls (F, STAR Methods) were shown. n = 3 biologically independent samples per group; unpaired one-tailed t test for each epitope peptide (LNP–ELFN1-AS1 or LNP–TYMSOS versus LNP–control).

(G) Tumor volumes in humanized mice treated with LNP–control, LNP–ELFN1-AS1, or LNP–TYMSOS on days 5 and 8 post-tumor inoculation. n = 6 biologically independent samples per group; unpaired two-tailed t test.

(H) Tumor growth curves of humanized mice treated with mRNA vaccines and PD-1–laIL-2. PD-1–laIL-2 (20 μg per mouse) was administered intraperitoneally on days 12 and 14 post tumor inoculation (indicated by arrows). n = 6 biologically independent samples per group; data are presented as the mean ± SEM. Two-way ANOVA test.

(I) Quantification of IFN-γ spots normalized by both positive and negative controls, in PBMCs from six colorectal cancer patients (co-incubated with pools of *ELFN1-AS1* or *TYMSOS* peptides, STAR Methods). n = 3 biological replicates per patient; unpaired one-tailed t test (each pooled peptides versus the negative control for each patient).

(C–I) ns, not significant; $^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. (C and D) Bars, mean; error bars, SD. (E–I) Bars, mean; error bars, SEM. See also Figures S13–S19 and Tables S16 and S17.

Please cite this article in press as: Xiao et al., Oncogenic roles of young human *de novo* genes and their potential as neoantigens in cancer immunotherapy, Cell Genomics (2025), https://doi.org/10.1016/j.xgen.2025.100928

In a follow-up experiment, after administrating four doses of the mRNA vaccines, we subcutaneously implanted SW480 tumor cells into humanized mice to assess the efficacy of the vaccines in suppressing tumor growth. To ensure physiological relevance and avoid complications from HLA mismatching, we selected HLA-A-matched donors for tumor inoculation (the HLA-A subtype of human $CD34^+$ stem cells transplanted in humanized mice and SW480 tumor cells: HLA-A*02:01, 24:02). Consistent with the observed T cell responses, both LNP–ELFN1-AS1 and LNP–TYMSOS vaccines effectively elicit targeted immune activation and significantly inhibited tumor progression during the early stages (Figures 5B and 5G).

However, the overall anti-tumor effects of the mRNA vaccines were modest—likely due to the limited functionality and expansion capacity of $CD8^+$ T cells in the humanized mice. We thus investigated the efficacy of combining the vaccines with PD-1–laIL-2 treatment, which selectively targets and delivers IL-2 signaling to $CD8^+$ T cells within the tumor microenvironment, thereby promoting their proliferation and activation.[75] Specifically, PD-1–laIL-2 was administered on days 12 and 14 post-tumor inoculation to enhance intratumoral $CD8^+$ T cell avidity. Notably, this combination therapy elicited a stronger anti-tumor effect, with treatment using either LNP–ELFN1-AS1 or LNP–TYMSOS vaccines, combined with PD-1–laIL-2, inducing robust immune responses and effectively suppressing tumor progression in humanized mice (Figures 5H and S19).

Since humanized mouse models may not completely replicate the complexity of the human immune response, their reconstructed immune systems may not have encountered human proteins during the development of the transplanted human immune systems, potentially resulting in immunogenic responses driven by the recognition of these proteins as "foreign." To better assess the relevance of these *de novo* genes for anti-tumor mRNA vaccine applications, we predicted and synthesized peptides with high binding affinity to the most common HLA-A subtypes for *ELFN1-AS1* and *TYMSOS* (Table S16) and performed IFN-γ ELISpot assays to assess antigen-specific T cell responses in PBMCs from six colorectal cancer patients (Table S17). Notably, PBMCs co-incubated with a pool of *ELFN1-AS1* or *TYMSOS* peptides showed significantly elevated IFN-γ responses in two patients, respectively, demonstrating that antigens derived from these young *de novo* genes are immunogenic and capable of eliciting antigen-specific T cell activation in the host (Figure 5I; STAR Methods; Table S16). These findings support the potential of young human *de novo* genes as neoantigens for novel anti-tumor mRNA vaccine strategies.

DISCUSSION

Although previous studies have identified human *de novo* genes, their accurate identification and annotation remain challenging due to several intrinsic complexities—frequent occurrence in repetitive genomic regions, low and spatiotemporally restricted expression, and limited cross-species conservation. Moreover, rapid sequence divergence, gene loss, and distant homology further complicate the identification process.[1,49,76] These issues underscore the necessity for precise definitions of *de novo* genes within an evolving genomic framework driven by ongoing advancements in comparative, evolutionary, and functional genomics.[1,27] In this study, we implemented a rigorous, multi-stage pipeline to identify young *de novo* genes in humans. First, we utilized synteny-based reconstruction of ancestral genomic sequences through whole-genome alignments across 120 mammalian species and systematically traced the origin and evolutionary trajectories of candidate genes at their genomic loci (STAR Methods), building upon recent methodological advances in comparative genomic methods.[12,47–50] To assess coding potential across ancestral lineages, we adopted a 70% threshold, according to previous practices in human *de novo* gene identification,[6–8,26,48–50] which has been demonstrated to be robust for detecting most *de novo* ORFs in previous studies.[7,8,48] Second, we used annotated protein sequences from humans and 14 representative out-group species to exclude candidates with significant homology to annotated proteins. Third, to further refine the candidate set, we conducted additional BLASTp analyses and iterative jackhmmer searches against the UniProtKB database.[6,8,12,48,50,77] This allowed us to further eliminate ambiguous cases, including those that appear as follows: (1) orthologs in newly sequenced species not included in the initial 120-mammal dataset; (2) orthologs in species with low-quality genome assemblies; and (3) paralogs in species outside the human and 14 representative out-group species considered in the primary filtering steps. Finally, by integrating genomic profiles derived from 1,630 transcriptomes, 279 Ribo-seq datasets, and 100 million in-house-generated MS spectra, we identified 37 young *de novo* genes with strong evidence for both transcriptional and translational activity. Collectively, this set of 37 genes represents the most rigorously validated catalog of young human *de novo* genes to date.

Emerging evidence implicates young *de novo* genes in tumorigenesis,[19,21,22,45] paralleling their tumor-like roles in neuronal development[16,17] and spermatogenesis.[39,40] Here, we provide a systematic investigation of this connection, revealing a general upregulation and temporospatial expansion of these genes in tumor tissues. This suggests that the stringent expression restrictions governing these genes in normal tissues are relaxed during tumorigenesis. Our analyses propose that genome architecture reorganization, particularly ecDNAs, may drive this phenomenon. Notably, 16.4% of upregulated and 16.2% of expression-expanded genes in tumors are localized on ecDNAs, a proportion significantly higher than the background and similar to that of known oncogenes. Although further studies are certainly required to establish causality, these findings suggest a mechanism by which tumors relieve expression restrictions on young *de novo* genes and subsequently hijack their functions under physiological conditions to support tumor development, such as repressing apoptosis or maintaining the stem cell pool.

Functionally, these young genes directly contribute to tumorigenesis. In our CRISPR-Cas9-based screening, the depletion of 66.7% of expressed *de novo* genes resulted in a significant reduction in tumor cell proliferation, with 77.8% of these positive hits independently validated by siRNA knockdown assays. Many of these genes have been overlooked in cancer genomics, often misclassified as non-coding due to inconsistent annotation and excluded from traditional functional assays. Their strong functional impacts make them promising candidates for anti-tumor

Please cite this article in press as: Xiao et al., Oncogenic roles of young human *de novo* genes and their potential as neoantigens in cancer immunotherapy, Cell Genomics (2025), https://doi.org/10.1016/j.xgen.2025.100928

drug targets or neoantigens, potentially informing broad-spectrum therapies and preventive mRNA vaccine development.

Personalized mRNA vaccines, which utilize patient-specific tumor antigens, have shown significant promise.[78,79] However, broad-spectrum mRNA vaccines targeting multiple cancer types across diverse populations could provide a more cost-effective approach. The scarcity of shared, tumor-specific neoantigens has hindered progress in this area. The unique expression patterns of certain young *de novo* genes—active during early development, silenced in adult tissues, and reactivated exclusively in tumors—position them as ideal candidates for such vaccines. In this study, we developed two mRNA vaccines targeting *ELFN1-AS1* and *TYMSOS*, observing specific immune responses and effective tumor suppression in humanized mice. The antigens derived from these young genes are immunogenic and capable of eliciting antigen-specific T cell activation in colorectal cancer patients. These findings collectively support the potential of young human *de novo* genes as neoantigens for cancer immunotherapy.

Limitations of the study

First, the stringent criteria applied in *de novo* gene identification may introduce a risk of false negatives. For instance, accurately determining the age of genes and distinguishing the nature of transcripts and proteins based on phylogenetic trees require high-quality syntenic information across multiple species. As a result, rapidly evolving *de novo* genes with ambiguous syntenic alignments may be excluded. Moreover, compared to synteny-based analyses, BLASTp and jackhmmer searches are more sensitive but may also misidentify distant homologs due to genome assembly errors, repetitive sequences, and inconsistent annotations.[80,81] Notably, recent advances from model organisms[12,15,48] suggest several methodological improvements for human studies, including (1) constructing whole-genome alignments *de novo* using updated primate genome assemblies[30,31] and progressive alignment tools,[82] rather than relying on existing syntenic blocks, may enhance detection accuracy; (2) incorporating quantitative metrics like reading frame conservation scores[49,83] may improve ancestral state inference; and (3) beginning analyses with all translated ORFs identified through Ribo-seq and MS data,[48,50] rather than pre-selected candidates, could enable more comprehensive discovery. Second, the findings linking mechanism of ecDNA to temporospatially expanded expression of *de novo* gene in tumors should be interpreted with caution due to the limited sample size of *de novo* genes in our analyses. Future studies incorporating larger number of confirmed *de novo* genes, additional independent datasets, and experimental validation will be crucial to confirm and extend these observations. Finally, although the tumor-specific expression and functional relevance position these young *de novo* genes as previously unrecognized targets for novel therapeutic strategies, further investigations are required to bridge the gap between their potential and actual clinical applications.

RESOURCE AVAILABILITY

Lead contact

Requests for further information, resources, and reagents should be directed to and will be fulfilled by the lead contact, Chuan-Yun Li (chuanyunli@genetics.ac.cn).

Materials availability

The unique reagents generated during this research are accessible through the lead contact upon completion of a material(s) transfer agreement.

Data and code availability

- This paper analyzes existing, publicly available transcriptome and whole-genome sequencing data. The accession numbers for the datasets are listed in the key resources table.
- Public Ribo-seq datasets analyzed in this study are listed in the key resources table and Table S4.
- In-house-generated MS data in this study have been deposited to iProX[84] with project identifier IPX0008436003.
- Hi-C data analyzed in this study are listed in the key resources table and Table S10.
- Raw CRISPR-Cas9 screening data and RNA-seq data have been deposited at the NCBI Sequence Read Archive (SRA) as PRJNA1196182 and are publicly available as of the date of publication.
- The code generated during this study has been deposited and is publicly available at GitHub: https://github.com/Chunfu-Shawn/Denovo_genes-tumors and Figshare: https://doi.org/10.6084/m9.figshare.29177168.v1.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by grants from the National Natural Science Foundation of China (32300510), and the Chinese Institute for Brain Research (Beijing) (2020-NKX-XM-11). We thank Dr. Huan Yang at the Flow Cytometry Core, National Center for Protein Sciences, Peking University for technical support.

AUTHOR CONTRIBUTIONS

C.-Y.L. conceived the idea. C.-Y.L., N.A.A., and Q.C. designed the study. C.X. performed most of the data analysis. X.L. performed most of the experiments. P.L. and Q.C. developed mRNA vaccines. X.X., C.Y., L.Z., and Y.D. performed part of the data analysis. C.L., Q.X., T.G., Y.Q., Y.D., and C.H. performed part of the experiments. C.W., Y.W., and Z.P. supported the experiments with patients. C.-Y.L., C.X., X.L., and N.A.A. wrote the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Mice
  - Cell lines
- METHOD DETAILS
  - Identification of young *de novo* genes in humans
  - Expression profiles of *de novo* genes
  - Translational evidence for *de novo* genes
  - Properties of *de novo* genes
  - Differential expression analyses between healthy and tumor samples
  - Monte Carlo simulation
  - Genome architecture and ecDNA
  - sgRNA library cloning
  - CRISPR-Cas9 screening assay
  - Validations with siRNA knockdown and *ELFN1-AS1* knockout
  - Prognosis analysis

- mRNA vaccine preparation
- Mice vaccination for immune response evaluation and tumor challenge
- Flow cytometry
- Epitope peptides prediction
- T2 cell binding assay
- *Ex vivo* IFN-γ ELISpot assay

- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2025.100928.

Received: December 11, 2024
Revised: March 14, 2025
Accepted: June 2, 2025

REFERENCES

1. Zhao, L., Svetec, N., and Begun, D.J. (2024). De Novo Genes. Annu. Rev. Genet. *58*, 211–232. https://doi.org/10.1146/annurev-genet-111523-102413.
2. Liu, X., Chunfu, X., Xinwei, X., Jie, Z., Mo, F., Delihas, J.-y.C.N., and Li, C.-y. (2024). Origin of functional de novo genes in humans from "hopeful monsters". WIREs RNA *15*, e1845. https://doi.org/10.1002/wrna.1845.
3. Weisman, C.M. (2022). The Origins and Functions of De Novo Genes: Against All Odds? J. Mol. Evol. *90*, 244–257. https://doi.org/10.1007/s00239-022-10055-3.
4. Knowles, D.G., and McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. Genome Res. *19*, 1752–1759. https://doi.org/10.1101/gr.095026.109.
5. Li, C.Y., Zhang, Y., Wang, Z., Zhang, Y., Cao, C., Zhang, P.W., Lu, S.J., Li, X.M., Yu, Q., Zheng, X., et al. (2010). A human-specific de novo protein-coding gene associated with human brain functions. PLoS Comput. Biol. *6*, e1000734. https://doi.org/10.1371/journal.pcbi.1000734.
6. Wu, D.D., Irwin, D.M., and Zhang, Y.P. (2011). De novo origin of human protein-coding genes. PLoS Genet. *7*, e1002379. https://doi.org/10.1371/journal.pgen.1002379.
7. Xie, C., Zhang, Y.E., Chen, J.Y., Liu, C.J., Zhou, W.Z., Li, Y., Zhang, M., Zhang, R., Wei, L., and Li, C.Y. (2012). Hominoid-Specific De Novo Protein-Coding Genes Originating from Long Non-Coding RNAs. PLoS Genet. *8*, e1002942. https://doi.org/10.1371/journal.pgen.1002942.
8. Chen, J.Y., Shen, Q.S., Zhou, W.Z., Peng, J., He, B.Z., Li, Y., Liu, C.J., Luan, X., Ding, W., Li, S., et al. (2015). Emergence, Retention and Selection: A Trilogy of Origination for Functional De Novo Proteins from Ancestral LncRNAs in Primates. PLoS Genet. *11*, e1005391. https://doi.org/10.1371/journal.pgen.1005391.
9. Guerzoni, D., and McLysaght, A. (2016). De novo genes arise at a slow but steady rate along the primate lineage and have been subject to incomplete lineage sorting. Genome Biol. Evol. *8*, 1222–1232. https://doi.org/10.1093/gbe/evw074.
10. Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., and Albà, M.M. (2009). Origin of primate orphan genes: A comparative genomics approach. Mol. Biol. Evol. *26*, 603–612. https://doi.org/10.1093/molbev/msn281.
11. Xie, C., Bekpen, C., Künzel, S., Keshavarz, M., Krebs-Wheaton, R., Skrabar, N., Ullrich, K.K., and Tautz, D. (2019). A de novo evolved gene in the house mouse regulates female pregnancy cycles. eLife *8*, e44392. https://doi.org/10.7554/eLife.44392.
12. Peng, J., and Zhao, L. (2024). The origin and structural evolution of de novo genes in Drosophila. Nat. Commun. *15*, 810–814. https://doi.org/10.1038/s41467-024-45028-1.
13. Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A.R., Yu, Y., Hou, G., Zi, J., Zhou, R., et al. (2019). Rapid evolution of protein diversity by de novo origination in Oryza. Nat. Ecol. Evol. *3*, 679–690. https://doi.org/10.1038/s41559-019-0822-5.
14. Vakirlis, N., Hebert, A.S., Opulente, D.A., Achaz, G., Hittinger, C.T., Fischer, G., Coon, J.J., and Lafontaine, I. (2018). A Molecular Portrait of De Novo Genes in Yeasts. Mol. Biol. Evol. *35*, 631–645. https://doi.org/10.1093/molbev/msx315.
15. Vakirlis, N., Acar, O., Hsu, B., Castilho Coelho, N., Van Oss, S.B., Wacholder, A., Medetgul-Ernar, K., Bowman, R.W., Hines, C.P., Iannotta, J., et al. (2020). De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. Nat. Commun. *11*, 781. https://doi.org/10.1038/s41467-020-14500-z.
16. An, N.A., Zhang, J., Mo, F., Luan, X., Tian, L., Shen, Q.S., Li, X., Li, C., Zhou, F., Zhang, B., et al. (2023). De novo genes with an lncRNA origin encode unique human brain developmental functionality. Nat. Ecol. Evol. *7*, 264–278. https://doi.org/10.1038/s41559-022-01925-6.
17. Qi, J., Mo, F., An, N.A., Mi, T., Wang, J., Qi, J.T., Li, X., Zhang, B., Xia, L., Lu, Y., et al. (2023). A Human-Specific De Novo Gene Promotes Cortical Expansion and Folding. Adv. Sci. *10*, e2204140. https://doi.org/10.1002/advs.202204140.
18. Bekpen, C., Xie, C., and Tautz, D. (2018). Dealing with the adaptive immune system during de novo evolution of genes from intergenic sequences. BMC Evol. Biol. *18*, 121. https://doi.org/10.1186/s12862-018-1232-z.
19. Samusik, N., Krukovskaya, L., Meln, I., Shilov, E., and Kozlov, A.P. (2013). PBOV1 is a human de novo gene with tumor-specific expression that is associated with a positive clinical outcome of cancer. PLoS One *8*, e56162. https://doi.org/10.1371/journal.pone.0056162.
20. Yu, J., Ou, Z., Lei, Y., Chen, L., Su, Q., and Zhang, K. (2020). LncRNA MYCNOS facilitates proliferation and invasion in hepatocellular carcinoma by regulating miR-340. Hum. Cell *33*, 148–158. https://doi.org/10.1007/s13577-019-00303-y.
21. Fang, L., Wu, S., Zhu, X., Cai, J., Wu, J., He, Z., Liu, L., Zeng, M., Song, E., Li, J., et al. (2019). MYEOV functions as an amplified competing endogenous RNA in promoting metastasis by activating TGF-β pathway in NSCLC. Oncogene *38*, 896–912. https://doi.org/10.1038/s41388-018-0484-9.
22. Zhao, X., Li, D., Pu, J., Mei, H., Yang, D., Xiang, X., Qu, H., Huang, K., Zheng, L., and Tong, Q. (2016). CTCF cooperates with noncoding RNA MYCNOS to promote neuroblastoma progression through facilitating MYCN expression. Oncogene *35*, 3565–3576. https://doi.org/10.1038/onc.2015.422.
23. Papamichos, S.I., Margaritis, D., and Kotsianidis, I. (2015). Adaptive Evolution Coupled with Retrotransposon Exaptation Allowed for the Generation of a Human-Protein-Specific Coding Gene That Promotes Cancer Cell Proliferation and Metastasis in Both Haematological Malignancies and Solid Tumours: The Extraordinary Case of MYEOV Gene. Scientifica *2015*, 984706–984710. https://doi.org/10.1155/2015/984706.
24. Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T., and Albà, M.M. (2015). Origins of De Novo Genes in Human and Chimpanzee. PLoS Genet. *11*, e1005721. https://doi.org/10.1371/journal.pgen.1005721.
25. Wilson, B.A., Foy, S.G., Neme, R., and Masel, J. (2017). Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth. Nat. Ecol. Evol. *1*, 0146. https://doi.org/10.1038/s41559-017-0146.
26. Broeils, L.A., Ruiz-Orera, J., Snel, B., Hubner, N., and van Heesch, S. (2023). Evolution and implications of de novo genes in humans. Nat. Ecol. Evol. *7*, 804–815. https://doi.org/10.1038/s41559-023-02014-y.
27. Xiao, C., Mo, F., Lu, Y., Xiao, Q., Yao, C., Li, T., Qi, J., Liu, X., Chen, J.-y., Zhang, L., et al. (2024). Reply to : Identification of old coding regions disproves the hominoid de novo status of genes. Nat. Ecol. Evol. *8*, 1831–1834. https://doi.org/10.1038/s41559-024-02515-4.

28. Zhang, Y.E., Landback, P., Vibranovski, M., and Long, M. (2012). New genes expressed in human brains: Implications for annotating evolving genomes. Bioessays *34*, 982–991. https://doi.org/10.1002/bies.201200008.

29. Christmas, M.J., Kaplow, I.M., Genereux, D.P., Dong, M.X., Hughes, G. M., Li, X., Sullivan, P.F., Hindle, A.G., Andrews, G., Armstrong, J.C., et al. (2023). Evolutionary constraint and innovation across hundreds of placental mammals. Science *380*, eabn3943. https://doi.org/10.1126/SCIENCE.ABN3943.

30. Kuderna, L.F.K., Gao, H., Janiak, M.C., Kuhlwilm, M., Orkin, J.D., Bataillon, T., Manu, S., Valenzuela, A., Bergman, J., Rousselle, M., et al. (2023). A global catalog of whole-genome diversity from 233 primate species. Science *380*, 906–913. https://doi.org/10.1126/science.abn7829.

31. Shao, Y., Zhou, L., Li, F., Zhao, L., Zhang, B.L., Shao, F., Chen, J.W., Chen, C.Y., Bi, X., Zhuang, X.L., et al. (2023). Phylogenomic analyses provide insights into primate evolution. Science *380*, 913–924. https://doi.org/10.1126/science.abn6919.

32. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44–53. https://doi.org/10.1126/science.abj6987.

33. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223. https://doi.org/10.1126/science.1168978.

34. Strober, B.J., Wen, X., Wucher, V., Kwong, A., Lappalainen, T., Li, X., and Liang, Y. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *18*, 1318–1330. https://doi.org/10.1126/science.aaz1776.

35. Przybyla, L., and Gilbert, L.A. (2022). A new era in functional genomics screens. Nat. Rev. Genet. *23*, 89–103. https://doi.org/10.1038/s41576-021-00409-w.

36. Chen, J., Brunner, A.D., Cogan, J.Z., Nuñez, J.K., Fields, A.P., Adamson, B., Itzhak, D.N., Li, J.Y., Mann, M., Leonetti, M.D., and Weissman, J.S. (2020). Pervasive functional translation of noncanonical human open reading frames. Science *367*, 1140–1146. https://doi.org/10.1126/science.aay0262.

37. Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802. https://doi.org/10.1016/j.cell.2011.10.002.

38. Chothani, S.P., Adami, E., Widjaja, A.A., Langley, S.R., Viswanathan, S., Pua, C.J., Zhihao, N.T., Harmston, N., D'Agostino, G., Whiffin, N., et al. (2022). A high-resolution map of human RNA translation. Mol. Cell *82*, 2885–2899.e8. https://doi.org/10.1016/j.molcel.2022.06.023.

39. Gubala, A.M., Schmitz, J.F., Kearns, M.J., Vinh, T.T., Bornberg-Bauer, E., Wolfner, M.F., and Findlay, G.D. (2017). The Goddard and Saturn Genes Are Essential for Drosophila Male Fertility and May Have Arisen de Novo. Mol. Biol. Evol. *34*, 1066–1082. https://doi.org/10.1093/molbev/msx057.

40. McLysaght, A., and Hurst, L.D. (2016). Open questions in the study of de novo genes: what, how and why. Nat. Rev. Genet. *17*, 567–578. https://doi.org/10.1038/nrg.2016.78.

41. Rivard, E.L., Ludwig, A.G., Patel, P.H., Grandchamp, A., Arnold, S.E., Berger, A., Scott, E.M., Kelly, B.J., Mascha, G.C., Bornberg-Bauer, E., and Findlay, G.D. (2021). A putative de novo evolved gene required for spermatid chromatin condensation in Drosophila melanogaster. PLoS Genet. *17*, e1009787. https://doi.org/10.1371/journal.pgen.1009787.

42. Kaneko, Y., Suenaga, Y., Islam, S.M.R., Matsumoto, D., Nakamura, Y., Ohira, M., Yokoi, S., and Nakagawara, A. (2015). Functional interplay between MYCN, NCYM, and OCT4 promotes aggressiveness of human neuroblastomas. Cancer Sci. *106*, 840–847. https://doi.org/10.1111/cas.12677.

43. Buhl, A.M., Jurlander, J., Jørgensen, F.S., Ottesen, A.M., Cowland, J.B., Gjerdrum, L.M., Hansen, B.V., and Leffers, H. (2006). Identification of a gene on chromosome 12q22 uniquely overexpressed in chronic lymphocytic leukemia. Blood *107*, 2904–2911. https://doi.org/10.1182/blood-2005-07-2615.

44. Suenaga, Y., Islam, S.M.R., Alagu, J., Kaneko, Y., Kato, M., Tanaka, Y., Kawana, H., Hossain, S., Matsumoto, D., Yamamoto, M., et al. (2014). NCYM, a Cis-Antisense Gene of MYCN, Encodes a De Novo Evolved Protein That Inhibits GSK3β Resulting in the Stabilization of MYCN in Human Neuroblastomas. PLoS Genet. *10*, e1003996. https://doi.org/10.1371/journal.pgen.1003996.

45. Ma, C., Li, C., Ma, H., Yu, D., Zhang, Y., Zhang, D., Su, T., Wu, J., Wang, X., Zhang, L., et al. (2022). Pan-cancer surveys indicate cell cycle-related roles of primate-specific genes in tumors and embryonic cerebrum. Genome Biol. *23*, 251. https://doi.org/10.1186/s13059-022-02821-9.

46. Camarena, M.E., Theunissen, P., Ruiz, M., Ruiz-Orera, J., Calvo-Serra, B., Castelo, R., Castro, C., Sarobe, P., Fortes, P., Perera-Bel, J., and Albà, M.M. (2024). Microproteins encoded by noncanonical ORFs are a major source of tumor-specific antigens in a liver cancer patient meta-cohort. Sci. Adv. *10*, eadn3628. https://doi.org/10.1126/sciadv.adn3628.

47. Vakirlis, N., and McLysaght, A. (2019). Computational Prediction of De Novo Emerged Protein-Coding Genes. In Computational Methods in Protein Evolution, T. Sikosek, ed. (Springer), pp. 63–81. https://doi.org/10.1007/978-1-4939-8736-8_4.

48. Vakirlis, N., Vance, Z., Duggan, K.M., and McLysaght, A. (2022). De novo birth of functional microproteins in the human lineage. Cell Rep. *41*, 111808. https://doi.org/10.1016/j.celrep.2022.111808.

49. Vakirlis, N., Acar, O., Cherupally, V., and Carvunis, A.R. (2024). Ancestral Sequence Reconstruction as a Tool to Detect and Study De Novo Gene Emergence. Genome Biol. Evol. *16*, evae151. https://doi.org/10.1093/gbe/evae151.

50. Sandmann, C.L., Schulz, J.F., Ruiz-Orera, J., Kirchner, M., Ziehm, M., Adami, E., Marczenke, M., Christ, A., Liebe, N., Greiner, J., et al. (2023). Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. Mol. Cell *83*, 994–1011.e18. https://doi.org/10.1016/j.molcel.2023.01.023.

51. Roginski, P., Grandchamp, A., Quignot, C., and Lopes, A. (2024). De Novo Emerged Gene Search in Eukaryotes with DENSE. Genome Biol. Evol. *16*, evae159. https://doi.org/10.1093/gbe/evae159.

52. Hecker, N., and Hiller, M. (2020). A genome alignment of 120 mammals highlights ultraconserved element variability and placenta-associated enhancers. GigaScience *9*, giz159. https://doi.org/10.1093/gigascience/giz159.

53. Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B., et al. (2012). Proto-genes and de novo gene birth. Nature *487*, 370–374. https://doi.org/10.1038/nature11184.

54. Li, Y., Li, C., Li, S., Peng, Q., An, N.A., He, A., and Li, C.Y. (2018). Human exonization through differential nucleosome occupancy. Proc. Natl. Acad. Sci. USA *115*, 8817–8822. https://doi.org/10.1073/pnas.1802561115.

55. Avgan, N., Wang, J.I., Fernandez-Chamorro, J., and Weatheritt, R.J. (2019). Multilayered control of exon acquisition permits the emergence of novel forms of regulatory control. Genome Biol. *20*, 141. https://doi.org/10.1186/s13059-019-1757-5.

56. Oliver, J.L., and Marín, A. (1996). A relationship between GC content and coding-sequence length. J. Mol. Evol. *43*, 216–223. https://doi.org/10.1007/bf02338829.

57. Monsellier, E., and Chiti, F. (2007). Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO Rep. *8*, 737–742. https://doi.org/10.1038/sj.embor.7401034.

58. Yu, J.F., Cao, Z., Yang, Y., Wang, C.L., Su, Z.D., Zhao, Y.W., Wang, J.H., and Zhou, Y. (2016). Natural protein sequences are more intrinsically disordered than random sequences. Cell. Mol. Life Sci. *73*, 2949–2957. https://doi.org/10.1007/s00018-016-2138-9.

59. Wright, P.E., and Dyson, H.J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. Nat. Rev. Mol. Cell Biol. *16*, 18–29. https://doi.org/10.1038/nrm3920.

60. Chen, S.C.-C., Chuang, T.-J., and Li, W.-H. (2011). The Relationships Among MicroRNA Regulation, Intrinsically Disordered Regions, and Other Indicators of Protein Evolutionary Rate. Mol. Biol. Evol. *28*, 2513–2520. https://doi.org/10.1093/molbev/msr068.

61. Kesner, J.S., Chen, Z., Shi, P., Aparicio, A.O., Murphy, M.R., Guo, Y., Trehan, A., Lipponen, J.E., Recinos, Y., Myeku, N., and Wu, X. (2023). Noncoding translation mitigation. Nature *617*, 395–402. https://doi.org/10.1038/s41586-023-05946-4.

62. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell *169*, 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038.

63. Cardoso-Moreira, M., Halbert, J., Valloton, D., Velten, B., Chen, C., Shao, Y., Liechti, A., Ascenção, K., Rummel, C., Ovchinnikova, S., et al. (2019). Gene expression across mammalian organ development. Nature *571*, 505–509. https://doi.org/10.1038/s41586-019-1338-5.

64. Witt, E., Benjamin, S., Svetec, N., and Zhao, L. (2019). Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in drosophila. eLife *8*, e47138. https://doi.org/10.7554/eLife.47138.

65. Hung, K.L., Yost, K.E., Xie, L., Shi, Q., Helmsauer, K., Luebeck, J., Schöpflin, R., Lange, J.T., Chamorro González, R., Weiser, N.E., et al. (2021). ecDNA hubs drive cooperative intermolecular oncogene expression. Nature *600*, 731–736. https://doi.org/10.1038/s41586-021-04116-8.

66. Turner, K.M., Deshpande, V., Beyter, D., Koga, T., Rusert, J., Lee, C., Li, B., Arden, K., Ren, B., Nathanson, D.A., et al. (2017). Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature *543*, 122–125. https://doi.org/10.1038/nature21356.

67. Wu, S., Turner, K.M., Nguyen, N., Raviram, R., Erb, M., Santini, J., Luebeck, J., Rajkumar, U., Diao, Y., Li, B., et al. (2019). Circular ecDNA promotes accessible chromatin and high oncogene expression. Nature *575*, 699–703. https://doi.org/10.1038/s41586-019-1763-5.

68. Verhaak, R.G.W., Bafna, V., and Mischel, P.S. (2019). Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. Nat. Rev. Cancer *19*, 283–288. https://doi.org/10.1038/s41568-019-0128-6.

69. Deng, S., Feng, Y., and Pauklin, S. (2022). 3D chromatin architecture and transcription regulation in cancer. J. Hematol. Oncol. *15*, 49. https://doi.org/10.1186/s13045-022-01271-x.

70. Zhao, S.G., Bootsma, M., Zhou, S., Shrestha, R., Moreno-Rodriguez, T., Lundberg, A., Pan, C., Arlidge, C., Hawley, J.R., Foye, A., et al. (2024). Integrated analyses highlight interactions between the three-dimensional genome and DNA, RNA and epigenomic alterations in metastatic prostate cancer. Nat. Genet. *56*, 1689–1700. https://doi.org/10.1038/s41588-024-01826-3.

71. Zhong, T., Wang, W., Liu, H., Zeng, M., Zhao, X., and Guo, Z. (2023). eccDNA Atlas: a comprehensive resource of eccDNA catalog. Brief. Bioinform. *24*, bbad037. https://doi.org/10.1093/bib/bbad037.

72. Sharma, A., Blériot, C., Currenti, J., and Ginhoux, F. (2022). Oncofetal reprogramming in tumour development and progression. Nat. Rev. Cancer *22*, 593–602. https://doi.org/10.1038/s41568-022-00497-8.

73. Polack, F.P., Thomas, S.J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J.L., Pérez Marc, G., Moreira, E.D., Zerbini, C., et al. (2020). Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. N. Engl. J. Med. *383*, 2603–2615. https://doi.org/10.1056/NEJMoa2034577.

74. Baden, L.R., El Sahly, H.M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D., Spector, S.A., Rouphael, N., Creech, C.B., et al. (2021). Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. N. Engl. J. Med. *384*, 403–416. https://doi.org/10.1056/NEJMoa2035389.

75. Ren, Z., Zhang, A., Sun, Z., Liang, Y., Ye, J., Qiao, J., Li, B., and Fu, Y.-X. (2022). Selective delivery of low-affinity IL-2 to PD-1+ T cells rejuvenates antitumor immunity with reduced toxicity. J. Clin. Investig. *132*, e153604. https://doi.org/10.1172/JCI153604.

76. Xia, S., Chen, J., Arsala, D., Emerson, J.J., and Long, M. (2025). Functional innovation through new genes as a general evolutionary process. Nat. Genet. *57*, 295–309. https://doi.org/10.1038/s41588-024-02059-0.

77. Shao, Y., Chen, C., Shen, H., He, B.Z., Yu, D., Jiang, S., Zhao, S., Gao, Z., Zhu, Z., Chen, X., et al. (2019). GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. Genome Res. *29*, 682–696. https://doi.org/10.1101/gr.238733.118.

78. Hu, Z., Ott, P.A., and Wu, C.J. (2018). Towards personalized, tumour-specific, therapeutic vaccines for cancer. Nat. Rev. Immunol. *18*, 168–182. https://doi.org/10.1038/nri.2017.131.

79. Lang, F., Schrörs, B., Löwer, M., Türeci, Ö., and Sahin, U. (2022). Identification of neoantigens for individualized therapeutic cancer vaccines. Nat. Rev. Drug Discovery *21*, 261–282. https://doi.org/10.1038/s41573-021-00387-y.

80. Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., et al. (2018). Ten things you should know about transposable elements. Genome Biol. *19*, 199. https://doi.org/10.1186/s13059-018-1577-z.

81. Tørresen, O.K., Star, B., Mier, P., Andrade-Navarro, M.A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A.V., Promponas, V.J., et al. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. *47*, 10994–11006. https://doi.org/10.1093/nar/gkz841.

82. Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I.T., Novak, A.M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., et al. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature *587*, 246–251. https://doi.org/10.1038/s41586-020-2871-y.

83. Wacholder, A., Parikh, S.B., Coelho, N.C., Acar, O., Houghton, C., Chou, L., and Carvunis, A.-R. (2023). A vast evolutionarily transient translatome contributes to phenotype and fitness. Cell Syst. *14*, 363–381.e8. https://doi.org/10.1016/j.cels.2023.04.002.

84. Ma, J., Chen, T., Wu, S., Yang, C., Bai, M., Shu, K., Li, K., Zhang, G., Jin, Z., He, F., et al. (2019). iProX: an integrated proteome resource. Nucleic Acids Res. *47*, D1211–D1217. https://doi.org/10.1093/nar/gky869.

85. Wang, Z.Y., Leushkin, E., Liechti, A., Ovchinnikova, S., Mößinger, K., Brüning, T., Rummel, C., Grützner, F., Cardoso-Moreira, M., Janich, P., et al. (2020). Transcriptome and translatome co-evolution in mammals. Nature *588*, 642–647. https://doi.org/10.1038/s41586-020-2899-z.

86. Duffy, E.E., Finander, B., Choi, G., Carter, A.C., Pritisanac, I., Alam, A., Luria, V., Karger, A., Phu, W., Sherman, M.A., et al. (2022). Developmental dynamics of RNA translation in the human brain. Nat. Neurosci. *25*, 1353–1365. https://doi.org/10.1038/s41593-022-01164-9.

87. Loayza-Puch, F., Rooijers, K., Buil, L.C.M., Zijlstra, J., Oude Vrielink, J.F., Lopes, R., Ugalde, A.P., Van Breugel, P., Hofland, I., Wesseling, J., et al. (2016). Tumour-specific proline vulnerability uncovered by differential ribosome codon reading. Nature *530*, 490–494. https://doi.org/10.1038/nature16982.

88. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461. https://doi.org/10.1038/nature12787.

89. Wang, J., Huang, T.Y.-T., Hou, Y., Bartom, E., Lu, X., Shilatifard, A., Yue, F., and Saratsis, A. (2021). Epigenomic landscape and 3D genome structure in pediatric high-grade glioma. Sci. Adv. *7*, eabg4126. https://doi.org/10.1126/sciadv.abg4126.

90. Iyyanki, T., Zhang, B., Wang, Q., Hou, Y., Jin, Q., Xu, J., Yang, H., Liu, T., Wang, X., Song, F., et al. (2021). Subtype-associated epigenomic

landscape and 3D genome structure in bladder cancer. Genome Biol. *22*, 105. https://doi.org/10.1186/s13059-021-02325-y.

91. Kim, T., Han, S., Chun, Y., Yang, H., Min, H., Jeon, S.Y., Kim, J.-i., Moon, H.-G., and Lee, D. (2022). Comparative characterization of 3D chromatin organization in triple-negative breast cancers. Exp. Mol. Med. *54*, 585–600. https://doi.org/10.1038/s12276-022-00768-2.
92. Du, Y., Gu, Z., Li, Z., Yuan, Z., Zhao, Y., Zheng, X., Bo, X., Chen, H., and Wang, C. (2022). Dynamic Interplay between Structural Variations and 3D Genome Organization in Pancreatic Cancer. Adv. Sci. (Weinh) *9*, 2270113. https://doi.org/10.1002/advs.202270113.
93. Akdemir, K.C., Le, V.T., Chandran, S., Li, Y., Verhaak, R.G., Beroukhim, R., Campbell, P.J., Chin, L., Dixon, J.R., Futreal, P.A., et al. (2020). Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. Nat. Genet. *52*, 294–305. https://doi.org/10.1038/s41588-019-0564-y.
94. Ooi, W.F., Nargund, A.M., Lim, K.J., Zhang, S., Xing, M., Mandoli, A., Lim, J.Q., Ho, S.W.T., Guo, Y., Yao, X., et al. (2020). Integrated paired-end enhancer profiling and whole-genome sequencing reveals recurrent CCNE1 and IGF2 enhancer hijacking in primary gastric adenocarcinoma. Gut *69*, 1039–1052. https://doi.org/10.1136/gutjnl-2018-317612.
95. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.
96. Nueda, M.J., Tarazona, S., and Conesa, A. (2014). Next maSigPro: Updating maSigPro bioconductor package for RNA-seq time series. Bioinformatics *30*, 2598–2602. https://doi.org/10.1093/bioinformatics/btu333.
97. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.
98. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
99. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. j. *17*, 10. https://doi.org/10.14806/ej.17.1.200.
100. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.
101. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.
102. Zhang, P., He, D., Xu, Y., Hou, J., Pan, B.F., Wang, Y., Liu, T., Davis, C.M., Ehli, E.A., Tan, L., et al. (2017). Genome-wide identification and differential analysis of translational initiation. Nat. Commun. *8*, 1749. https://doi.org/10.1038/s41467-017-01981-8.
103. Chi, H., Liu, C., Yang, H., Zeng, W.F., Wu, L., Zhou, W.J., Wang, R.M., Niu, X.N., Ding, Y.H., Zhang, Y., et al. (2018). Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. Nat. Biotechnol. *36*, 1059–1066. https://doi.org/10.1038/nbt.4236.
104. Erdős, G., and Dosztányi, Z. (2024). AIUPred: Combining energy estimation with deep learning for the enhanced prediction of protein disorder. Nucleic Acids Res. *52*, W176–W181. https://doi.org/10.1093/nar/gkae385.
105. Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. *32*, 896–902. https://doi.org/10.1038/nbt.2931.
106. Osorio, D., Rondón-Villarreal, P., and Torres, R. (2015). Peptides: A Package for Data Mining of Antimicrobial Peptides. The R Journal *7*, 4–14. https://doi.org/10.32614/RJ-2015-001.
107. Luebeck, J., Huang, E., Kim, F., Liefeld, T., Dameracharla, B., Ahuja, R., Schreyer, D., Prasad, G., Adamaszek, M., Kenkre, R., et al. (2024). AmpliconSuite: an end-to-end workflow for analyzing focal amplifications in cancer genomes. Preprint at bioRxiv. https://doi.org/10.1101/2024.05.06.592768.
108. Kruse, K., Hug, C.B., and Vaquerizas, J.M. (2020). FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. Genome Biol. *21*, 303. https://doi.org/10.1186/s13059-020-02215-9.
109. Salameh, T.J., Wang, X., Song, F., Zhang, B., Wright, S.M., Khunsriraksakul, C., Ruan, Y., and Yue, F. (2020). A supervised learning framework for chromatin loop detection in genome-wide contact maps. Nat. Commun. *11*, 3428. https://doi.org/10.1038/s41467-020-17239-9.
110. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.
111. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.
112. Li, W., Xu, H., Xiao, T., Cong, L., Love, M.I., Zhang, F., Irizarry, R.A., Liu, J.S., Brown, M., and Liu, X.S. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. *15*, 554. https://doi.org/10.1186/s13059-014-0554-4.
113. Hubisz, M.J., Pollard, K.S., and Siepel, A. (2011). PHAST and RPHAST: phylogenetic analysis with space/time models. Brief. Bioinform. *12*, 41–51. https://doi.org/10.1093/bib/bbq072.
114. Löytynoja, A., and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. Proc. Natl. Acad. Sci. USA *102*, 10557–10562. https://doi.org/10.1073/pnas.0409137102.
115. Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput. Biol. *7*, e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
116. UniProt Consortium (2025). UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res. *53*, D609–D617. https://doi.org/10.1093/nar/gkae1010.
117. Eisenberg, E., and Levanon, E.Y. (2013). Human housekeeping genes, revisited. Trends Genet. *29*, 569–574. https://doi.org/10.1016/j.tig.2013.05.010.
118. Hounkpe, B.W., Chenou, F., de Lima, F., and de Paula, E.V. (2021). HRT Atlas v1.0 database: Redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. Nucleic Acids Res. *49*, D947–D955. https://doi.org/10.1093/nar/gkaa609.
119. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. Science *347*, 1260419. https://doi.org/10.1126/science.1260419.
120. Hennig, T., Michalski, M., Rutkowski, A.J., Djakovic, L., Whisnant, A.W., Friedl, M.S., Jha, B.A., Baptista, M.A.P., L'Hernault, A., Erhard, F., et al. (2018). HSV-1-induced disruption of transcription termination resembles a cellular stress response but selectively increases chromatin accessibility downstream of genes. PLoS Pathog. *14*, e1006954. https://doi.org/10.1371/journal.ppat.1006954.
121. Uversky, V.N., and Dunker, A.K. (2010). Understanding protein non-folding. Biochim. Biophys. Acta *1804*, 1231–1264. https://doi.org/10.1016/j.bbapap.2010.01.017.
122. Aran, D., Camarda, R., Odegaard, J., Paik, H., Oskotsky, B., Krings, G., Goga, A., Sirota, M., and Butte, A.J. (2017). Comprehensive analysis of normal adjacent to tumor transcriptomes. Nat. Commun. *8*, 1077. https://doi.org/10.1038/s41467-017-01027-z.
123. Suehnholz, S.P., Nissan, M.H., Zhang, H., Kundra, R., Nandakumar, S., Lu, C., Carrero, S., Dhaneshwar, A., Fernandez, N., Xu, B.W., et al.

(2024). Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer. Cancer Discov. *14*, 49–65. https://doi.org/10.1158/2159-8290.Cd-23-0467.

124. Rennie, S., Dalby, M., Van Duin, L., and Andersson, R. (2018). Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions. Nat. Commun. *9*, 487. https://doi.org/10.1038/s41467-017-02798-1.

125. Chen, H., Li, C., Peng, X., Zhou, Z., Weinstein, J.N., Liang, H., Felau, I., Kasapi, M., Ferguson, M.L., et al.; Cancer Genome Atlas Research Network (2018). A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. Cell *173*, 386–399.e12. https://doi.org/10.1016/j.cell.2018.03.027.

126. Ma, J., Köster, J., Qin, Q., Hu, S., Li, W., Chen, C., Cao, Q., Wang, J., Mei, S., Liu, Q., et al. (2016). CRISPR-DO for genome-wide CRISPR design and optimization. Bioinformatics *32*, 3336–3338. https://doi.org/10.1093/bioinformatics/btw476.

127. Wang, T., Lander, E.S., and Sabatini, D.M. (2016). Viral Packaging and Cell Culture for CRISPR-Based Screens. Cold Spring Harb. Protoc. https://doi.org/10.1101/pdb.prot090811.

Please cite this article in press as: Xiao et al., Oncogenic roles of young human *de novo* genes and their potential as neoantigens in cancer immunotherapy, Cell Genomics (2025), https://doi.org/10.1016/j.xgen.2025.100928

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Brilliant Violet 650 anti-human CD3	Biolegend	Cat#317324; RRID: AB_2563352
FITC anti-human CD45	Biolegend	Cat#304005; RRID: AB_314393
Brilliant Violet 570 anti-human CD4	Biolegend	Cat#300533; RRID: AB_10896788
Brilliant Violet 421 anti-human CD137 (4-1BB)	Biolegend	Cat#309819; RRID: AB_10895902
PE-Cy7 Anti-Human CD8a	Biolegend	Cat#300913; RRID: AB_314117
Fixable Viability Dye eFluor 780	Invitrogen	Cat#65-0865-14; RRID: AB_3113075
PE anti-human CD11c	Biolegend	Cat#337205; RRID: AB_1236439
FITC anti-human CD83	Biolegend	Cat#305305; RRID: AB_314513
Brilliant Violet 421 anti-human CD80	Biolegend	Cat#305221; RRID: AB_10899567
APC anti-human CD86	Biolegend	Cat#374207; RRID: AB_2721448
7-AAD Viability Staining Solution	Biolegend	Cat#420404; RRID: AB_408812
PE HLA-A2 Monoclonal Antibody (BB7.2)	Invitrogen	Cat#MA1-80303; RRID: AB_931640
Monoclonal ANTI-FLAG M2 antibody	Sigma-Aldrich	Cat#F1804; RRID: AB_262044
GAPDH Mouse Monoclonal antibody (2BB)	Biodragon	Cat#B3029
Chemicals, peptides, and recombinant proteins		
Fetal Bovine Serum	Gibco	Cat#10091-148
Penicillin-Streptomycin	Gibco	Cat#15070-063
Trypsin-EDTA	Gibco	Cat#25200-056
DMEM	Gibco	Cat#C11995500BT
RPMI-1640	Gibco	Cat#C11875500BT
FcR Blocking Reagent human	Milteny	Cat#130-059-901
DPBS	Gibco	Cat#C14190500BT
Elispot Medium	Dakewei	Cat#6115012
SODIUM DODECYL SULFATE (SDS)	AMRESCO	Cat# 0227-1KG
RIPA buffer	Solarbio	Cat#R0010
Proteinase K	TIANGEN	Cat#RT403
Q5 HiFi DNA polymerase	NEB	Cat#M0491L
Glycine	AMRESCO	Cat#0167-1KG
HEPES free acid	AMRESCO	Cat#0511-1KG
Sodium chloride (NaCl)	SIGMA	Cat#S7653
0.5M EDTA PH = 0.8	Invitrogen	Cat#AM9260G
Amicon Ultra-0.5ML Centrifuge Filters	Millipore	Cat#UFC5010BK
Pur-A-Lyzer Midi Dialysis Kits	Sigma-Aldrich	Cat#PURD35100-1KT
SM-102 (8-[(2-hydroxyethyl)[6-oxo-6-(undecyloxy)hexyl]amino]-octanoic acid, 1-octylnonyl ester; CAS:2089251-47-6)	SINOPEG	Cat#06040008800
DSPC (1,2-distearoyl-*sn*-glycero-3-phosphocholine; CAS:816-94-4)	SINOPEG	Cat#06030001100
Cholesterol	SINOPEG	Cat#06040010300
DMG-PEG2000(1,2-dimyristoyl-rac-*glycero*-3-methoxypolyethylene glycol-2000; CAS:160743-62-4)	SINOPEG	Cat#06020112402
KOD One PCR Master Mix	Toyobo	Cat#KMM-201
ClonExpress II One Step Cloning Kit	Vazyme	Cat#C112-02

(*Continued on next page*)

Please cite this article in press as: Xiao et al., Oncogenic roles of young human *de novo* genes and their potential as neoantigens in cancer immunotherapy, Cell Genomics (2025), https://doi.org/10.1016/j.xgen.2025.100928

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
T7 High Yield RNA Transcription Kit (Nl-Me-Pseudo UTP)	Vazyme	Cat#DD4202-01
EasyCap GAG m7G(5′)ppp(5′)(2′OMeA)pG	Synthgene	Cat#CAP3011
Critical commercial assays		
DNA Clean & Concentrator-25 Kit	Zymo	Cat#D4006
Human IFN-γ ELISpot plus kit	MABTECH	Cat#3420-4HST-2
Deposited data		
120-mammals whole-genome alignments	Hecker et al.[52]	https://bds.mpi-cbg.de/hillerlab/120MammalAlignment/
Primate net alignments	UCSC Genome Browser GRCh38/hg38 assembly	https://hgdownload.soe.ucsc.edu/goldenPath/hg38/
Annotated proteomes of 14 representative outgroup species	Ensembl	https://ensembl.org/biomart/martview
GTEx RNA sequencing data for 27 normal tissue types	dbGAP	phs000424.v10.p2
RNA sequencing data for 6 organs (forebrain, hindbrain, heart, kidney, liver and testis) across developmental stages	Cardoso-Moreira et al.[63]	ArrayExpress: E-MTAB-6814
TCGA RNA sequencing data for 22 tumor types	Genomic Data Commons	phs000178.v11.p8
RNA sequencing data for 5 cell lines (U87, SW480, HeLa, Cal62, and OS-RC-2)	This paper	SRA: PRJNA1196182
Ribosome profiling sequencing data for multiple human tissues	Chothani et al.[38]	PRJNA756018: SRR15513148-SRR15513226
Ribosome profiling sequencing data for human brain, liver and testis	Wang et al.[85]	ArrayExpress: E-MTAB-7247
Ribosome profiling sequencing data for human embryonic stem cell neuronal cultures and human cortex	Duffy et al.[86]	PRJNA743949: SRR15175563-SRR15175568, SRR19165065-SRR19165070, SRR15906422-SRR15906494
Ribosome profiling sequencing data for human kidney and kidney tumors	Loayza-Puch et al.[87]	PRJNA256316: SRR1528686-SRR1528697
Raw in-house-generated mass spectrometry data	This paper	iProX: IPX0008436003
Peptide-spectrum matches identified in MassIVE	MassIVE	https://massive.ucsd.edu/ProteoSAFe/massive_search.jsp (released in 11/30/2023)
Peptide-spectrum matches identified in PeptideAtlas	PeptideAtlas	https://peptideatlas.org/builds/human/202401/APD_Hs_all.fasta
TCGA whole-genome sequencing data for BLCA samples	Genomic Data Commons	phs000178.v11.p8
Raw CRISPR/Cas9 screening data	This paper	SRA: PRJNA1196182
FANTOM enhancer annotation	Andersson et al.[88]	https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/extra/enhancer/F5.hg38.enhancers.bed.gz
GBM cell line and NHA cell line raw Hi-C data	Wang et al.[89]	GSE162976
BLCA cell line raw Hi-C data	Iyyanki et al.[90]	GSE148079
BRCA cell line raw Hi-C data	Kim et al.[91]	GSE167150
PAAD cell line raw Hi-C data	Du et al.[92]	GSE185069
COAD & ESCA cell line raw Hi-C data	Akdemir et al.[93]	GSE116694
STAD primary sample raw Hi-C data	Ooi et al.[94]	GSE118391

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ecDNA annotations from eccDNA Atlas	Zhong et al.[71]	https://lcbb.swjtu.edu.cn/eccDNAatlas/download/Homo%20sapiens.xlsx
Experimental models: Cell lines		
HEK293T cells	From Laboratory of Dr. Ying Liu	CL-0063
CAL62 cell	Procell Life Science & Technology Co.,Ltd.	CL-0618
HeLa cell	Procell Life Science & Technology Co.,Ltd.	CL-0101
SW480 cell	Procell Life Science & Technology Co.,Ltd.	CL-0223
OS-RC-2 cell	Procell Life Science & Technology Co.,Ltd.	CL-0177
U-87 MG cell	Procell Life Science & Technology Co.,Ltd.	CL-0238
Experimental models: Organisms/strains		
Mouse: huHSC-NKG-ProF	Cyagen Biosciences Co.,Ltd	N/A
Oligonucleotides		
CRISPR screening 1st-F	TSINGKE	This paper
CRISPR screening 1st-R	TSINGKE	This paper
CRISPR screening 2nd-F	TSINGKE	This paper
CRISPR screening 2nd-R-index1	TSINGKE	This paper
CRISPR screening 2nd-R-index2	TSINGKE	This paper
CRISPR screening 2nd-R-index3	TSINGKE	This paper
CRISPR screening 2nd-R-index4	TSINGKE	This paper
CRISPR screening 3rd-F	TSINGKE	This paper
CRISPR screening 3rd-F	TSINGKE	This paper
Recombinant DNA		
Plasmid: LentiCrispr-V2	From Laboratory of Dr. Shaokun Shu	N/A
Software and algorithms		
PHAST v1.4	Cold Spring Harbor Laboratory	http://compgen.cshl.edu/phast/
PRANK v170427	Löytynoja lab	https://ariloytynoja.github.io/prank-msa/
featureCounts v2.0.2	Liao et al.[95]	http://subread.sourceforge.net/
maSigPro v1.70.0	Nueda et al.[96]	https://www.bioconductor.org/packages/release/bioc/html/maSigPro.html
fastp v0.23.4	Chen et al.[97]	https://github.com/OpenGene/fastp
Trimmomatic v0.39	Bolger et al.[98]	http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.39.zip
Cutadapt v4.5	Martin[99]	https://cutadapt.readthedocs.io/en/stable/installation.html
Bowtie2 v.2.5.3	Langmead et al.[100]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
STAR v2.7.9a	Dobin et al.[101]	https://github.com/alexdobin/STAR/releases
RiboTISH v0.2.7	Zhang et al.[102]	https://github.com/zhpn1024/ribotish
pFind v3.2	Chi et al.[103]	http://pfind.org/downloads.html
AIUPred v0.9	Erdos et al.[104]	https://aiupred.elte.hu/download
RUVSeq v.1.31.0	Risso et al.[105]	https://github.com/drisso/RUVSeq
Peptides v2.4.6	Osorio et al.[106]	https://github.com/dosorio/Peptides/
AmpliconSuite-pipeline v1.3.4	Luebeck et al.[107]	https://github.com/AmpliconSuite/AmpliconSuite-pipeline
CycleViz v0.2.0	From Jens Luebeck	https://github.com/AmpliconSuite/CycleViz
runHiC v0.8.7	From Xiaotao Wang	https://github.com/XiaoTaoWang/HiC_pipeline

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
JuicerTools v1.13.02	From aidenlab	https://github.com/aidenlab/JuicerTools
FAN-C v0.9.26	Kruse et al.[108]	https://github.com/vaquerizaslab/fanc
Peakachu v1.2.0	Salameh et al.[109]	https://github.com/tariks/peakachu
HiCPeaks v0.3.7	From Xiaotao Wang	https://github.com/XiaoTaoWang/HiCPeaks
BEDTools v2.26.0	Quinlan et al.[110]	https://bedtools.readthedocs.io/en/latest/
SAMtools v1.16.1	Danecek et al.[111]	https://github.com/samtools/samtools
MAGeCK v0.5.9.5	Li et al.[112]	https://sourceforge.net/p/mageck/wiki/install/
survival v3.5.8	From Terry M. Therneau	https://github.com/therneau/survival
NetMHCIIpan 4.1	The Immune Epitope Database	http://tools.iedb.org/mhcii/
Python v3.9.18	Python Software Foundation	https://www.python.org/
R v4.2.1	The R Foundation for Statistical Computing	https://www.r-project.org/

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Mice

HuHSC-C-NKG-ProF mice were obtained from Cyagen Biosciences. Three-day-old female C-NKG mice received 1 Gy irradiation and were then injected with umbilical cord blood-derived hematopoietic stem cells *via* the superficial temporal vein. Human immune cell reconstitution was assessed ten weeks post-injection using flow cytometry. Humanized female mice after 12 weeks of successful transplantation are used for follow-up experiments. All animal experiments were approved by the Peking University Institutional Animal Care and Use Committee (IMM-LiCY-1).

Cell lines

The human COAD cell line (SW480), thyroid anaplastic carcinoma cell line (Cal62), cervical carcinoma cell line (HeLa), glioblastoma cell line (U87) renal carcinoma cell line (OS-RC-2) and T2 cell line were obtained from Procell Life Science & Technology Co., Ltd. SW480, Cal62, and HeLa cells were cultured in Dulbecco's modified eagle medium (DMEM) medium (Gibco), U87 cells in minimum essential media (MEM) medium supplemented withnon-essential amino acids (Gibco), and OS-RC-2 and T2 cells in RPMI-1640 medium (Gibco). All media were supplemented with 10% fetal bovine serum (Gibco), 100 U/mL penicillin (Gibco), and 100 μg/mL streptomycin (Gibco). Cells were maintained at 37°C in 5% CO_2.

METHOD DETAILS

Identification of young *de novo* genes in humans

De novo gene candidates were compiled from An et al.,[16] Broesil et al.,[26] and 12 additional studies reporting *de novo* genes in humans (Table S1). To ensure gene annotation consistency within an updated genomic context, we performed systematic sequence similarity searches for the coding sequences (CDS) of all candidates against the GENCODE comprehensive gene annotation set (Version 43) using BLASTn. This analysis identified annotated transcripts containing intact CDSs, which were subsequently processed by merging duplicate entries and manually verifying genomic loci, resulting in a final set of 100 candidate genes (Table S1).

We first assessed the origins of these candidates using orthologous genomic sequences derived from a 120-mammals whole-genome alignments.[52] High-quality orthologous sequences (alignment coverage >95%) were extracted using the maf_parse tool from PHAST[113] (v1.4), and ancestral genomic sequences were reconstructed using PRANK[114] (v170427) with the following parameters: -showanc -showevents -prunetree -prunedata -F -once, based on the pre-built phylogenetic tree of 120 mammals. For each candidate gene, we systematically evaluated the coding potential across all common ancestors in the phylogenetic tree by assessing both the structural integrity and length of intact ORFs. Specifically, if the ancestral sites aligned precisely or within an in-frame window of six nucleotides downstream of the translation initiation site of the candidate, and the reconstructed ancestral sequence encoded a putative ORF longer than 70% of the candidate, the orthologous region in that ancestor was considered coding.[7,8,26,48–50] Only candidates showing a complete absence of coding potential in all older ancestors prior to the first emergence of coding potential were retained, highlighting the non-coding to coding transition characteristic of *de novo* gene origination, as proposed in recent practices.[26,50,48,49]

Based on the syntenic genomic alignments across multiple species, we further determined the evolutionary age of ORFs encoded by these candidates. Specifically, if an outgroup species encoded a putative ORF longer than 70% of the candidate ORF and aligns to

it in the same frame without frameshifts, the ORF was considered present in that species. The age of these candidate genes was then defined by assessing the presence or absence of long, in-frame ORFs in outgroup species along the phylogenetic tree, and only candidates with young ORFs newly originated in the hominoid lineage were retained. To further confirm that the finding of these candidates represents recent gene origination events in humans rather than ancestral ORFs lost in outgroups, we required evidence of "common disablers." These are shared mutations—such as start codon losses, frameshifts or stop codon gains—present in orthologous sequences across multiple outgroup species. Only candidates with at least one common disablers during the evolution were retained. The candidates were further manually curated using the latest primate synteny data from UCSC Primate Genomes (*Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla*, *Pongo abelii*, *Nomascus leucogenys*, *Macaca mulatta*, *Callithrix jacchus*, and *Otolemur garnettii*; Data S2).

Second, to rule out gene duplication as the origin of these candidates, homolog searches were performed against the human genome (hg38) and annotated transcripts (GENCODE V43) using BLASTn with the criteria: E-value $<10^{-6}$, identity >50%, and query coverage >50%. Additional searches were conducted in representative outgroup species (e.g., *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla*, *Pongo abelii*, *Nomascus leucogenys*, *Macaca mulatta*, *Macaca fascicularis*, *Callithrix jacchus*, *Otolemur garnettii*, *Oryctolagus cuniculus*, *Mus musculus*, *Canis lupus familiaris*, *Loxodonta africana*, and *Monodelphis domestica*) against their annotated proteomes (Ensembl version 114) using BLASTp with the criteria: E-value $<10^{-5}$, identity >40%, and query coverage >50%. Candidates with multiple homologs mapping to distinct genomic locations, transcripts, or proteins were excluded.

Finally, to further eliminate potential distant homologs,[12] we performed additional BLASTp analyses and iterative jackhmmer[115] searches (HMMER v3.1b2) against the UniProtKB database[116] (2025_01 release, 253,206,170 entries, 1,333,558 species). Iterative jackhmmer searches were conducted using sequence and domain E-values of 10^{-5} as inclusion thresholds for subsequent iteration. Putative homologous hits were then identified using jackhmmer with an E-value cutoff of 10^{-3} or BLASTp with an E-value $<10^{-5}$ and sequence identity >40%. We carefully curated these entries individually to eliminate ambiguous cases, including those that appear as orthologs in newly sequenced species not included in the initial 120-mammal dataset; orthologs in species with low-quality genome assemblies; and paralogs in species outside the human and 14 representative out-group species. Finally, we identified 37 *de novo* genes that encode human- or hominoid-specific proteins (Table S2).

Evolutionary ages of canonical protein-coding genes were obtained from GenTree,[77] with genes predating the divergence of old-world monkeys and hominoids classified as "older genes". In addition, 257 ''older'' *de novo* genes were also identified from GenTree and evaluated using the same workflow applied for the 37 young *de novo* genes. A total of 1,456 housekeeping genes were compiled from three sources: Eisenberg et al.,[117] the Housekeeping and Reference Transcript (HRT) Atlas database (v1.0)[118] and Uhlén et al.[119]

Expression profiles of *de novo* genes

Expression levels of *de novo* genes were quantified by uniquely mapped reads using featureCounts[95] (v2.0.2) with default parameters, based on 6,908 RNA-seq datasets across 27 normal tissues (GTEx V8), 22 tumor types (TCGA), and 21 developmental stages in six organs (Cardoso-Moreira et al.[63]; Table S3). For strand-non-specific RNA-seq data, unique representative regions for each *de novo* transcript were identified using BEDTools[110] (v2.26.0), excluding overlaps with other genes or regions to reduce false positives caused by pervasive natural antisense transcripts in primates. The regions shorter than 100 bp were excluded. Genes were considered unexpressed if their maximum median TPM across tissues or tumor types was less than 0.2.

We performed hierarchical clustering (Ward algorithm) of developmental expression profiles for *de novo* genes, with gene similarity measured by Pearson correlation coefficients. For each organ, clusters containing fewer than two genes were iteratively merged with the most similar adjacent cluster. Expression profiles were grouped into clusters 1–6, representing varied patterns for predominant gene expression from late to early developmental stages (Figures S3 and S4). Genes with a maximum TPM below 0.5 across all time points in an organ were unclassified (cluster 0).

We quantified spatial expression specificity using the *Tau* metric, calculated from TPM-normalized expression values across multiple tissue types in GTEx datasets. For temporal specificity assessment within individual organs, we applied the same *Tau* metric to expression profiles across developmental stages using data from Cardoso-Moreira et al.[63]

$$Tau = \frac{\sum_{i=1}^{N}\left(1 - \frac{Exp_i + 0.01}{Exp_{max} + 0.01}\right)}{N - 1}$$

where Exp_i indicates the median TPM value of a gene in all samples of i-th tissue type (or i-th time point of an organ), and Exp_{max} indicates the maximun Exp_i among all tissue types (or time points of an organ). A pseudo-value of 0.01 was added to prevent division by zero and minimize low-expression noise. N represents the total number of tissue types (or time points of an organ) considered for the calculation.

Temporally dynamic genes were identified using the maSigPro package[96] (v1.70.0), where TPM values were regressed against log-transformed days post-conception using a third-degree polynomial model. Genes were classified as temporally dynamic if they showed an $R^2 > 0.3$ in regression analysis or a temporal specificity index >0.8. Early-expressed and late-expressed genes were further defined based on their predominant expression during early developmental stages (cluster 4–cluster 6) or late stages (cluster 1–cluster 3).

To quantify the expression levels of *de novo* genes in the five cancer cell lines (U87, SW480, HeLa, Cal62, and OS-RC-2), total RNA was extracted and subjected to library preparation using the VAHTS Universal V8 RNA-seq Library Prep Kit (Vazyme, China). The prepared libraries were subsequently sequenced on the DNBSEQ-T7 platform (MGI Tech, Shenzhen, China). Raw reads containing 3′ adaptor or low-quality bases (Phred score < Q20) were trimmed or removed using fastp[97] (v0.23.4). Reads were then aligned to the human genome (hg38, GENCODE V43) using STAR[101] (v2.7.9a) with default parameters. Uniquely mapped reads were counted using featureCounts[95] (v2.0.2), and TPM values were calculated to quantify gene expression levels across all generated and incorporated public data (Table S13). Genes were considered expressed if the maximum TPM across all samples in a cell line exceeded 0.2.

Translational evidence for *de novo* genes

Evidence supported by Ribo-seq data

We re-analyzed 279 Ribo-seq datasets from various human tissues, embryonic and neural stem cells, and cell lines (Table S4). Low-quality reads (Phred score < Q20) and reads containing 3′ adaptors were removed or trimmed using one of the recommended tools—fastp[97] (v0.23.4), Trimmomatic[98] (v0.39), or Cutadapt[99] (v4.5)—as specified in the original publications. Trimmed reads were mapped to human ribosomal RNA (rRNA) index libraries using Bowtie 2[100] (v.2.5.3) and rRNA-aligned reads were excluded. Clean reads were then aligned to the human genome (hg38, GENCODE V43 annotations) using STAR[101] (v2.7.9a) with the following parameters: –alignEndsType EndToEnd –seedSearchStartLmax 15 –outFilterMismatchNmax 2 –outSJfilterOverhangMin 30 8 8 8, retaining only uniquely mapped reads. Reads longer than 32 nt or shorter than 27 nt were discarded. To analyze ribosomal profiling periodicity in candidate ORFs, we first determined P-site positions using RiboTISH (v0.2.7)[102]. The P-site offset in each read length was calculated using canonical ORFs. For reads on the positive strand, P-site positions were determinated as the read start position + offset, while for reads on the negative strand, the positions were calculated as read start position + read length – offset – 1. The 3-nucleotide periodicity within ORFs encoded by canonical and *de novo* genes was then quantified by calculating the proportion of P-site reads in reading frame 0. Translation of *de novo* genes was then predicted using RiboTISH with default parameters and P-site offset file generated before (RiboPvalue <0.05; Table S5), based on the ribosome footprint patterns. This was supplemented with annotations from four databases (openProt, RPFdb v2, sORFs.org and nORFs.org), eight Ribo-seq studies, and a reference catalog for human translated ORFs (Tables S6 and S7). Together, these sources provide evidence for the translation of a subset of *de novo* genes from the perspective of Ribo-seq data.

Evidence supported by MS data

In-house MS data were analyzed using pFind (v3.2)[103] against a combined protein database, which included Swiss-Prot proteins (March 2024, 20,433 entries), *de novo* proteins, and built-in contaminants. The search parameters were set as follows: open modification search, peptide mass tolerances of 10 ppm, fragment mass tolerance of 20 ppm, up to two missed cleavages for fully tryptic peptides, and a 5% false discovery rate threshold for peptide-spectrum matches (PSMs). For publicly available MS data, PSMs were obtained from PeptideAtlas (Human 2024-1) and MassIVE (released in 11/30/2023), and synthetic spectra were excluded from subsequent analyses. Peptides completely mapped to *de novo* genes were identified through similarity searches. Finally, all above PSMs uniquely mapped to *de novo* genes were used to confirm their *in vivo* translation, and those mapped to multiple genomic locations or genes identified by BLAT and BLASTp were excluded. The peptide evidence is provided in Table S8 (in-house generated MS) and Table S9 (PeptideAtlas and MassIVE). Additionally, Western blot analyses supporting the translation of selected *de novo* genes were also included (Table S1).

Properties of *de novo* genes

RNA nuclear export activity was quantified as the ratio of RNA abundance in the nucleus-to-cytoplasm using RNA-seq data from Hennig et al.[120] Uniquely mapped reads were retained and counted, and weakly expressed transcripts ($TPM_{nucleus}$ + $TPM_{cytoplasm}$ < 0.8) were discarded. The nucleus-to-cytoplasm TPM ratio was log_2-transformed, with a pseudo-count of 0.1 added to each TPM value. Intrinsic structural disorder was assessed using AIUPred (v0.9)[104] with default parameters, excluding cysteines to account for potential disulfide bond uncertainties.[121] The disorder scores were then averaged across all remaining amino acids for each protein. C-terminal hydrophobicity of protein was analyzed for ORFs longer than 300 nt (100 amino acids), following the methodology described previously.[61] The average hydrophobicity of the last 30 amino acids near the C-terminal was calculated using the hydrophobicity function (Miyazawa scale) in the R package Peptides (v2.4.6)[106]. To specifically assess C-terminal hydrophobicity independent of known functional domains, which are often hydrophilic, we analyzed only proteins lacking annotated domains in their final 100 amino acids. Domain identification was performed using BLASTp searches against the NCBI Conserved Domain Database (E-value $<10^{-9}$). Translation efficiency was calculated using RNA-seq and Ribo-seq data of human brain, liver and testis samples from Wang et al.[85] Reads uniquely mapped to ORFs were counted, and translation efficiency was defined as the count ratio of total ribosome-bound to total RNA-bound reads in all samples. A pseudo-count of 1 was added respectively to prevent division by zero and minimize noise. As a note, we applied an expression threshold (mean TPM >0.2) to ensure reliable estimates, excluding low-abundance transcripts. For comparison, transcript and ORFs of other categories (Canonical, lncRNA and Intergenic) were shuffled from the GENCODE gene annotation set (Version 43) and extracted using the EMBOSS getorf (v6.5.7.0) with the ATG as the start codon. The properties of these groups were analyzed in parallel.

Differential expression analyses between healthy and tumor samples

For differential expression analysis, we employed matched GTEx samples as healthy controls to identify tumor-associated transcriptional changes across TCGA cancer types (Table S3), following established methodologies.[122] The integration of GTEx and TCGA datasets was validated by comparing the expression profiles of 1,456 housekeeping genes between GTEx normal tissues and TCGA normal adjacent tissues, showing strong correlation (Pearson correlation coefficient $r = 0.964$, p-value $<1 \times 10^{-15}$ for median expression; coefficient $r = 0.94$, p-value $<1 \times 10^{-15}$ for SD; Figures S5 and S6). Batch effects were controlled using the RUVg method (RUVSeq package[105] v.1.31.0), with one factor of unwanted variation (k = 1) and 1,456 housekeeping genes serving as the negative control set. Normalizations were performed separately for paired healthy and tumor samples within each tissue type (Table S3). Finally, upregulated genes in each tumor type were identified by comparing TCGA tumor transcriptomes with GTEx normal tissue transcriptomes under the following criteria: 1) Bonferroni-Hochberg corrected p-value <0.05 (Wilcoxon rank-sum test), and 2) fold change in upper quantile TPM >1.5. Upregulated genes which expressed in tumors but not in the corresponding normal tissue samples (median TPM <0.5) were classified as exhibiting tumor-specific expression expansion. Moreover, a total of 363 oncogenes were retrieved from the OncoKB knowledgebase (released in 10/29/2024).[123]

Monte Carlo simulation

We employed the Monte Carlo simulation to assess the statistical significance of expression expansion and ecDNA preference for *de novo* genes, comparing them to whole-genome background genes and known oncogenes. Specifically, we generated 10,000 distinct randomized datasets, each consisting of 37 background genes or oncogenes. To ensure a fair comparison for expression expansion, background genes were selected based on a probability distribution constrained by the low expression levels characteristic of *de novo* genes. Finally, statistical significance for *de novo* genes was determined by calculating the fraction of simulated datasets in which the examined metrics (e.g., proportion of genes upregulated or involved in ecDNAs in tumors) exceeded those observed for *de novo* genes.

Genome architecture and ecDNA

Raw Hi-C sequencing data, 11 from normal tissues (as controls) and 47 from tumors (Table S10), were processed using the runHiC pipeline (v0.8.7). Normalization was performed with JuicerTools addNorm (v1.13.02) using VC_SQRT, and A/B compartments were identified using the FAN-C[108] compartments tool (v0.9.26). Genes undergoing consistent compartment transitions in at least two samples *per* tumor type were retained. Tumor-specific loops, representing enhancer–promoter interactions, were identified with Peakachu (v1.2.0)[109] with depth-appropriate models at 10 kb resolution (Peakachu score >0.95 in tumor samples, with average scores <0.55 in 20 kb flanking regions in control samples) and verified using aggregate peak analysis with HiCPeaks (v0.3.7) (Table S11; Figure S8). Only loops connect active enhancers annotated in the FANTOM5 project[88,124] and reproducibly detected (present in $\geq$2 samples *per* tumor type) were included for analysis of genome architecture alterations in tumorigenesis.

To identify enhancer–promoter interactions with increased enhancer activity in tumors, we utilized 60,215 active enhancers from the FANTOM5 project.[124] To precisely measure enhancer activity through transcription,[125] enhancers overlapping known transcripts (GENCODE v43) or intronic regions were excluded, leaving 19,457 enhancers. RNA-seq reads from normal tissue (GTEx) and tumor samples (TCGA) overlapping these enhancers and their 2 kb flanking regions were quantified. Enhancer activity levels were reported as reads *per* million mapped reads (RPM) and batch effects were minimized using the RUVg method, as described in the previous section. Enhancers with higher activity in tumors compared to normal samples were defined by a fold change in median RPM >1 and a Bonferroni-Hochberg adjusted P-value <0.05 (Wilcoxon rank-sum test).

Annotated ecDNAs were retrieved from the eccDNA Atlas v1.05[71]. *De novo* genes located on at least two independent ecDNAs within each tumor type were identified. To validate these ecDNAs and trace their sequences in patients, whole-genome sequencing data from 114 BLCA samples were downloaded from TCGA and processed using the AmpliconSuite pipeline (v1.3.4)[107] to detect ecDNAs (copy-number variants regions >50 kb, copy number >5). Data with paired-read rates below 95% were excluded to ensure reliable use of discordant read pairs. Circular ecDNAs were identified as single amplicons containing ecDNA or resulting from breakage-fusion-bridge events. Finally, the identified ecDNAs were visualized using CycleViz (v0.2.0).

sgRNA library cloning

The sgRNA library comprised 352 sgRNAs targeting young human *de novo* genes as described by An et al.,[16] 14 sgRNAs targeting 3 oncogenes genes, 2 sgRNAs targeting one tumor suppressor gene and 20 non-targeting control sgRNAs with no matches in the screened cell lines. Candidate sgRNA sequences were selected from the 150 bp upstream of the ORF to the 19 bp immediately preceding the NGG protospacer adjacent motif. The CRISPR-DO tool was employed to optimize sgRNA specificity and cleavage efficiency.[126] Sequences of sgRNAs targeting 27 *de novo* genes are provided in Table S12.

The sgRNA library was synthesized by Azenda Life Science Co., Ltd. Synthesized sgRNAs were amplified, purified, and cloned into the LentiCRISPR-V2 vector using the BsmBI restriction enzyme digestion followed by NEBuilder HiFi DNA assembly cloning. The pooled library was transformed into an electrocompetent strain (Stbl4) to ensure at least 300$\times$ library coverage. Transformed colonies were expanded in LB medium, and high-quality plasmid DNA was extracted using the genElute HP plasmid maxiprep kit (Tiangen). The library was further validated by next-generation sequencing to confirm uniform sgRNA distribution.

CRISPR-Cas9 screening assay

Lentiviral particles were generated by polyethylenimine (PEI)-mediated transfection of the sgRNA library into HEK293T cells. The multiplicity of infection (MOI) was optimized for each cell line (U87, SW480, HeLa, Cal62, and OS-RC-2) through pilot infections with obtained lentiviruses. Based on titration results, cells were infected at a standard MOI of 0.3 to ensure single-copy viral integration events. At 24 h post-infection, transduced cells were selected with 1 μg/mL puromycin (Gibco) for 72 h, followed by cultivation for ten passages (P0–P10) in complete medium refreshed every 48 h.

Genomic DNA was isolated at passages P0 and P10. Cell pellets (1×10^6 cells) were resuspended in 500 μL lysis buffer (200 mM NaCl, 100 mM Tris-HCl, pH 8.5, 50 mM EDTA, pH 8.0, 0.5% SDS) and incubated at 55°C for 16 h. The lysate was then treated sequentially with 40 μL RNase A (10 mg/mL) at 37°C for 2 h and 20 μL proteinase K (20 mg/mL) at 55°C for 2 h. DNA purification was performed through organic extraction (25:24:1 phenol:chloroform:isoamyl alcohol), sodium acetate precipitation, and 70% ethanol washes. Purified DNA was dissolved in nuclease-free water and quantified using a Nanodrop spectrophotometer. SgRNA inserts were amplified by PCR[127], with products purified using AMPure XP beads (Beckman Coulter) and quantified *via* Qubit dsDNA HS Assay (Thermo Fisher Scientific). Final libraries were sequenced (150 bp paired-end) on an Illumina NovaSeq X Plus platform.

Downstream analysis identified 335 sgRNAs targeting 34 *de novo* genes. Raw sequencing reads were preprocessed to remove low-quality sequences and adapter contaminants. Using MAGeCK[112] (v0.5.9.5) with parameters -norm-method control –gene-lfc-method secondbest, we calculated fold changes in sgRNA abundance between P0 and P10 and the associated statistical significance. Significant hits of changed sgRNA abundance were defined by dual thresholds: p-value <0.05 and fold change >0.7.

Validations with siRNA knockdown and *ELFN1-AS1* knockout

The siRNAs targeting 14 candidate genes (*ODC1-DT*, *AATBC*, *TIPARP-AS1*, *MOCS2-DT*, *TENM3-AS1*, *ELFN1-AS1*, *MYEOV*, *SMIM45-107aa*, *PAX8-AS1*, *RNF32-DT*, *ADORA2A-AS1*, *SERP1*, *EXOC3-AS1*, *YIF1B*) and a positive control (*MYC*) were synthesized by Berry Genomics Co., Ltd. Three independent siRNAs were designed for each gene, with siRNAs targeting *MYC* and non-targeting scrambled siRNAs serving as the positive and negative controls, respectively. The siRNA sequences are provided in Table S14. To knockout *ELFN1-AS1*, an sgRNA was designed to target the coding sequence of *ELFN1-AS1* (5′-TGCGTCTCGGAGT GAATGAC-3′) and inserted into PX458 vector *via* BbsI restriction sites.

For transient siRNA transfection, U87, SW480, HeLa, Cal62, and OS-RC-2 cells were cultured to 60% confluency and transfected with either control siRNA or siRNA targeting *de novo* genes using lipofectamine RNAiMAX (Invitrogen). To ensure efficient knockdown, three siRNAs designed against each gene were pooled and transfected together. 24 h post-transfection, U87, SW480, HeLa, Cal62, and OS-RC-2 cells were seeded into 96-well plates at a density of 2×10^3 cells *per* well. CCK-8 reagent (Solarbio) was then added to the plates, followed by a 2-h incubation at 37°C. This procedure was repeated for five consecutive days. Cell proliferation was assessed by measuring absorbance at 450 nm, and the resulting data were used to generate the growth curve.

For the validation experiment with *ELFN1-AS1* knockout in SW480 cells, sgRNA targeting *ELFN1-AS1* and non-targeting scrambled sgRNA were transfected into SW480 cells using lipofectamine 2000 (Invitrogen), respectively. After 24 h, GFP-positive cells were sorted and seeded into 96-well plates. Sorted cells were cultured in complete medium for 14 days to establish monoclonal populations. Complete gene knockout in clonal cell lines was confirmed by Sanger sequencing of target PCR products from genomic DNA. The clone exhibiting frameshift mutations was selected for further validation, with a non-targeting sgRNA-infected clone serving as negative control. Both selected frameshift mutant clones and negative controls were plated in 96-well plates at 2×10^3 cells *per* well. Cell proliferation was assessed using the identical CCK-8 assay protocol described previously.

Prognosis analysis

Patients were stratified into high- and low-expression groups based on the median TPM values of *de novo* genes for each tumor type. Cox proportional hazards models were used to evaluate statistical significance (survival, v3.5.8). Survival curves were generated using the survminer package (v0.4.9).

mRNA vaccine preparation

ELFN1-AS1-3×FLAG and TYMSOS-3×FLAG mRNAs were synthesized using an *in vitro* transcription kit (Vazyme). PCR-amplified sequences were cloned into a pVAX1 vector, with optimized 5′ and 3′ UTR and a poly(A) tail. IVT reactions were performed according to the manufacturer's protocol, replacing uridine triphosphate with N1-methylpseudouridine-5′-triphosphate and adding Cap-1 for mRNA capping. Synthesized mRNAs were stored at −80°C until use. Sequence details are provided in Table S15.

mRNA–LNP formulations were prepared using a microfluidic device. Lipids were dissolved in ethanol, and mRNA was dissolved in 10 mM citrate buffer (pH 4.0). The lipid mixtures were prepared at a molar ratio of 50/10/38.5/1.5 (SM102/DSPC/cholesterol/DMG-PEG2K) and mixed with the mRNA solution at a 3:1 volume ratio in the microfluidic system. The final lipid-to-mRNA weight ratio was 20:1. Control formulations replaced the mRNA solution with citrate buffer. The LNP solution was dialyzed against 1×PBS for two hours using Pur-A-Lyzer Midi Dialysis Kits (WMCO 3.5 kDa) and concentrated by ultra-filtration (10 kD, Millipore) to a lipid concentration of 12 μg/μL (mRNA concentration: 0.6 μg/μL).

Mice vaccination for immune response evaluation and tumor challenge

Mice were randomly assigned to different groups before treatment. For immune response evaluation, female huHSC-C-NKG-ProF mice were administered intramuscular injections of LNP–control, LNP–ELFN1-AS1 or LNP–TYMSOS. Four doses of 30 μg *per* mouse were administered weekly. One week after the last final vaccine dose, PBMCs and spleen mononuclear cells (SMCs) of mice were collected for subsequent flow cytometry analysis and *ex vivo* IFN-γ ELISpot assay. PBMCs and SMCs were isolated using lymphocyte separation medium (Solarbio) and mouse spleen lymphocyte separation medium (Solarbio), respectively, following the protocol provided by manufacturer.

For tumor challenge and therapeutic treatment, female huHSC-C-NKG-ProF mice were administered intramuscular injections of LNP–control, LNP–ELFN1-AS1 or LNP–TYMSOS, with receiving a total of four doses of 30 μg *per* mouse weekly. One week after following the last dose of vaccination, SW480 tumor cells (5×10^6) were implanted subcutaneously into the right flank of the humanized mice. Tumor growth was monitored by measuring the tumor dimensions, including length (a) and width (b), were assessed to determine the tumor volume, which was measured using the formula: Tumor Volume = $ab^2/2$.

For the combination therapy with PD-1–laIL-2, female huHSC-C-NKG-ProF mice received the same vaccination regimen and tumor challenge as described above (four weekly intramuscular doses of 30 μg *per* dose *per* mouse of LNP–control, LNP–ELFN1-AS1, or LNP–TYMSOS). Tumor-bearing mice subsequently received intraperitoneal administration of PD-1–laIL-2 (20 μg/mouse) at two timepoints: days 12 and 14 post-tumor inoculation, followed by tumor size measurements.

Flow cytometry

Spleens were harvested from mice and filtered through a 70 μm cell strainer. PBMCs and SMCs were isolated using mouse spleen lymphocyte separation medium (Solarbio) and washed three times with PBS. DC maturation and activation were assessed by staining cells with Fixable Viability Dye eFluor 506 (Biolegend), PE anti-human CD11c (Biolegend), Brilliant Violet 421 anti-human CD80 (Biolegend), APC anti-human CD86 (Biolegend), and FITC anti-human CD83 (Biolegend) antibodies. CD137-positive T cells were identified using Fixable Viability Dye eFluor 506, FITC anti-human CD45 (Biolegend), Brilliant Violet 650 anti-human CD3 (Biolegend), PE-Cy7 anti-human CD8a (Biolegend), and Brilliant Violet 421 anti-human CD137 (4-1BB) (Biolegend) antibodies. Flow cytometry was performed on a CYTEK Aurora system.

Epitope peptides prediction

The HLA binding affinity of *ELFN1-AS1* and *TYMSOS* peptides was predicted using the IEDB-recommended predictor, NetMHCpan 4.1 BA, from the IEDB MHC-I binding predictions (v2.24) using 9–10mers for all MHC-I A/B allele reference set. Peptides with an IC_{50} <500 nM or a %Rank <2.0 were classified as HLA binders, while those with an IC_{50} <50 nM or a %Rank <0.5 were identified as strong binders. The optimal peptides were selected for synthesis (Table S16).

T2 cell binding assay

High-affinity peptides were synthesized by GeneScript and purified to ≥90% purity, with endotoxin removal and trifluoroacetic acid content verified. The lyophilized peptides were dissolved in DMSO to a concentration of 5 mg/mL. Purity and identity were confirmed by high-performance liquid chromatography, and the peptides were stored at −80°C until used in T2 cell binding assay and T cell response assays.

The affinity of peptides to HLA-A*02:01 molecules was assessed using T2 cell binding assay. T2 cells were resuspended in serum-free RPMI-1640 medium and seeded into U-bottom 96-well plates at a density of 1×10^5 cells *per* well. Subsequently, 50 μg/mL of the epitope peptides and 5 μg/mL of human β2-microglobulin (Sigma) were added to the culture medium, and the cells were incubated at 37°C in a 5% CO_2 atmosphere for 16 h. After incubation, the cells were stained with PE HLA-A2 monoclonal antibody (Ebioscience) for 20 min at 4°C and were washed three times with PBS subsequently. Mean fluorescence intensity (MFI) was then determined by flow cytometry. The validated peptide (CDC73-NIFAILESV) was used as a positive control and the peptide (KRAS(A11)-G12C-VGACGVGK) was used as a negative control. The unstimulated T2 cells treated with DMSO alone were used as the background control. Peptide binding affinity was expressed as the fluorescence index (FI), calculated using the following formula: FI = (MFI with peptide − MFI without peptide)/MFI without peptide. Peptides with an FI >1 were considered high-affinity epitopes.

Ex vivo IFN-γ ELISpot assay

Peptide-reactive IFN-γ-secreting T cells were identified using a human IFN-γ ELISpot kit (Mabtech). Plates were washed five times with PBS and blocked with ELISpot serum-free medium (Dakewei) for 0.5 h. A total of 1×10^5 PBMCs *per* well were stimulated for 20 h with 25 μg/mL of selected 9–10 amino-acid-length peptides in ELISpot serum-free medium. Each assay was performed in duplicate, with DMSO as the negative control and anti-CD3 (Mabtech) as the positive control. IFN-γ binding was detected using a biotin-conjugated antibody, followed by incubation with a streptavidin-HRP secondary antibody and visualization with TMB substrate. Spots were quantified (Dakewei, China) and normalized as spot-forming units *per* 1×10^5 PBMCs, firstly divided by the positive control, and then by the negative control.

For IFN-γ ELISpot assay in PBMCs from patients, a total of six colorectal cancer patients were recruited. PBMCs were isolated from fresh whole blood, and 5×10^4 PBMCs *per* well were stimulated for 20 h with 25 μg/mL pools of 9–10 amino acid-length peptides

in ELISpot serum-free medium, covering most HLA-A subtypes. Each assay was performed in triplicate. Spots were quantified and normalized as described above. The patient information and the peptide sequences included in each peptide pool are provided in the Table S17.

QUANTIFICATION AND STATISTICAL ANALYSIS

All quantification and statistical details are indicated in the method details, results, or figure legends.