Single-cell RNA-seq-based proteogenomics identifies glioblastoma-specific transposable elements encoding HLA-I-presented peptides

Pierre-Emmanuel Bonté, Yago Arribas, Antonela Merlotti, Montserrat Carrascal, Jiasi Vicky Zhang, Elina Zueva, Zev Binder, Cécile Alanio, Christel Goudot, Sebastian Amigorena

Version of Record: https://www.sciencedirect.com/science/article/pii/S2211124722006933
Manuscript_df4de08368fed45bab5b27cb63889f27

Single cell RNAseq-based proteogenomics identifies glioblastoma-specific transposable elements encoding HLA-I-presented peptides

Pierre-Emmanuel Bonté[1,*], Yago A. Arribas[1,*,] Antonela Merlotti[1], Montserrat Carrascal[2], Jiasi Vicky Zhang[3], Elina Zueva[1], Zev A. Binder[3], Cécile Alanio[1,4,5], Christel Goudot[1,#,§], Sebastian Amigorena[1,6,#,§]

[1] Institut Curie, PSL University, Inserm U932, Immunity and Cancer, 75005 Paris, France

[2] Biological and Environmental Proteomics, Institut d'Investigacions Biomèdiques de Barcelona-CSIC, IDIBAPS, Roselló 161, 6a planta, 08036 Barcelona, Spain.

[3] GBM Translational Center of Excellence, Department of Neurosurgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA.

[4] Laboratoire d'immunologie clinique, Institut Curie, Paris, 75005, France

[5] Parker Institute of Cancer Immunotherapy, San Francisco, USA

[6] Lead contact

[*] These authors contributed equally.

[#] These authors contributed equally.

[§] Correspondance: christel.goudot@curie.fr, sebastian.amigorena@curie.fr

Summary

We analyze transposable elements (TE) in glioblastoma (GBM) patients using a proteogenomic pipeline that combines single cell transcriptomics, bulk RNAseq samples from tumors and healthy tissue cohorts, and immunopeptidomic samples. We thus identify 370 HLA-I-bound peptides encoded by TEs differentially expressed in GBM. Some of the peptides are encoded by repeat sequences from intact open reading frames (ORFs) present in up to several hundred TEs from recent LINE-1, LTR and SVA subfamilies. Other HLA-I-bound peptides are encoded by single copies of TEs from old subfamilies that are expressed recurrently in GBM tumors and not expressed, or very infrequently and at low levels, in healthy tissues (including brain). These peptide-coding, GBM-specific, highly recurrent TEs represent potential tumor-specific targets for cancer immunotherapies.

Introduction

T cells can control, and sometimes reject, solid tumors, especially after reprograming by immune checkpoint blockade (ICB) (Morotti et al., 2021; Waldman et al., 2020). The nature of the tumor antigens targeted by these T cells, however, remains unclear. After identification of differentiation and tumor-testis antigens decades ago (Almeida et al., 2009; Boon and van der Bruggen, 1996; Simpson et al., 2005; van der Bruggen et al., 1991), a new family of antigens derived from tumor somatic mutations was discovered (Coulie et al., 1995; Robbins et al., 1996; Tran et al., 2015; van Rooij et al., 2013). Defined sets of mutations in single cells, occurring before or after oncogenic transformation, are amplified by clonal expansion of tumor cells (Castle et al., 2012). This "amplified" set of mutations, becomes "visible" to the immune system, and triggers T cell immune responses (Lennerz et al., 2005; Robbins et al., 2013; van Rooij et al., 2013). Unlike differentiation and tumor testis antigens that are, by definition, also expressed in certain normal cells, mutational neo-antigens are strictly tumor specific. However, most such mutations are passenger events and are largely specific to individual patients. The presence of mutation-specific T cells in ICB-treated cancer patients, the high rate of clinical responses to ICB in patients with microsatellite instability, and the correlation between the median number of mutations in certain cancer types and the rate of response to ICB, all indicate that passenger mutations can be effectively targeted by T cells in cancer patients (Carreno et al., 2015; Chauvin et al., 2015; Gubin et al., 2014; Le et al., 2015; Rizvi et al., 2015; Schadendorf et al., 2015; Snyder et al., 2014).

Several lines of evidence, however, also suggest that point mutations are not the only antigens seen by T cells on human tumors. First, there are exceptions to the correlation between the frequency of mutations and the rates of response to ICB (McGrail et al., 2021). Renal cell carcinoma (RCC), for example, has a mutational burden around 2 mutations per megabase (MB), and a response rate to ICB around 25%, as compared to squamous non-small cell lung cancer (LUSC), with around 9 mutations/MB and a response rate to ICB of 17% (Yarchoan et al., 2019; Yarchoan et al., 2017). Second, at the level of individual patients, the number of mutations is not highly predictive of clinical responses to ICB (Gromeier et al., 2021; McGrail et al., 2021). Finally, there are multiple examples in the literature of T cell responses to non-mutational antigens in cancer patients, including differentiation and tumor-testis antigens (Novellino et al., 2005; Rapoport et al., 2015; Rooney et al., 2015).

Different teams have recently used proteogenomic approaches to search broadly for tumor-specific non-canonical open reading frames (ORFs) that encode peptides presented by HLA-I molecules on tumor cells (Chong et al., 2020; Laumont et al., 2016). Most of the identified peptides in these studies derive from non-coding genomic regions. Some of these potential tumor-specific antigens are found in multiple patients and can induce immune responses *in vitro* or in mouse models (Ehx et al., 2021); Laumont et al. (2018). A large fraction of the non-coding genome is composed of transposable elements (TEs). TEs include DNA transposons (Bourque et al., 2018; Burns, 2017) as well as three main classes of retrotransposons (short interspersed nuclear elements -SINE, long interspersed nuclear

elements -LINE and long terminal repeats -LTRs). Each class is sub-divided into families and subfamilies that arose during evolution from common ancestors and are classified according to sequence homology of the individual copies in the genome. The age of TE subfamilies can be estimated based on the conservation of their repeat motifs (Choudhary et al., 2020). A small proportion of young LINE-1 in the human genome can still be active for retro-transposition (Cowley and Oakey, 2013; Lanciano and Cristofari, 2020; Mills et al., 2007; Zhang et al., 2020). Retro-transposition can compromise the stability of the genome, and mammalian differentiated cells in tissues protect themselves against TE-induced genome instability through epigenetic repression of TE transcription (Burns, 2017) (Slotkin and Martienssen, 2007). As a result, TE transcription is low in most adult cells, and more active during embryonic development, in stem cells and, intriguingly, in tumors (Garcia-Perez et al., 2016). TE de-repression in tumors occurs through multiple epigenetic changes to TE loci, including DNA and histone de-methylation (Anwar et al., 2017; Grundy et al., 2021; Lynch-Sutherland et al., 2020). Both epigenetic changes can be associated with oncogenesis, resulting in different levels of epigenetic de-regulation.

TE overexpression in tumors compared to healthy tissue has prompted multiple teams to search for anti-TE T cell responses in cancer, and there is clear evidence that this can occur (Neukirch et al., 2019; Rycaj et al., 2015; Saini et al., 2020; Wang-Johanning et al., 2008). One recent study showed presentation of TE-derived peptides on HLA-I molecules (Kong et al., 2019). This study, however, only analyzed peptides derived from TE-subfamilies and did not address the cellular origin of the identified HLA-I-presented peptides. Whether TEs de-repressed in tumors can be a source of truly tumor-specific antigens is therefore still an open question. Here, we propose an original TE-centered proteogenomic approach based on a combination of single cell transcriptomics, bulk RNAseq analyses in tumor and healthy tissues, together with immunopeptidomics, to identify single and recurrent, tumor-selective TE-derived peptides presented by HLA-I molecules on GBM tumors.

Results

Single cell TE-expression resolves all cell populations in tumors

We reasoned that a powerful way to identify TEs expressed specifically in tumor cells would be to compare TE expression in tumor and in tumor-infiltrating cells from the same patient. To do so, we used single cell transcriptomics (scRNAseq) of all cells present in the tumor microenvironment. We initiated the study on a public data set including tumor and juxta-tumor samples from four GBM patients analyzed by SMARTseq2 (Figure S1A). Consistent with the analysis performed in the original article (Darmanis et al., 2017), dimensionality reduction and t-SNE visualization based on gene expression resolves the seven sorted cell populations from the tumor core and the surrounding tissue (tumor and periphery in Figure S1A): immune cells (mostly macrophages), neoplastic cells and oligodendrocyte precursor cells (OPCs) are the most numerous (Figure 1A, left panel and Figure S1B).

To investigate TE expression in single cells, we mapped scRNAseq reads to either TE subfamilies (as shown previously (Kong et al., 2019)) or to individual genomic TEs (Figure S1C). Because mapping of TEs to individual genomic locations can be affected by high conservation of their repeat motifs, we compared the use of uniquely and multi-mapping reads. Uniquely mapping reads allow accurate estimation of the expression of oldest TE subfamilies, but underestimates the expression for youngest TE subfamilies, as compared to multi-mapping reads, which reflect more accurately expression of young TE subfamilies (Figure S1D) (Lanciano and Cristofari, 2020).

tSNE based on expression on all 992 TE subfamilies, or 5000 most variable individual TEs in single cells, like gene expression, resolves all cell populations in the tumor microenvironment (Figure 1A, middle panel). Neoplastic cells and OPCs are mostly present in tumor and juxta-tumor (Figure S1A right panel) samples, respectively, while immune cells are present in both. Individually mapped TEs allow better resolution of the different cell populations than TE subfamilies (Figure 1A, right panel). These results show that expression of individual TEs can be resolved at the single cell level and is sufficient to distinguish different cell populations in the tumor microenvironment.

TE subfamilies are differentially expressed in neoplastic and immune cells

To better understand the nature of these TEs, we performed differential expression (DE) analyses of TEs in each cell population against all others, thus defining population-specific TE signatures (Figure S1E). These signatures are selective for neoplastic cells, immune cells (Figure 1B), and for each of the other cell populations present in the tumor microenvironment (Figure S1F). Heatmap representation of the 20 most differentially expressed TEs based on the average log2 fold change shows selective expression in each cell population, including in neoplastic cells (Figure 1C). To further investigate the nature of the TEs differentially expressed in each cell population, we compared each signature to all TEs expressed in the data set (130,028). TEs differentially expressed in neoplastic cells are depleted in SINEs (51.7% vs. 44.5%) and enriched in LTRs (8.3% vs. 12.1%), while TEs in immune cells are

depleted in LINEs (30.3% vs. 26.5%) and LTRs (8.3% vs. 5.6%) and enriched in SINEs (51.7% vs. 59.2%) (Figure S2A), confirming the results from direct mapping of TE subfamilies (Figure S2B). Statistical analyses by subfamily show strong enrichment for several LTR subfamilies in neoplastic cells (mainly HERV), while immune cells differentially express several SINE subfamilies (mainly Alu) (Figure 1D). We conclude that the different cell types present in the tumor environment express distinct patterns of TE subfamilies that can be analyzed from individually mapped TEs by single cell transcriptomics.

Gain of chromosome 7 and loss of chromosome 10 are recurrent genomic copy number alterations in GBM (Kurscheid et al., 2015). As an internal control for TE mapping to chromosomal loci, we quantified genes and TEs in each cell type-specific signature to their respective chromosomes. As shown in Figure 1E, TEs differentially expressed in neoplastic cells, but not in other cell populations, present a clear bias for chromosome 7 (Figure 1E, Figure S2C, Figure S2D). The bias for chromosome 7 in neoplastic cells is even stronger for TEs than for genes while the loss of chromosome 10 is similar in the TE and gene signatures (Figure S2C). A chromosome 7 bias is also observed when considering only the expression of distal TEs i.e., TEs located at more than 2Kb from the nearest protein-coding genes (mostly intergenic), indicating that this bias is not due to high contamination with intron retained TEs in the scRNAseq data sets (Figure S2E). We conclude that individual TEs can be accurately mapped from scRNAseq and as expected, show a chromosome 7 bias selectively in neoplastic GBM cells.

TE expression in neoplastic cells is enriched in elements independent of their closest gene

To better understand the control of TE expression in different cell populations, we first analyzed TE genomic locations. As compared to all expressed TEs in the data set, TEs differentially expressed in neoplastic cells show reduced intronic locations (77% vs. 38.7%), including when compared to the proportion of intronic TEs differentially expressed in immune cells (68.8%) or astrocytes (71%) (Figure S2F). Neoplastic TEs also show a marked increase in 3'UTR encoded TEs (25.3%), compared to all expressed TEs (5%) or to immune cell TEs (11.3%) (Figure S2F. These results show that, while TEs differentially expressed in immune cells are largely intronic, TEs differentially expressed in neoplastic cells are more frequently from intergenic and 3'UTRs regions.

Consistent with these results, the proportion of distal TEs is higher in the neoplastic cell signature (22.3%) that in the immune cell signature (13%, Figure 1F). t-SNE analysis based on distal TEs resolves all cell populations (Figure 1G), suggesting that cell type-specific TE expression may not be exclusively due to gene-driven transcription. Consistently, the TE-gene distances are increased for TEs differentially expressed in neoplastic cells, especially for LINE and LTRs (Figure S2G, as compared to the TEs differentially expressed in immune cells. Higher distances from the closest genes for TEs expressed selectively in neoplastic cells could reflect gene-independent TE expression, including enhancer-dependent or long non-coding (Lnc) RNA-dependent read-through transcription. We therefore next analyzed the

correlation between expression of TEs and their closest genes, in neoplastic and immune cells. Figure S2H shows examples of proximal and distal TEs, expressed together or independently of their closest gene. Quantification of the proportions of TEs in the two categories shows that the proportion of both proximal and distal TEs that are expressed while their closest gene is silent ($TE^{+}gene^{-}$), is higher in the neoplastic cells (39%) signature as compared to the immune cells (24%) signature (Figure 1H). These results show that higher proportions of TEs differentially expressed in neoplastic cells are distant and transcribed independently of their closest gene neighbor, suggesting a higher level of autonomy in TE transcription in GBM tumoral cells.

Tumor-enrichment and patient recurrence of the single cell neoplastic TE-signature

To validate the single cell-based neoplastic TE-signature, we next analyzed bulk RNAseq from the TCGA (155 GBM patients) and GTEx (1080 healthy samples from 25 tissues, Figure S3A and S3B) cohorts. As previously observed within the single cell RNA-seq data, the proportion of intronic TEs is higher in normal tissue than in GBM: 53.7% vs 68.6% (Figure S2F). These results indicate that neoplastic GBM cells express higher proportions of non-intronic TEs than non-neoplastic cells, and that this difference is detected in both bulk and scRNAseq data sets.

We next performed Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) based on neoplastic TE-signature and we show that GBM samples cluster away from normal tissue GTEx samples (Figure 2A and 2B, Figure S3C and S3D). Heatmap Z-score representation in TCGA and GTEx samples show higher expression of the 2000 top TEs from the single cell neoplastic TE-signature in TCGA GBM samples, and reduced expression in healthy tissues (Figure 2C). Gene Set Enrichment Analysis (GSEA) shows that expression of the neoplastic TE-signature is highly enriched in GBM vs. normal brain samples (NES=1.67 and FDR < 0.05, Figure 2D) and vs. other normal tissues samples in GTEx (Figure S3E and S3F). The median neoplastic TE-signature expression level is also higher in GBM samples, compared to normal tissue GTEx samples (Figure 2E and Figure S3G). Examples of individual TEs overexpressed in both data sets, bulk RNAseq (Figure 2F, top panels) and scRNAseq (bottom panels) illustrates the selective expression of certain TEs in GBM cells as compared to EGFR, a known GBM marker. We conclude that analysis of individual TEs from scRNAseq is accurate and allows the identification of recurrent, tumor-enriched individual TEs.

TE-derived peptides are presented on HLA-I and are immunogenic *in vitro*

To investigate if TE-derived peptides are presented by HLA-I molecules in GBM cells, we used 30 mass spectrometry (MS)-based immunopeptidomic samples from GBM primary tumors and cell lines (Forlani et al., 2021; Sarkizova et al., 2020; Shraibman et al., 2018; Shraibman et al., 2016) (Figure 3A). Multi-mapping (3428) or uniquely mapping (1945) differentially expressed TEs from the neoplastic TE-signature were *in silico* translated in the 6 reading frames (RF) and concatenated to the human annotated proteome. We thus obtained

370 TE-derived peptides, including 63 peptides identified in both signatures, 147 only in the multi-mapped reads signature and 160 only in the uniquely-mapped reads signature (Figure 3B-3C and Supplementary Data S1). Heatmap representation of all identified TE-derived peptides shows that the number of peptides varies among samples, and that some peptides are found in several patients and cell lines (Figure 3D).

TE-derived peptides showed similar SEQUEST quality scores and peptide length distribution as Uniprot-annotated peptides (Figure 3E and Figure S4A). TE-derived peptides binding to HLA-A3 (the most abundant HLA-I among all TE-derived peptides, n=96) contained the expected binding motif obtained from the Immune Epitope Database (IEDB) (Figure 3F) (Vita et al., 2019). In addition, TE-derived peptides maintained the correlation between hydrophobicity and retention time (3 representative examples in Figure S4B). These results indicate that TE-derived peptidome is reliable and contains similar characteristics to the canonical peptidome. In addition, 23 TE-derived peptides were synthesized and validated by comparison with the endogenous MS/MS spectra (out of 24 tested, Supplementary Data S1-S2). Confirming the robustness of our pipeline, the identified peptides, similar to the neoplastic TE-signature, are preferentially encoded by TEs from chromosome 7 and depleted from TEs on chromosome 10 (Figure 3G). We conclude that HLA-I molecules on GBM neoplastic cells present peptides encoded by differentially expressed TEs.

To investigate the possibility that TE-encoded peptides can represent potential tumor antigens, we searched for T cell precursors in healthy donors. Using a tetramer-formation assay, we first experimentally tested the binding for HLA-A*02:01 (6 peptides from immunopeptidomics and 17 from NetMHC predictions on *in silico* translated TEs from the neoplastic signature, Figure S4C) and for HLA-B*07:02 (2 peptides from the immunopeptidomics) (Figure S4D and Supplementary Data S1). 19 peptides were confirmed as HLA-I binders and were used to test immunogenicity *in vitro*. Peptide-loaded monocyte-derived dendritic cells were cultured with autologous $CD4^+$ and $CD8^+$ T cells from 7 healthy donors and tetramer staining was used as read-out (Figure 3H and Figure S4E). Figure 3H shows examples of expanded populations of TE-specific, tetramer-positive, $CD8^+$ T cells. Mutated Melan-A peptide, a strong binder to HLA-A*02:01 and with high T cell precursor frequency in most healthy donors (Pittet et al., 1999), was used as positive control for cell expansions. Three HLA-A*02:01-binding peptides from canonical proteins not specifically expressed in GBM were also included as negative controls. The 3 peptides derived from canonical proteins induced very weak or no responses, although mutated Melan-A derived peptide (also a non-TE-derived non-GBM-specific protein) induced high T cell responses (Figure 3I and Figure S4F). Expanded tetramer-positive populations were observed for 15 TE-derived peptides (including 5 from the immunopeptidomic identifications), in at least one donor. These results demonstrate that a subgroup of TEs differentially expressed in GBM can encode HLA-I-binding peptides that are immunogenic *in vitro* in healthy donors and could potentially represent a source of tumor antigens.

Young L1, LTR and SVA subfamilies are main source of TE-derived HLA-I peptides

To investigate the nature of the neoplastic-enriched TEs that encode HLA-I-presented peptides in GBM, we next mapped the peptide sequences to all differentially expressed TEs from the single cell neoplastic TE-signature. In doing so, we realized that although 85.4% of the 370 peptides are encoded by one single TE per peptide, the remaining 15% of peptides could potentially be encoded by 2-200 neoplastic differentially expressed TEs per peptide (Figure 4A). We will refer to these peptides as "single-TE" or "multi-TE" encoded peptides, respectively. Several TEs coding for the same peptide will be referred to as "redundant". For further analyses, regarding redundant TEs, since we cannot determine which TE or TEs encodes the peptide, we considered either all the TEs bearing the peptide-coding nucleotide sequence ("all assignments"), or only one (chosen uniformly) of these TEs per peptide ("single assignment").

We first analyzed the genomic location of the peptide-coding TEs relative to the nearest gene. Among TEs coding for HLA-I-presented peptides, 37.9% and 31.9% (for all and single assignments, respectively) are distal compared to all expressed TEs (12.1%) or to neoplastic differentially expressed TEs (22.3%) (Figure 4B). Analysis of the genomic locations of peptide-coding TEs revealed increased proportions of intergenic TEs (35% and 28.9% for all and single assignments, respectively, compared to 15.2% in the neoplastic TE-signature) (Figure 4C). The proportion of intronic TEs is also increased, but not as much (50% and 50.7% for all and single assignments, respectively, compared to 38.7% in TEs expressed in neoplastic cells). 3' UTR encoded TEs are less frequent in peptide-coding TEs: 25.3% of TEs in the neoplastic TE-signature, and only 5.8% and 7% for all and single assignments, respectively, among peptide-coding TEs. These results establish selectivity in the genomic location of peptide-coding TEs, which are preferentially distal, intergenic, and not present in 3'UTRs.

We next sought to investigate if the identified peptides are preferentially derived from certain TE classes. Based on both all and single assignments, peptide-coding TEs are significantly enriched for LINE elements, which represent around 30% of all expressed or neoplastic differentially expressed TEs, and from 52 to 64%, for all and single assignments of peptide-coding TEs, respectively (Figure 4D, statistics in Figure S5A, and individually for each immunopeptidomic sample in Figure S5B). These TE class analyses also revealed that TEs classified as "Other" are also enriched. This category includes SVA elements and other types of repeats codified in RepeatMasker as RC, RNA, Satellite and Unknown. Among the 51 TE-derived peptides the "Other" category, 23 are from SVA elements (Figure S5C). SINE elements, in contrast, are depleted among peptide-coding TEs (from 51.7% and 44.5% in all expressed TEs and neoplastic differentially expressed TEs, to around 11% in peptide-coding TEs). Therefore, neoplastic differentially expressed LINE elements are a major source of TE-derived peptides presented on HLA-I in GBM.

TEs within each class are classified in families and subfamilies. The evolutionary "age" of these subfamilies can be estimated from the degeneration of their characteristic repeat motifs (Choudhary et al., 2020). We reasoned that the peptides that can be redundantly encoded by multiple TEs could be derived from conserved sequences present in different young TEs from

the same subfamilies. Figure 4E shows the age of all TEs from each class in RepeatMasker, as well as all expressed TEs, neoplastic differentially expressed TEs and peptide-coding TEs. The median age of the peptide-coding SINE and DNA TEs are similar to all genomic TEs annotated in RepeatMasker, and to all expressed and neoplastic TE-signature. For LTRs, the proportion of younger TEs is increased among peptide-coding TEs (decreasing the median age of the peptide-coding TEs compared to other categories), but older TEs are also presented on HLA-I. For LINE and "Other" (see above a more detailed analysis of this category) classes, a bi-modal distribution is observed, with a clear enrichment in peptides encoded by TEs from young subfamilies (under 50 M years) that are rare in RepeatMasker, in all expressed and in neoplastic differentially expressed TEs (Figure 4E). We conclude that among LINE and LTR classes, recent TEs are more prone to provide peptides for HLA-I presentation.

Conserved viral proteins are a source of HLA-I-presented peptides

A few of the youngest subfamilies include TEs that contain intact viral protein ORFs, including a few "active" TEs in terms of retro-transposition (Burns, 2017; Rodic et al., 2015; Scott et al., 2016). We next investigated if peptides from TEs are derived from validated Endogenous Viral Elements (EVE) in the gEVE database (Nakagawa and Takahashi, 2016). These EVEs of at least 80 amino acids were identified processing both RepeatMasker annotations and conserved known motifs from viral proteins such as Gag and Pol. Mapping peptide-coding TEs to gEVE shows that, for both LINEs and LTRs, TEs with an annotated EVE are significantly enriched among peptide-coding TEs (based on both all and single assignments), as compared to RepeatMasker, all expressed and the neoplastic TE-signature (Figure 4F). Consistent with these results, mapping of the TE-derived peptides to annotated EVE protein sequences shows selectivity for Alu among SINEs, L1PA/B/x and L2 among LINEs, ERV1, ERVK, ERVL and ERV-MaLR among LTRs and SVA among "Other" (Figure S5C). Allowing one or two nucleotide mismatches (to take into account possible mutations or polymorphisms) increases markedly the proportion of TE-derived peptides that map to annotated EVE protein sequences, including for classes and families (Figure S5C, middle and right panels), suggesting that recently mutated TEs are also a major source of peptides for HLA-I presentation. Most peptides are derived from ORFs bearing a start codon, either ATG (canonical) or CTG/GTG/TTG (non-canonical) (Figure 4G and Figure S5D). We conclude that TEs from young subfamilies, preferentially bearing retroviral protein motifs, are more prone to provide peptides for presentation on HLA-I molecules in GBM cells. Figure S5E shows an example of 3 peptides encoded by an SVA-family member, SVA_B_dup189. Peptides can be encoded by different reading frames (RFs) and rarely outside of ORFs. In these cases, it could be that the ORF is shorter than 30 nucleotides, that the start codon for this ORF is not among the 4 start codons used in the pipeline or that the start codon is upstream the TE sequence.

Analysis of the length of the ORFs encoding HLA-I-presented peptides shows that among L1PA/B/x, but not among other TE families, ORFs generating peptides and containing a canonical ATG start codon are longer than the ORFs beginning with a non-canonical start

codon (Figure 4H). Among peptide-coding LTRs mapping to a gEVE annotated ORF, ORFs from all retroviral proteins are found in the peptide-coding TE sequences (Figure 4I), with an enrichment for Gag (which represents 10.6% of LTR EVE annotated, vs. 28% of LTR peptide-coding TEs). In the case of LINEs, Pol are the only gEVE annotated proteins (Figure 4I). Blast of the peptide-coding sequences shows that the majority of LINE encoded peptides are not derived from the two major LINE ORFs, ORF1p (3.1%) and ORF2p, (10.8%) (Figure 4J). Therefore, TE-derived peptides are derived from 10 to 1000 amino acids long ORFs bearing canonical or alternative start codons.

TE subfamilies share HLA-I-presented peptide coding sequences

To investigate if some types of TEs are more prone to provide HLA-I-binding peptides than others, we next compared the proportions of TE families among the ones differentially expressed in GBM (and used for the immunopeptidomic search) and the proportions found among the TEs that code for peptides. Figure S6A shows that for LTRs, SINEs and "Other", the proportions of most families are similar between the neoplastic TE-signature and peptide-coding TEs (both with all or single assignments) (Fig S6A, middle and right panels). For LINEs, in contrast, peptides are preferentially derived from L1PA/B/x: 25.3% in the neoplastic TE-signature vs. 76.6% or 49.7% for all and single assignments, respectively. Other LINE families are depleted among peptide-coding TEs (especially L2, which represent 25.1% of LINE in the neoplastic TE-signature and provide for only 7.4% or 15.4% of LINE peptide-coding TEs, with all and main assignments, respectively) (Figure S6A, left panel). Statistical analysis shows significant enrichment in peptide-coding TEs over neoplastic differentially expressed TEs for L1PA/B/x and SVA considering either all or single assignments (Figure 5A and Figure S6B). ERV1 and "Other L1" are enriched with all assignments while on the other hand ERVK are enriched with single assignment. "Other repeats", classified in RepeatMasker as RC, RNA, Satellite and Unknown are also enriched. L2, SINEs (including Alu and MIR) and ERVs (including ERVL and ERVL-MalR) are all significantly depleted among peptide-coding TEs, as compared to neoplastic differentially expressed TEs (Figure 5A and Figure S6B). L1PA/B/x include L1HS (or L1PA1, TE subfamily with few members still active in human genome) and their closely related subfamilies L1PA(x) and L1PB(x), which are all among the younger subfamilies compared to other LINE-1 subfamilies. We conclude that some recent, mainly LINE-1, TE families preferentially generate HLA-I-presented peptides in GBM.

Because recent TEs have more conserved repeat motifs, we next sought to investigate if multi-TE encoded HLA-presented peptides corresponded to shared subfamily motifs. We represented the 152 TE subfamilies coding for the 370 identified HLA peptides in 2-dimensional plots coloring the intersections between 2 subfamilies according to the number of shared peptides. The green diagonal in this plot indicates that most subfamilies code for only one peptide (Figure 5B). The three main groups of TE subfamilies coding shared peptides, or "redundancy clusters", appear as large squares and are enlarged in Figure 5B (bottom panel). The first redundancy cluster corresponds to a group of L1HS and L1PA(x) which are young subfamilies of LINE-1 elements that share up to 25 peptides, pairwise. The second cluster

identifies relatively young SINE elements (mainly Alu) that share single peptides. The third cluster corresponds to a group of young subfamilies of SVA elements that share variable numbers of peptides. Therefore, redundancy occurs within multiple TEs from the same recent related subfamilies that could all potentially code for multiple peptides presented on HLA-I molecules. Redundancy in peptide-coding TEs is therefore limited to a small number of recent TE subfamilies.

To investigate further the links between redundancy and age of TEs, we extended the analysis to all TEs in the genome (redundancy was so far analyzed among the neoplastic TE-signature). Genomic TE-redundancy analysis shows that 49.5% of the 370 peptides identified by immunopeptidomics are encoded by only one TE in the genome (Figure 5C) (as compared to 85.4% in the neoplastic TE-signature, Figure 4A). At the opposite end, 15.9% of these peptides could potentially be encoded by 201-13500 TE occurrences in the genome. A plot of each peptide according to the number of TEs it can potentially be encoded by, and the age of the corresponding subfamilies is shown in Figure 5D. Among SINEs, Alu-derived peptides are highly redundant and from recent subfamilies, while the MIR-derived peptides are encoded by single TEs from older subfamilies. The same correlation is observed among LINE-1 peptides, with young L1HS-, L1PA(x)- and L1PB(x)-derived peptides being encoded by multiple elements, and peptides derived from older L2 and "Other L1" subfamilies by single elements. The negative correlation between the number of TEs potentially encoding single peptides and the age of the corresponding TE subfamilies is confirmed across all TE families (r=-0.61, Figure 5E). We conclude that regardless of TE classes (LINE, SINE, LTR or DNA), subfamilies of young TEs bear shared (redundant) sequences that could code for the same HLA-I peptide, while peptides encoded by TEs from older, more degenerated subfamilies are vastly derived from one genomic sequences.

Ancient single-TE encoded peptides are more tumor-enriched

To investigate how redundancy of TE-derived peptides affects tumor specificity, we next calculated for each TE-derived peptide the ratio between the aggregate expression of all TEs coding for the same peptide in TCGA GBM versus all healthy tissues from GTEx samples (Figure 6A, brown for higher expression in GBM, blue for the opposite). Unsupervised clustering of the aggregate TE expression identifies two main groups of TE-derived peptides, group 1 and 2, dominated by peptide-coding TEs overexpressed in GTEx and in GBM, respectively. Group 1 contain higher proportions of LINEs and “Others” (including all 23 peptide-coding SVA elements), while group 2 contains more LTRs and DNA transposons (Figure 6A, right panels). Moreover, group 1 contains a majority of multi-TE encoded peptides (63.5%), compared to only 26.6% in group 2 (Figure 6A). Consistently, the median age of group 1 TEs is significantly lower than the median age of group 2 (Figure 6B). These results show that single-TE encoded peptides from older TE subfamilies are more likely to be overexpressed in GBM, than TEs from younger subfamilies containing multi-TE encoded peptides.

Can we, then, identify tumor-specific TE peptides? Figure 6C shows expression of the top 50 tumor-enriched peptide-coding TEs in GBM and all GTEx healthy tissues (as 90 percentile expression, left panel, and percentage of samples with higher expression than GBM median expression, right panel). The most tumor-enriched TEs are from diverse classes, but are preferentially derived from ORFs containing a canonical start codon (right histograms in Figure 6C and Figure S7A). Some of these TEs are expressed in a majority of GBM tumors, and undetectable in all, or in a majority, of GTEx healthy tissues (including brain) (Figure 6D). For some of these TEs, over 90% of the cells expressing the TEs are GBM neoplastic cells from all four patients in the scRNAseq data sets (pie charts in Figure 6D, violin plots in Figure S7B). We conclude that a subset of non-redundant peptide-coding TEs are highly tumor-enriched and recurrent in cancer patients. These non-redundant peptide-coding TEs represent interesting potential targets for immunotherapy.

Discussion

In search for tumor specific recurrent antigens, we use a TE-centered proteogenomic approach to investigate HLA-I presentation of TE-derived peptides. We first analyze scRNAseq from total live cells of four primary GBM tumors, to identify individual TEs expressed selectively in GBM tumor cells, and not in hematopoietic or stromal cells. We show that the TEs differentially expressed in neoplastic cells are overexpressed in a cohort of 155 bulk RNAseq samples from GBM patients (TCGA), as compared to all tissues, including brain tissue from healthy donors (GTEx). This neoplastic-enriched TE-signature is used to interrogate MS-based immunopeptidomic data sets from 30 cell lines and primary GBM tumors. We identify 370 TE-derived peptides with reliable profiles and motif compliance to HLA-I alleles of the corresponding samples. These peptides are encoded by 568 TEs, whose analysis revealed some interesting aspects of the biology of HLA-I presentation of peptides from TEs in GBM cells.

Our study relies on scRNAseq mapping of TEs. Several recent papers have analyzed TEs in scRNAseq data sets. Although a few early studies pointed to possible bias and limitations (He et al., 2021; Shao and Wang, 2021), reliable pipelines and guidelines are now available, and have been applied in our study. Our results are also supported by internal controls that confirm the robustness of our TE scRNAseq analyses. First, we show that the TEs expressed in neoplastic GBM cells, but not in other cell populations, are biased for TEs encoded by chromosome 7 (all or intergenic TEs, suggesting that the bias is not due to intronic TE expression) (Figure 1E). Second, the neoplastic TE-signature based on scRNAseq is overexpressed in GBM bulk RNAseq patient cohorts compared to healthy tissues (Figure 2D and 2E). Importantly, the peptides identified in immunopeptidomic data bases are also biased for chromosome 7, further and independently validating our peptide discovery and validation pipelines.

One conclusion of our study is that the proportions of intronic and intergenic TE are increased among peptide-coding TEs, as compared to the neoplastic TE-signature (the database used to identify the peptides), at the expense of 3'UTR TEs. HLA-I-presented peptides can therefore be derived from both gene-dependent and gene-independent transcription and translation, but the reasons why intronic TEs provide proportionally more peptides than 3'UTR TEs is worth further analyses. Previous studies have found that 3'UTRs can code for HLA-presented peptides (Laumont et al., 2016; Ruiz Cuevas et al., 2021; Zhao et al., 2020) but these studies did not consider TEs from other genomic locations, as we do here. We also find that LINE-1 elements are the major source of HLA-I presented peptides in GBM. LINE-1 represent around 30% of TEs in the human genome, of all TEs expressed in GBM, and of neoplastic TE-signature, but over 50% of the TE encoded peptides presented on HLA-I. SVA-derived peptides are also strongly enriched, while the proportion of SINE-derived peptides is reduced (as compared to genomic, expressed and differentially expressed SINEs in GBM). LINE-1 elements with and without intact ORFs are preferentially represented among peptide-

generating TEs and this bias is observed whether TEs are assigned to multiple or to single locations, indicating that the bias is not due to TE mapping issues.

Another conclusion from our study is that HLA-I molecules present peptides that can be encoded by either one or multiple TEs (bearing nucleotide sequence encoding the exact same peptide). Redundancy, in most cases, occurs within TE subfamilies, and in some cases within different subfamilies that are always from the same TE class. The most redundant TEs (from several hundred to several thousand occurrences) are from L1PA/B/x and often bear intact annotated ORFs. Peptides derived from Alu (a SINE family member), ERV1 (an LTR family) and SVA (an intermediate length independent family), which are all among the youngest TE families in humans, are also highly represented and redundant. Redundancy is negatively correlated with the age of the TE subfamily, suggesting that the recurrent sequences encoding HLA-I-binding peptides are part of the ancestral TE insertion event, which subsequently degenerated by mutations and disappeared with time as members of the subfamilies diverged. This scenario is supported by the observation that if 1 or 2 nucleotide mismatches are allowed, the number of redundant TEs is even larger (Figure S5C). This is an intriguing observation and we do not know yet if the peptides identified by mass spectrometry are derived from multiple or single TE loci.

Analysis of the peptide-coding TE ORFs reveal that peptides are generally encoded by 10-100 amino acid long ORFs (with the exception of around half of the LINE-encoded peptides that are derived from longer ORFs). In LTRs, peptides are derived from all viral ORF types, with a positive bias for gag-derived peptides, as compared to the proportion of gag genes annotated in the databases (Figure 4I). Among LINE-derived peptides, only a small proportion (around 10%) are derived from the known ORF1p and ORF2p proteins. The TE-coding ORFs bear either canonical or alternative start codons, with exception of the longer LINE1 ORFs (over 100 amino acids) which are all driven by canonical ATG start codons.

How, then, can we use this knowledge to identify tumor specific TE-derived antigens? Analysis of the relative expression of individual peptide-coding TEs in GBM tumors and a wide series of healthy tissues reveal that redundant TEs from younger subfamilies are generally less tumor-enriched than single TEs from older ones (Figure 6A). Because of their wide tissue expression, it is most likely that the immune system is more tolerized to the antigens from these TEs (although this would need to be addressed specifically). Redundant TEs are therefore probably not the best candidates for tumor-specific targets for immunotherapy, although vaccination with LINE-1 intact ORFs has been shown to be both immunogenic and safe in mice and monkeys (Sacha et al., 2012). Our results, however, also identify non-redundant peptide-coding TEs, that are preferentially from MIR, LINE-1 and -2 and some ERV oldest subfamilies. These non-redundant peptide-coding TEs are in majority from relatively old TE subfamilies (over 50 M years), and tBLASTn analysis show that some of these sequences are present only once in the genome. Some of these TEs are from subfamilies recurrently and selectively de-repressed in tumors, mostly through local DNA demethylation (Brocks et al., 2017; Chiappinelli et al., 2017; Lavie et al., 2005; Ohtani et al., 2020; Roulois et al., 2015; Sacha et al., 2012). We show that some of these peptide-coding

TEs that are expressed in a majority of GBM tumors, are either not detected in healthy tissues or detected at low frequencies and/or low levels (Figure 6). Further studies will investigate if these tumor-enriched peptide-coding TEs can be expressed in other pathological conditions, such as apoptosis or inflammation, in which TE de-repression can be observed.

Our results of *in vitro* stimulation with some of the TE-derived peptides indicate that the TCR repertoire for TEs in healthy individuals exists, opening the possibility that these TEs are immunogenic in patients. Previous studies, however, have shown T cell reactivity against tumor-expressed TEs, establishing the proof of concept that TEs, including ERVs, can be immunogenic in cancer patients (Saini et al., 2020; Smith et al., 2018; Wang-Johanning et al., 2008). In this context, mapping the expression of individual TEs from single-cell and bulk RNAseq in cancer patients proved efficient in defining individual TE occurrences that yield HLA-I-presented peptides. The tumor-enrichment and high recurrence of these peptide-coding TEs opens perspectives for immunotherapies in many cancer types with de-repressed TEs and beyond, in other pathologies in which TEs expression is de-regulated.

Limitations of study

First, our study maps RNAseq reads to annotated TEs. At least in part because TEs are largely repetitive, TE annotation in the human genome is far from perfect (even it has made significant progress in the last few years). Consequently, it is possible that the genomic location and the assignment to a specific locus or subfamily will change in the coming years. Second, even if our analysis of T cell stimulation in healthy PBMCs shows potential immunogenicity, the actual direct demonstration of TE immunogenicity would come from analyses of T cell responses in GBM patients. This is not trivial, as TILs in GBM are rare and difficult to amplify ex vivo. Thirdly, our demonstration that numerous transcribed TEs can potentially code for some of the identified HLA-I-bound peptides, we still don't know which and how many of those TEs actually encode the peptides. Finally, even if the expression of some TEs seems truly specific (absent completely from all healthy tissues), these TEs are very rare (maybe 3 in this study). Tumor-enriched TEs in contrast are quite numerous, and could be sufficiently immunogenic to develop effective immunotherapy tools.

Acknowledgements

We thank all the U932 members, Olivier Lantz, Maude Delost, Nathalie Amzallag and Florence Faure for helpful discussions. Also, we are grateful to Joshua J. Waterfall for helpful discussion and bioinformatics advice. In addition, we thank the core facilities at Curie Institute, including the Bioinformatics groups. S.A received funding from the Institute Curie; Institut National de la Santé et de la Recherche Médicale; Centre National de la Recherche Scientifique. This work has received support under the program «Investissements d'Avenir» launched by the French Government. We thank the Ligue Nationale Contre Le Cancer for the fellowship to Y.A.A.

Author Contributions

C.G. and S.A. conceived and designed the project. P.B. and C.G. designed, performed bioinformatics analyses and interpreted data. A.M. designed, performed *in vitro* vaccination experiments and analyzed the results. Y.A.A. performed and analyzed proteomics data. M.C. and Y.A.A. carried out experimental work about validation of peptides. J.V.X. and Z.B. contributed to experimental work. C.A., Z.B. and E.Z. provided critical revisions to the manuscript. S.A., C.G., P.B., A.M. and Y.A.A. wrote the manuscript and interpreted data.

Declaration of interests

C.A. is a consultant for Biotherapy Partners. SA is shareholder and consultant for Mnemo therapeutics. P.B, Y.A.A, A.M, E.Z, C.G are consultants for Mnemo therapeutics. A patent related to this work has been deposited under number EP22305355 in European Patent Application.

Main figure legends

Figure 1. Single cell TE expression distinguishes cell populations in GBM tumors. (A) tSNE visualizing all single cells after filtering (n = 3,167) segregated based on gene expression (left), TE subfamily expression (middle) and individual TE copy expression (right). Cells are color-coded based on cell population. (B) Violin plots representing TE specific signatures for neoplastic cells (top) and immune cells (bottom). (C) Unsupervised heatmap showing expression of top 20 differentially expressed TEs for each cell population. (D) Plot showing TE subfamily enrichment analysis using all expressed TEs (left), neoplastic (middle) and immune (right) signatures. Red dashes represent adjusted P-value <0.05 on x-axis. (E) Radar plots displaying the rate of genes (top) and TEs (bottom) along all chromosomes. (F) Barplot showing the number of TEs in proximal or distal regions of nearest protein-coding genes in neoplastic and immune signatures. (G) tSNE visualizing cell populations using the individual distal TE copy expression. (H) Plots summarizing the association between TEs and genes described above in neoplastic (top) and immune signatures (bottom).

Figure 2. Single cell neoplastic TE signature is highly enriched in TCGA GBM samples compared to GTEx normal tissues. (A-B) PCA and UMAP projection of TCGA. GBM tumor samples and healthy tissue samples from GTEx based on single cell neoplastic TE signature. (C) Heatmap and hierarchical clustering on TCGA GBM tumor and GTEx normal samples representing z-score of top 2000 TEs from the neoplastic TE-signature. (D) GSEA was performed to assess the specific enrichment of the neoplastic TE-signature in TCGA GBM tumor samples compared to GTEx normal brain samples. (E) Violin plot showing the median expression of single cell neoplastic TE-signature in TCGA GBM and GTEx samples (brain and other tissues) (F) Violin plots showing specific expression of EGFR gene and five individual TEs in bulk and single cell data sets.

Figure 3. GBM-enriched, TE-derived, immunogenic peptides are presented on HLA-I molecules. (A) Workflow for the identification of TE-derived peptides using MS-based immunopeptidomics. (B-C) Venn diagrams summarizing the overlap between neoplastic TE-signatures or TE-derived peptides obtained from uniquely or multi-mapped analysis. (D) Heatmap summarizing TE-derived peptides found in each immunopeptidomic sample analyzed. (E) Boxplot showing the peptide-spectrum identification score (SEQUEST score) from annotated-canonical and TE-derived peptides. (F) HLA-A3 binding motif obtained by GibbsCluster2.0 from TE-derived peptidome (top) and IEDB reference peptides (bottom). (G) Radar plot showing the percentage of peptide-coding TEs among all chromosomes. (H) Examples of expanded tetramer positive CD8 T cells for TE-derived peptides after *in-vitro* immunogenicity assay. (I) Total frequency of tetramer positive populations for HLA-A*02:01 predicted or MS-derived peptides and HLA-B*07:02 MS-derived peptides in each evaluated donor. Lines below indicate peptide mixes used for each donor (n=7). P#: predicted TE-derived peptides; pMS#: MS-derived peptides; mutated Melan-A peptide and N#: normal proteome-derived peptides.

Figure 4. TE-derived peptides are located in long ORFs starting with canonical and non-canonical start codons. (A) Barplot showing the proportion of peptides encoded by one or several TEs from the single cell neoplastic signature. (B-D) Barplots displaying proportions of proximal and distal TEs (B), genomic location proportions (C) and TE class proportions (D) at RNA and peptide levels. (E) Violin plots representing the TE age distribution per class and subset. (F) Barplots showing for different subsets the quantification of LINE and LTR TEs with an Endogenous Viral Element ORF documented in gEVE database. (G) Pie charts showing the percentage of TE-derived peptides found in an ORF with a canonical or non-canonical start codon. (H) Plot showing the peptide-coding TE ORF length distribution depending on the type of start codon. (I) Barplots displaying the proportions of LTR and LINE elements matching an HMM profile of a known viral protein motif in gEVE. (J) Pie chart representing the percentage of LINE-derived peptides matching ORF1p and ORF2p proteins (from Uniprot) using blastp alignment.

Figure 5. TE-derived peptide redundancy depends on the age of TEs. (A) Plot showing TE family enrichment analysis using peptide-coding TEs with all or single assignment(s). (B) Plot showing the number of shared TE-derived peptides among TE subfamilies. A closed-up representation is displayed on the bottom showing the number of shared peptides between L1PA/B/x, SVA and Alu subfamilies. (C) Pie charts displaying the percentage of redundancy of TE-derived peptides. (D) Dot plot representing the median age of peptide-coding TEs in each family classified by TE classes. (E) Correlation plot between the total number of peptide-coding TEs and their median age.

Figure 6. A subset of non-redundant peptide-coding TEs are highly tumor-enriched and recurrent in cancer patients (A) Heatmap displaying the log2 ratio between TCGA GBM and GTEx samples of TE-derived peptides aggregate RNA related expression. GTEx tissues are classified into 5 normal tissue categories defined in (Bradley et al., 2020) (left). Hierarchical clustering identified two groups. Distribution of redundancy and TE classes is shown for each group (right). (B) Plot showing median age of peptide-coding TEs for each group. (C) Heatmaps of top 50 TEs from group 2 coding for single-TE encoded peptide displaying their 90^{th} percentile expression (left) and frequency (middle) in GBM tumor samples and 25 normal tissues from GTEx. ORF length are plotted (right) for each TE. (D) Plots showing TE expression for four examples marked with a star in (C). Median expression for each tissue is indicated with a black line. The percentage of positive cells in each cell type described in scRNAseq is represented using pie charts.

Main table legends

Data S1. Table summarizing several information on TE-derived peptides from immunipeptidomics or on predicted peptides by NetMHCpan, Related to Figure 3.

STAR Methods

Resource Availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Sebastian Amigorena (sebastian.amigorena@curie.fr) and Christel Goudot (christel.goudot@curie.fr).

Materials availability

This study did not generate new unique reagents.

Data and Code availability

All data used in the paper are listed in the key resources table.

All original code has been deposited at Mendeley and is publicly available as of the date of publication. DOIs are listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Experimental Model and Subject Details

Human research participants

Buffy coats from healthy donors were obtained from Etablissement Français du Sang (Paris, France) in accordance with INSERM ethical guidelines. According to French Public Health Law (art L 1121-1-1, art L 1121-1-2), written consent and IRB approval are not required for human non-interventional studies.

In-vitro vaccinations assay

PBMCs were obtained by density gradient separation using Lymphoprep (StemCell Technologies) and phenotyped by FACS using anti-HLA-A2 (clone BB7.2, BD Biosciences) and anti-HLA-B7 antibodies (clone BB7.1, Biolegend). Only HLA-A2+ and/or HLA-B7+ donors were used.
Monocytes and lymphocytes from the same donor were purified as CD14+, CD4+ and CD8+ cells, respectively, by positive selection using magnetic beads (Miltenyi Biotec). Monocyte-derived dendritic cells (mo-DCs) were obtained by differentiation of CD14+ fraction during 5 days at 10^6 cells/ml in RPMI-1650/Glutamax (Gibco), 10% FBS, penicillin (100

U/ml)/streptomycin (100 µg/ml), recombinant human IL-4 (50ng/mL, Miltenyi Biotec) and GM-CSF (10ng/mL, Miltenyi Biotec). Isolated CD4+ and CD8+ T cells were cryopreserved after purification.
After differentiation, mo-DCs were seeded in 24 well plates at $1x10^6$ cells/ml and maturated overnight with LPS (100 ng/ml). Then, culture media was removed, and LPS treated mo-DCs were pulsed during 3h at 37°C with a mix of selected good-binder TE-derived peptides (either predicted or from HLA-I peptidomic data). Each peptide was added at 1 µg/mL final concentration. Finally, peptide-loaded mo-DCs were harvested, pelleted and counted.
Cryopreserved lymphocyte fractions were thawed, and co-cultures were performed by mixing $1x10^6$ CD8+ T cells, $0,1x10^6$ CD4+ T cells and $0,1x10^6$ peptide-loaded mo-DCs (CD8-CD4-mo-DCs ratio: 10:1:1, respectively) in a final volume of 2ml in 24 well plate. Each well was considered as an independent replicate. Total number of replicates was limited by the total number of CD8+ T cells. Without disturbing the cells, half of the media was changed after 5 days and then, the culture was monitored every 3 days until day 15-20. Expansion of specific CD8+ T cell populations was evaluated by FACS using tetramer staining.
X-vivo 15 medium (Lonza) supplemented with penicillin (100 U/ml)/streptomycin (100 µg/ml) (Gibco), 10% FBS, IL-2 (10 U/ml, Novartis) and IL-7 (10 ng/ml, PeproTech) was used as culture media.
As negative control, a replicate using non-peptide pulsed mo-DCs was included. For HLA-A2+ donors, a positive control of T-cell expansions (1 or 2 replicates) using mo-DCs pulsed with mutated Melan-A peptide (ELAGIGILTV) was included. 3 HLA-A*02:01 binding peptides derived from normal proteins were included.

Peptide binding to HLA-A*02:01 and HLA-B*07:02 by tetramer formation.

Predicted peptides were synthetized by GeneCust with a >98% purity. HLA-A*02:01 and HLA-B*07:02 monomers were purchased as easYmers from Immunaware (Copenhagen, Denmark). The binding to HLA-A*02:01 and HLA-B*07:02 of the predicted and MSTE-derived peptides was measured by HLA-I-tetramer complex formation following manufacturer's instructions. Briefly, biotinylated monomers were incubated with synthetic peptides (100 mM) at 18°C during 48h. Then, they were bound to streptavidin-coated beads and stained with PE-conjugated anti-β2-microglobulin antibody. As positive controls for HLA-A*0201-complex formation, CMV pp65 495-503 (NLVPMVATV) and mutated Melan-A (ELAGIGILTV) were used. CMV pp65 417-426 (TPRVTGGGAM) peptide was used for HLA-B*07:02. Binding is represented as percentage of HLA-I-complex formation relative to CMV positive controls. Peptides with HLA-I-complex formation of at least 50% relative to positive control were used in *in-vitro* vaccination experiments.

For tetramer formation, peptide-HLA-I-complexes were tetramerized using different fluorescent streptavidins (PE, APC, BV421, BV711, PE-CF549 and PECy5) at a final concentration of 8 mg/ml. All tetramers were kept at 4°C and used within 2 months.

Tetramer staining and analysis.

Tetramer staining was performed on total cells after in-vitro vaccination experiments by combining 1µl of each tetramer specificity, and two different streptavidin-labelled tetramers per specificity. The staining was performed during 20 min at RT in a final volume of 100 µl of PBS 1% BSA per 1M cells. Then, 100 µl of surface antibody mix containing anti-CD3 BV650 (BD Biosciences) and anti-CD8 PECy7 (BD Biosciences) at 1/200 final dilution was added and incubated for further 20 min at 4°C. Finally, cells were washed twice with PBS-1% BSA and analyzed by flow cytometry. Live/Dead Aqua-405nm (ThermoFisher) was used to exclude dead cells. Data was collected using a ZE5 Cell Analyzer (Bio-Rad) and analyzed using FlowJo v10.3.
Tetramer analysis was done on live, single cells, CD3+CD8+ cells following the strategy described by Andersen *et al*. (Andersen et al., 2012). Expansions were considered positive when positive for both streptavidin-labelled tetramers. Expanded populations for each peptide are represented either as frequencies of total CD8+ cells in each replicate or as total tetramer frequencies among total CD8+ T cells evaluated in all replicated for one donor.

Method Details

Transposable Element annotations

Classification and TE metadata

Transposable Element annotations have been retrieved from two different databases: from Homer repeat gtf annotation file (v4.11.1) based on hg19 (v6.4) UCSC annotations, and from TEtranscript (Jin et al., 2015) hg19 gtf annotation file. Both annotations are based on RepeatMasker database and have been merged based on identical coordinates (Chr, Start, End) to obtain following information on each repeat: Class, Family, Subfamily, Divergence, coordinates. L1 family was subdivided into 2 families: (1) "L1PA/B/x" that include TEs from closely related L1HS, L1PA(x), L1PB(x), L1P(x) subfamilies; and (2) "Other L1" regrouping all other L1 that are not present in "L1PA/B/x". All DNA transposons were classified as DNA. annotatePeaks.pl from Homer was performed to obtain genomic locations (intron, exon, 3'UTR, 5'UTR, intergenic, other) for each individual TE. Closest and intersect tools from bedtools (v2.29.2) have been used to retrieve, for each TE, the distance from closest protein-coding genes from gencode gtf annotation file (Release 19 GRCh37.p13) (Harrow et al., 2006).

Age of TEs

Repeat age was calculated using percentage of divergence for human repeats: Divergence / $(2.2 * 10^{-9})$, following the formula from this article (Choudhary et al., 2020).

Ancient viral protein motif identification

Open reading frame (ORF) locations from Endogenous Viral Elements (EVE) were retrieved from gEVE database. As analyses were performed on human genome version hg19, hg38 gEVE annotations were formatted and adjusted to hg19 using "Lift Genome annotations" tool from UCSC available here: https://genome.ucsc.edu/cgi-bin/hgLiftOver. ORFs coordinates from gEVE annotations and from all individual TEs in the genome were matched to assign an EVE ORF to individual TEs in case of coordinate overlap. 30517 individual TEs overlapped an EVE ORF, with most of them being L1 (mostly L1PA/B/x) and ERV (mostly ERV1, ERVK, ERVL) elements. To identify amino acid sequence similarity between canonical TE proteins from gEVE database and peptides from immunopeptidomics, a blastp (McGinnis and Madden, 2004) was performed between gEVE protein sequences and the immunopeptidomic sequences. No threshold on Evalue was set, and the similarity was estimated and classified in 3 categories: (1) 100% match: no mismatch, no gap and query coverage per HSP to 100%; (2) at most 1 mismatch: 1 mismatch, no gap and query coverage per HSP above 85%; and (3) at most 2 mismatches: 2 mismatches, no gap and query coverage per HSP above 85%.

Analysis of known TE proteins

LTR and LINE proteins

LTRs coding for peptides overlapping an intact ORF were classified as Env, Gag, Pol or Pro using RetroTector annotations from gEVE. For LINE elements, a blastp (v2.12.0+) was performed between LINE-derived peptides and either ORF1p and ORF2p protein sequences found in Uniprot (accession numbers Q9UN81 and O00370). LINE and LTR coding for a peptide were also compared to gEVE HMM profile annotations to classify the TE protein motif found in those TEs.

TE ORF annotations

A homemade R script was used to identify and annotate ORFs from TE sequences. (1) TE nucleotide sequences were formatted to obtain 6 frames using R package Biostrings (v2.58.0) and its functions DNAStringSet and reverseComplement. (2) 6-frame sequences were translated with translate function from Biostrings. (3) Stop codons and methionines were detected using matchPDict function from Biostrings. (4) Peptides from immunopeptidomics were mapped using matchPDict function. (5) ORFik R package (v1.10.13) (Tjeldnes et al., 2021) was used to detect ORF with at least 30bp (3 for the start codon, 8AA*3 for the sequence, 3 for the stop codon) and to keep only the longest ORF. Two different start codon patterns were submitted to detect ORFs: "ATG" for canonical start codons and "ATG|CTG|GTG|TTG" for canonical and non-canonical start codons. ORFs only found using the second pattern were classified as "CTG|GTG|TTG". (6) Length of ORFs were calculated using start and end positions. (7) R package ggplot2 was used to represent all identified ORFs, stop codons, methionines and peptide locations in all 6 frames of the TEs.

Single-cell data analysis

Downloading data and read alignment to genome

Smart-seq2 data (GEO accession number: GSE84465) were downloaded from the Sequence Read Archive (SRA) database using prefetch from SRA Toolkit (v2.10.0). SRA files were converted to fastq files using fastq-dump. Fastq files were 75bp paired-end unstranded reads. Raw RNA reads were mapped to the human genome (hg19) using the 2-pass mode of STAR (version 2.7.1.a) (Dobin et al., 2013) (parameters: --quantMode GeneCounts, --twopassMode Basic, --alignSJDBoverhangMin 1, --bamRemoveDuplicatesType UniqueIdentical, --winAnchorMultimapNmax 1000, --outFilterMultimapNmax 1000, --outFilterScoreMinOverLread 0.33, --outFilterMatchNminOverLread 0.33, --outFilterMismatchNoverLmax 0.04, --outMultimapperOrder Random, --sjdbOverhang 76).

Quantification of gene and TE expression

To compute quantification of TE and gene expression, featureCounts (Liao et al., 2014) from Subread (v1.6.4) was computed on each genome-mapped read files. Different parameters were used depending on the analysis : (1) for gene expression : –p –ignoreDup –g gene_id using gencode gtf annotation file; (2) for TE expression on individual copies (a) considering only uniquely mapping reads: –p –ignoreDup –g transcript_id using TEtranscript hg19 gtf annotation file; (b) considering uniquely and multi-mapping reads : –p –ignoreDup –g transcript_id -M --primary; (3) for TE expression on subfamilies with uniquely and multi-mapping reads : –p –ignoreDup –g gene_id -M --primary. Cell count files were merged into a matrix with a homemade python script (Python 3.6).

Filtering features and cells, normalization and batch correction

Cell metadata and feature raw count matrices were imported to R (v4.0.3) to create a SingleCellExperiment R object. CPM, FPKM and TPM values on gene and TE expression were calculated on raw counts prior to any filtering using scuttle R package (v1.0.4) and its functions: calculateCPM, calculateFPKM, calculateTPM. Cells with low number of counts and low number of features (3 times lower than MAD) were removed using Scater and Scran packages. To remove low expressed features, several filters have been applied depending on the analysis. For TE expression using uniquely-mapped reads (1), individual TEs with less than 1 count/cell in average were removed [22000 individual remaining TEs]. For TE expression using multi-mapped reads (2), individual TEs with less than 5 counts in at least 20 cells were removed to take into account expression in small populations [130028 individual TEs]. For gene expression (3), genes with less than 5 counts in at least 20 cells were removed [19867 genes remaining]. For TE subfamily expression, no filtering was performed [992 subfamilies]. Raw count matrices were then normalized using logNormCounts function from scater R package. After several verifications, a batch effect linked to the plate ID of the cells was identified. To correct it, removeBatchEffect function from limma R package was used providing the plate ID as batch and the cell type as design.

Dimensionality reduction

A single Seurat object was created importing raw, normalized and normalized + corrected feature matrices into different assays. CPM, FPKM and TPM matrices were also imported. Seurat v3 was used for the uniquely mapped read analysis; Seurat v4 was used for the multi-mapped read analysis, the subfamily analysis and the gene analysis. From Seurat, FindVariableFeatures was performed to distinguish the 5000 most variable genes or individual TEs; ScaleData to scale feature expression, RunPCA to compute 75 Principal Components, RunTSNE to perform t-SNE dimension reduction on 50 Principal Components. Dimensionality reduction step was performed on normalized + corrected assay.

Differential expression analysis and enrichment tests

From Seurat, FindAllMarkers was performed on annotated cell types with a threshold of 0.25 foldchange (either natural log with Seurat v3 or log2 with v4) on features expressed in at least 10% of all cells in 1 cell type. Genes, subfamily and individual TE signatures were designed based on FindAllMarkers results using differentially expressed features with an adjusted p-value lower or equal to 0.05. Signature scores were computed with the Seurat function AddModuleScore using the feature signature of interest. This function calculates, for each individual cell, the average expression of each feature from the signature, subtracted by the aggregated expression of control feature sets. TE subfamily enrichment was performed using all annotated individual TEs in the genome (4.6 million TEs) as a reference and either all expressed TEs or individual TE signatures from each population as ours queries. A hypergeometric test was computed using phyper from stats R package (v4.0.3). Then, a False Discovery Rate correction was applied using p.adjust from stats R package.

Radarplot and chromosome distribution

Radarplots representing feature distribution on chromosomes were made using radarchart function from fmsb R package (v0.7.1). Genomic proportions were calculated using all annotated genes and individual TEs from gencode and TEtranscript annotations, respectively.

Bulk RNA-seq data analysis

Downloading, alignment to genome and quantification

Around 50 samples from each GTEx tissue (Consortium, 2013) were randomly targeted and their fastq read files were downloaded using prefetch and fasterq-dump from sratoolkit (v2.10.0). Fastq reads from TCGA-GBM project (Brennan et al., 2013) were downloaded using gdc-client (v1.6.1). Alignment and feature quantification (genes, individual TEs, TE subfamilies) were done in the same protocol described for the Smart-seq2 analysis. Expression was normalized using estimateSizeFactors from DESeq2 R package (v1.30.1) to obtaine normalized counts. TPM values were also computed using calculateTPM function from scuttle. Two subsets of TE expression matrices were obtained for each database: (1) Expression matrices with only TEs from the neoplastic single-cell TE signatures; (2)

Expression matrices with only expressed TEs. TEs were considered as expressed if we could observe at least 5 counts for 20% or more of the samples (considering separately either all samples from TCGA or GTEx database). 130640 TEs were retained for the TCGA samples whereas 192243 TEs were kept for the GTEx samples. Among those, 103585 TEs were common to both databases.

Downstream analysis of bulk RNA-seq samples

Merged neoplastic signature specific matrix with all samples from TCGA and GTEx was imported in a Seurat object. DESeq2 normalized counts and TPM values were both imported. Using normalized counts, ScaleData, RunPCA and RunUMAP were applied to obtain UMAP representations. To assess signature expression in the samples, median expression of all TEs from the neoplastic signature was done using TPM values.

Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) was performed using DESeq2 normalized count matrices of common expressed TEs between TCGA and GTEx databases (103585 TEs) to test enrichment of neoplastic single-cell signature in either Normal or Tumor samples. GSEA (v4.2.1) was run with default parameters. GSEA results were imported to R and ggplot2 was used to make representations.

Mass spectrometry based immunopeptidomics

Mass spectrometry data analysis

MS-based immunopeptidomic files were obtained from PXD020079, PXD008127, PXD003790 and MSV000084442. They were analysed with ProteomeDiscoverer 2.5 (ThermoFisher) using the following parameters: no-enzyme, precursor mass tolerance 20ppm and fragment mass tolerance 0.02 Da. Methionine oxidation and N-acetylation were enabled as variable modifications. Using Percolator, a false discovery rate (FDR) of 1% was applied at peptide level and no FDR was used at protein level. Spectra were searched against the human Uniprot/SwissProt with isoforms (updated 06/03/2020) concatenated with the 6 reading frame in silico translated neoplastic enriched TE database (from uniquely- or multiple-mapped analysis). Identified potential TE-derived peptides were filtered afterwards with UniProt/TrEMBL database, considering leucine-isoleucine and lysine-glutamine as equivalent, respectively. Finally, spectrums from TE-derived peptides were manually verified.

Peptide hydrophobicity index (HI) calculation

For retention time versus hydrophobicity comparisons, HI was predicted using SSRCalc (Krokhin et al. 2004) web server (http://hs2.proteome.ca/SSRCalc/SSRCalcX.html).

Single and all assignments definition

As multiple TEs can code for the same peptides, we made two different categories to make observations on TE-encoded peptide features. *All assignments* correspond to all TEs coding for a peptide (all 568 TEs for 370 peptides). *Single assignment* corresponds to a random selection for each peptide of an individual TE that can encode the corresponding peptide (370 TEs for 370 peptides).

Identifying potential peptide-coding TEs

To identify all TEs coding for peptides identified with immunopeptidomic results, peptide amino acid sequences were aligned to all annotated individual TEs in the genome in all 6 reading frames using tblastn (v2.11.0+). Sequences from all TEs in the genome were retrieved using getfasta from bedtools (v2.30.0) (Quinlan and Hall, 2010) with TETranscript gtf processed into BED format. No restriction on Evalue was requested. All hits with a number of mismatches equal to 0, a number of gap openings equal to 0 and a query coverage per HSP of 100 were kept and considered as peptide-coding TEs in addition to those from the neoplastic signature identified with ProteomeDiscoverer.

Spectrum validation with synthetic peptides

To validate the spectra, 24 of the identified peptides were synthesized (GeneCust) with an HPLC purity of 95% and were injected in a Velos Orbitrap (CID or HCD). Raw files were analysed with ProteomeDiscoverer 2.5 (ThermoFisher). Spectrums were exported and compared to the spectrums derived from the immunopeptidomic analysis. Only PSM with the same charge between synthetic and endogenous and without modifications were analysed. The same fragmentation type (CID or HCD) between both spectrums was prioritized when possible.

Assessing related RNA expression of TE-derived peptides

Identification of tumor-enriched TE-derived peptides

TPM expression of all possible TEs from the genome that can potentially code for the identified peptides was retrieved, and 90^{th} percentile values were calculated for each tissue. TEs coding for each specific peptide were selected and their 90^{th} percentile values were summed to obtain the total transcript expression related to these peptides. For single-TE encoded peptides, related transcript expression was directly the 90^{th} percentile value of the TE coding for the peptides. A log2 ratio was then performed between peptide related expression in GBM samples compared to each GTEx tissue to assess if the related expression of these peptides were higher in GBM samples compared Normal tissues. Using median TPM expression in GBM samples as a threshold, the percentage of expression in normal samples with an equal or higher expression was also calculated for each tissue. Pheatmap function from ComplexHeatmap R package (v2.6.2) was then used to represent the log2 ratio, the 90^{th}

percentile values as well as the percentage of expression in normal samples. Clustering method used in the heatmap with the log2 ratio was ward.D2.

Quantification and Statistical Analysis

Figures

Most figures were made using R (v4.0.3). Piecharts, lollipop charts, barplots, violin plots, boxplots, jitterplots, volcano plots, density plots, scatter plots and dimensionality reduction plot were made using either ggplot2 R package (v3.3.3) (Wickham, 2016) or functions from Seurat package (Butler et al., 2018). Pie donut chart was made with PieDonut function from webR package (v0.1.6). Heatmaps were built with Pheatmap R package (v1.0.12) (Kolde, 2019) and ComplexHeatmap (v2.6.2) (Gu et al., 2016). Clustering method used was ward.D2. IGV (v2.8.10) (Robinson et al., 2017) was used to visualize read coverage of bulk RNA-seq samples.

Statistical Analyses

Wilcoxon tests were performed with R package ggpubr (version 0.4.0) and its function stat_compare_means (1) to compare distance to closest gene between Immune and Neoplastic signatures; (2) to compare mean expression of the neoplastic signature in bulk RNA-seq samples; and (3) to compare length of canonical and non-canonical TE-derived peptides ORFs. Pearson correlation scores were computed using stat_cor from ggpubr : (1) to assess the correlation between TEs and their closest protein-coding gene; and (2) to assess the correlation between median age of TEs coding for a peptide and the number of TEs that can code for the peptide. Two proportions z-test were computed to compare LINE proportions in different subsets of individual TEs. The correspondence between p-values and symbols is as follows: ns: p > 0.05; *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001; ****: p <= 0.0001.

References

Almeida, L.G., Sakabe, N.J., deOliveira, A.R., Silva, M.C., Mundstein, A.S., Cohen, T., Chen, Y.T., Chua, R., Gurung, S., Gnjatic, S.*, et al.* (2009). CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. Nucleic Acids Res *37*, D816-819.

Andersen, R.S., Kvistborg, P., Frosig, T.M., Pedersen, N.W., Lyngaa, R., Bakker, A.H., Shu, C.J., Straten, P., Schumacher, T.N., and Hadrup, S.R. (2012). Parallel detection of antigen-specific T cell responses by combinatorial encoding of MHC multimers. Nat Protoc *7*, 891-902.

Anwar, S.L., Wulaningsih, W., and Lehmann, U. (2017). Transposable Elements in Human Cancer: Causes and Consequences of Deregulation. Int J Mol Sci *18*.

Boon, T., and van der Bruggen, P. (1996). Human tumor antigens recognized by T lymphocytes. J Exp Med *183*, 725-729.

Bradley, S.D., Talukder, A.H., Lai, I., Davis, R., Alvarez, H., Tiriac, H., Zhang, M., Chiu, Y., Melendez, B., Jackson, K.R.*, et al.* (2020). Vestigial-like 1 is a shared targetable cancer-placenta antigen expressed by pancreatic and basal-like breast cancers. Nat Commun *11*, 5332.

Brennan, C.W., Verhaak, R.G., McKenna, A., Campos, B., Noushmehr, H., Salama, S.R., Zheng, S., Chakravarty, D., Sanborn, J.Z., Berman, S.H.*, et al.* (2013). The somatic genomic landscape of glioblastoma. Cell *155*, 462-477.

Brocks, D., Schmidt, C.R., Daskalakis, M., Jang, H.S., Shah, N.M., Li, D., Li, J., Zhang, B., Hou, Y., Laudato, S.*, et al.* (2017). DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. Nat Genet *49*, 1052-1060.

Burns, K.H. (2017). Transposable elements in cancer. Nat Rev Cancer *17*, 415-424.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol *36*, 411-420.

Carreno, B.M., Magrini, V., Becker-Hapak, M., Kaabinejadian, S., Hundal, J., Petti, A.A., Ly, A., Lie, W.R., Hildebrand, W.H., Mardis, E.R.*, et al.* (2015). Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. Science *348*, 803-808.

Castle, J.C., Kreiter, S., Diekmann, J., Lower, M., van de Roemer, N., de Graaf, J., Selmi, A., Diken, M., Boegel, S., Paret, C.*, et al.* (2012). Exploiting the mutanome for tumor vaccination. Cancer Res *72*, 1081-1091.

Chauvin, J.M., Pagliano, O., Fourcade, J., Sun, Z., Wang, H., Sander, C., Kirkwood, J.M., Chen, T.H., Maurer, M., Korman, A.J.*, et al.* (2015). TIGIT and PD-1 impair tumor antigen-specific CD8(+) T cells in melanoma patients. J Clin Invest *125*, 2046-2058.

Chiappinelli, K.B., Strissel, P.L., Desrichard, A., Li, H., Henke, C., Akman, B., Hein, A., Rote, N.S., Cope, L.M., Snyder, A.*, et al.* (2017). Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. Cell *169*, 361.

Chong, C., Muller, M., Pak, H., Harnett, D., Huber, F., Grun, D., Leleu, M., Auger, A., Arnaud, M., Stevenson, B.J.*, et al.* (2020). Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. Nat Commun *11*, 1293.

Choudhary, M.N., Friedman, R.Z., Wang, J.T., Jang, H.S., Zhuo, X., and Wang, T. (2020). Co-opted transposons help perpetuate conserved higher-order chromosomal structures. Genome Biol *21*, 16.

Consortium, G.T. (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580-585.

Coulie, P.G., Lehmann, F., Lethe, B., Herman, J., Lurquin, C., Andrawiss, M., and Boon, T. (1995). A mutated intron sequence codes for an antigenic peptide recognized by cytolytic T lymphocytes on a human melanoma. Proc Natl Acad Sci U S A *92*, 7976-7980.
Darmanis, S., Sloan, S.A., Croote, D., Mignardi, M., Chernikova, S., Samghababi, P., Zhang, Y., Neff, N., Kowarsky, M., Caneda, C.*, et al.* (2017). Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. Cell Rep *21*, 1399-1410.
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.
Ehx, G., Larouche, J.D., Durette, C., Laverdure, J.P., Hesnard, L., Vincent, K., Hardy, M.P., Theriault, C., Rulleau, C., Lanoix, J.*, et al.* (2021). Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. Immunity *54*, 737-752 e710.
Forlani, G., Michaux, J., Pak, H., Huber, F., Marie Joseph, E.L., Ramia, E., Stevenson, B.J., Linnebacher, M., Accolla, R.S., and Bassani-Sternberg, M. (2021). CIITA-Transduced Glioblastoma Cells Uncover a Rich Repertoire of Clinically Relevant Tumor-Associated HLA-II Antigens. Mol Cell Proteomics *20*, 100032.
Garcia-Perez, J.L., Widmann, T.J., and Adams, I.R. (2016). The impact of transposable elements on mammalian development. Development *143*, 4101-4114.
Gromeier, M., Brown, M.C., Zhang, G., Lin, X., Chen, Y., Wei, Z., Beaubier, N., Yan, H., He, Y., Desjardins, A.*, et al.* (2021). Very low mutation burden is a feature of inflamed recurrent glioblastomas responsive to cancer immunotherapy. Nat Commun *12*, 352.
Grundy, E.E., Diab, N., and Chiappinelli, K.B. (2021). Transposable element regulation and expression in cancer. FEBS J.
Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847-2849.
Gubin, M.M., Zhang, X., Schuster, H., Caron, E., Ward, J.P., Noguchi, T., Ivanova, Y., Hundal, J., Arthur, C.D., Krebber, W.J.*, et al.* (2014). Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. Nature *515*, 577-581.
Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D.*, et al.* (2006). GENCODE: producing a reference annotation for ENCODE. Genome Biol *7 Suppl 1*, S4 1-9.
He, J., Babarinde, I.A., Sun, L., Xu, S., Chen, R., Shi, J., Wei, Y., Li, Y., Ma, G., Zhuang, Q.*, et al.* (2021). Identifying transposable element expression dynamics and heterogeneity during development at the single-cell level with a processing pipeline scTE. Nat Commun *12*, 1456.
Jin, Y., Tam, O.H., Paniagua, E., and Hammell, M. (2015). TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics *31*, 3593-3599.
Kolde, R. (2019). pheatmap: Pretty Heatmaps. R package version 1.0. 12. R Packag version 10 *8*.
Kong, Y., Rose, C.M., Cass, A.A., Williams, A.G., Darwish, M., Lianoglou, S., Haverty, P.M., Tong, A.J., Blanchette, C., Albert, M.L.*, et al.* (2019). Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. Nat Commun *10*, 5228.
Kurscheid, S., Bady, P., Sciuscio, D., Samarzija, I., Shay, T., Vassallo, I., Criekinge, W.V., Daniel, R.T., van den Bent, M.J., Marosi, C.*, et al.* (2015). Chromosome 7 gain and DNA hypermethylation at the HOXA10 locus are associated with expression of a stem cell related HOX-signature in glioblastoma. Genome Biol *16*, 16.

Lanciano, S., and Cristofari, G. (2020). Measuring and interpreting transposable element expression. Nat Rev Genet *21*, 721-736.
Laumont, C.M., Daouda, T., Laverdure, J.P., Bonneil, E., Caron-Lizotte, O., Hardy, M.P., Granados, D.P., Durette, C., Lemieux, S., Thibault, P.*, et al.* (2016). Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. Nat Commun *7*, 10238.
Laumont, C.M., Vincent, K., Hesnard, L., Audemard, E., Bonneil, E., Laverdure, J.P., Gendron, P., Courcelles, M., Hardy, M.P., Cote, C.*, et al.* (2018). Noncoding regions are the main source of targetable tumor-specific antigens. Sci Transl Med *10*.
Lavie, L., Kitova, M., Maldener, E., Meese, E., and Mayer, J. (2005). CpG methylation directly regulates transcriptional activity of the human endogenous retrovirus family HERV-K(HML-2). J Virol *79*, 876-883.
Le, D.T., Uram, J.N., Wang, H., Bartlett, B.R., Kemberling, H., Eyring, A.D., Skora, A.D., Luber, B.S., Azad, N.S., Laheru, D.*, et al.* (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med *372*, 2509-2520.
Lennerz, V., Fatho, M., Gentilini, C., Frye, R.A., Lifke, A., Ferel, D., Wolfel, C., Huber, C., and Wolfel, T. (2005). The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. Proc Natl Acad Sci U S A *102*, 16013-16018.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.
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.
Lynch-Sutherland, C.F., Chatterjee, A., Stockwell, P.A., Eccles, M.R., and Macaulay, E.C. (2020). Reawakening the Developmental Origins of Cancer Through Transposable Elements. Front Oncol *10*, 468.
McGinnis, S., and Madden, T.L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res *32*, W20-25.
McGrail, D.J., Pilie, P.G., Rashid, N.U., Voorwerk, L., Slagter, M., Kok, M., Jonasch, E., Khasraw, M., Heimberger, A.B., Lim, B.*, et al.* (2021). High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. Ann Oncol *32*, 661-672.
Morotti, M., Albukhari, A., Alsaadi, A., Artibani, M., Brenton, J.D., Curbishley, S.M., Dong, T., Dustin, M.L., Hu, Z., McGranahan, N.*, et al.* (2021). Promises and challenges of adoptive T-cell therapies for solid tumours. Br J Cancer *124*, 1759-1776.
Nakagawa, S., and Takahashi, M.U. (2016). gEVE: a genome-based endogenous viral element database provides comprehensive viral protein-coding sequences in mammalian genomes. Database (Oxford) *2016*.
Neukirch, L., Nielsen, T.K., Laursen, H., Daradoumis, J., Thirion, C., and Holst, P.J. (2019). Adenovirus based virus-like-vaccines targeting endogenous retroviruses can eliminate growing colorectal cancers in mice. Oncotarget *10*, 1458-1472.
Novellino, L., Castelli, C., and Parmiani, G. (2005). A listing of human tumor antigens recognized by T cells: March 2004 update. Cancer Immunol Immunother *54*, 187-207.
Ohtani, H., Orskov, A.D., Helbo, A.S., Gillberg, L., Liu, M., Zhou, W., Ungerstedt, J., Hellstrom-Lindberg, E., Sun, W., Liang, G.*, et al.* (2020). Activation of a Subset of Evolutionarily Young Transposable Elements and Innate Immunity Are Linked to Clinical Responses to 5-Azacytidine. Cancer Res *80*, 2441-2450.
Pittet, M.J., Valmori, D., Dunbar, P.R., Speiser, D.E., Lienard, D., Lejeune, F., Fleischhauer, K., Cerundolo, V., Cerottini, J.C., and Romero, P. (1999). High frequencies of naive Melan-

A/MART-1-specific CD8(+) T cells in a large proportion of human histocompatibility leukocyte antigen (HLA)-A2 individuals. J Exp Med *190*, 705-715.
Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.
Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res *42*, W187-191.
Rapoport, A.P., Stadtmauer, E.A., Binder-Scholl, G.K., Goloubeva, O., Vogl, D.T., Lacey, S.F., Badros, A.Z., Garfall, A., Weiss, B., Finklestein, J.*, et al.* (2015). NY-ESO-1-specific TCR-engineered T cells mediate sustained antigen-specific antitumor effects in myeloma. Nat Med *21*, 914-921.
Rizvi, N.A., Hellmann, M.D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J.J., Lee, W., Yuan, J., Wong, P., Ho, T.S.*, et al.* (2015). Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science *348*, 124-128.
Robbins, P.F., El-Gamil, M., Li, Y.F., Kawakami, Y., Loftus, D., Appella, E., and Rosenberg, S.A. (1996). A mutated beta-catenin gene encodes a melanoma-specific antigen recognized by tumor infiltrating lymphocytes. J Exp Med *183*, 1185-1192.
Robbins, P.F., Lu, Y.C., El-Gamil, M., Li, Y.F., Gross, C., Gartner, J., Lin, J.C., Teer, J.K., Cliften, P., Tycksen, E.*, et al.* (2013). Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. Nat Med *19*, 747-752.
Robinson, J.T., Thorvaldsdottir, H., Wenger, A.M., Zehir, A., and Mesirov, J.P. (2017). Variant Review with the Integrative Genomics Viewer. Cancer Res *77*, e31-e34.
Rodic, N., Steranka, J.P., Makohon-Moore, A., Moyer, A., Shen, P., Sharma, R., Kohutek, Z.A., Huang, C.R., Ahn, D., Mita, P.*, et al.* (2015). Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. Nat Med *21*, 1060-1064.
Rooney, M.S., Shukla, S.A., Wu, C.J., Getz, G., and Hacohen, N. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell *160*, 48-61.
Roulois, D., Loo Yau, H., Singhania, R., Wang, Y., Danesh, A., Shen, S.Y., Han, H., Liang, G., Jones, P.A., Pugh, T.J.*, et al.* (2015). DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. Cell *162*, 961-973.
Ruiz Cuevas, M.V., Hardy, M.P., Holly, J., Bonneil, E., Durette, C., Courcelles, M., Lanoix, J., Cote, C., Staudt, L.M., Lemieux, S.*, et al.* (2021). Most non-canonical proteins uniquely populate the proteome or immunopeptidome. Cell Rep *34*, 108815.
Rycaj, K., Plummer, J.B., Yin, B., Li, M., Garza, J., Radvanyi, L., Ramondetta, L.M., Lin, K., Johanning, G.L., Tang, D.G.*, et al.* (2015). Cytotoxicity of human endogenous retrovirus K-specific T cells toward autologous ovarian cancer cells. Clin Cancer Res *21*, 471-483.
Sacha, J.B., Kim, I.J., Chen, L., Ullah, J.H., Goodwin, D.A., Simmons, H.A., Schenkman, D.I., von Pelchrzim, F., Gifford, R.J., Nimityongskul, F.A.*, et al.* (2012). Vaccination with cancer- and HIV infection-associated endogenous retrotransposable elements is safe and immunogenic. J Immunol *189*, 1467-1479.
Saini, S.K., Orskov, A.D., Bjerregaard, A.M., Unnikrishnan, A., Holmberg-Thyden, S., Borch, A., Jensen, K.V., Anande, G., Bentzen, A.K., Marquard, A.M.*, et al.* (2020). Human endogenous retroviruses form a reservoir of T cell targets in hematological cancers. Nat Commun *11*, 5660.
Sarkizova, S., Klaeger, S., Le, P.M., Li, L.W., Oliveira, G., Keshishian, H., Hartigan, C.R., Zhang, W., Braun, D.A., Ligon, K.L.*, et al.* (2020). A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nat Biotechnol *38*, 199-209.
Schadendorf, D., Hodi, F.S., Robert, C., Weber, J.S., Margolin, K., Hamid, O., Patt, D., Chen, T.T., Berman, D.M., and Wolchok, J.D. (2015). Pooled Analysis of Long-Term Survival Data From Phase II and Phase III Trials of Ipilimumab in Unresectable or Metastatic Melanoma. J Clin Oncol *33*, 1889-1894.

Scott, E.C., Gardner, E.J., Masood, A., Chuang, N.T., Vertino, P.M., and Devine, S.E. (2016). A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res *26*, 745-755.
Shao, W., and Wang, T. (2021). Transcript assembly improves expression quantification of transposable elements in single-cell RNA-seq data. Genome Res *31*, 88-100.
Shraibman, B., Barnea, E., Kadosh, D.M., Haimovich, Y., Slobodin, G., Rosner, I., Lopez-Larrea, C., Hilf, N., Kuttruff, S., Song, C.*, et al.* (2018). Identification of Tumor Antigens Among the HLA Peptidomes of Glioblastoma Tumors and Plasma. Mol Cell Proteomics *17*, 2132-2145.
Shraibman, B., Kadosh, D.M., Barnea, E., and Admon, A. (2016). Human Leukocyte Antigen (HLA) Peptides Derived from Tumor Antigens Induced by Inhibition of DNA Methylation for Development of Drug-facilitated Immunotherapy. Mol Cell Proteomics *15*, 3058-3070.
Simpson, A.J., Caballero, O.L., Jungbluth, A., Chen, Y.T., and Old, L.J. (2005). Cancer/testis antigens, gametogenesis and cancer. Nat Rev Cancer *5*, 615-625.
Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet *8*, 272-285.
Smith, C.C., Beckermann, K.E., Bortone, D.S., De Cubas, A.A., Bixby, L.M., Lee, S.J., Panda, A., Ganesan, S., Bhanot, G., Wallen, E.M.*, et al.* (2018). Endogenous retroviral signatures predict immunotherapy response in clear cell renal cell carcinoma. J Clin Invest *128*, 4804-4820.
Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J.M., Desrichard, A., Walsh, L.A., Postow, M.A., Wong, P., Ho, T.S.*, et al.* (2014). Genetic basis for clinical response to CTLA-4 blockade in melanoma. N Engl J Med *371*, 2189-2199.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.
Tjeldnes, H., Labun, K., Torres Cleuren, Y., Chyzynska, K., Swirski, M., and Valen, E. (2021). ORFik: a comprehensive R toolkit for the analysis of translation. BMC Bioinformatics *22*, 336.
Tran, E., Ahmadzadeh, M., Lu, Y.C., Gros, A., Turcotte, S., Robbins, P.F., Gartner, J.J., Zheng, Z., Li, Y.F., Ray, S.*, et al.* (2015). Immunogenicity of somatic mutations in human gastrointestinal cancers. Science *350*, 1387-1390.
van der Bruggen, P., Traversari, C., Chomez, P., Lurquin, C., De Plaen, E., Van den Eynde, B., Knuth, A., and Boon, T. (1991). A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. Science *254*, 1643-1647.
van Rooij, N., van Buuren, M.M., Philips, D., Velds, A., Toebes, M., Heemskerk, B., van Dijk, L.J., Behjati, S., Hilkmann, H., El Atmioui, D.*, et al.* (2013). Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. J Clin Oncol *31*, e439-442.
Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., and Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res *47*, D339-D343.
Waldman, A.D., Fritz, J.M., and Lenardo, M.J. (2020). A guide to cancer immunotherapy: from T cell basic science to clinical practice. Nat Rev Immunol *20*, 651-668.
Wang-Johanning, F., Radvanyi, L., Rycaj, K., Plummer, J.B., Yan, P., Sastry, K.J., Piyathilake, C.J., Hunt, K.K., and Johanning, G.L. (2008). Human endogenous retrovirus K triggers an antigen-specific immune response in breast cancer patients. Cancer Res *68*, 5869-5877.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).
Wickham, H., François, R., Henry, L., and Müller, K. (2022). dplyr: A Grammar of Data Manipulation.
Wickham, H., and Girlich, M. (2022). tidyr: Tidy Messy Data.
Yarchoan, M., Albacker, L.A., Hopkins, A.C., Montesion, M., Murugesan, K., Vithayathil, T.T., Zaidi, N., Azad, N.S., Laheru, D.A., Frampton, G.M., *et al.* (2019). PD-L1 expression and tumor mutational burden are independent biomarkers in most cancers. JCI Insight *4*.
Yarchoan, M., Hopkins, A., and Jaffee, E.M. (2017). Tumor Mutational Burden and Response Rate to PD-1 Inhibition. N Engl J Med *377*, 2500-2501.
Zhao, Q., Laverdure, J.P., Lanoix, J., Durette, C., Cote, C., Bonneil, E., Laumont, C.M., Gendron, P., Vincent, K., Courcelles, M., *et al.* (2020). Proteogenomics Uncovers a Vast Repertoire of Shared Tumor-Specific Antigens in Ovarian Cancer. Cancer Immunol Res *8*, 544-555.

Figure 1

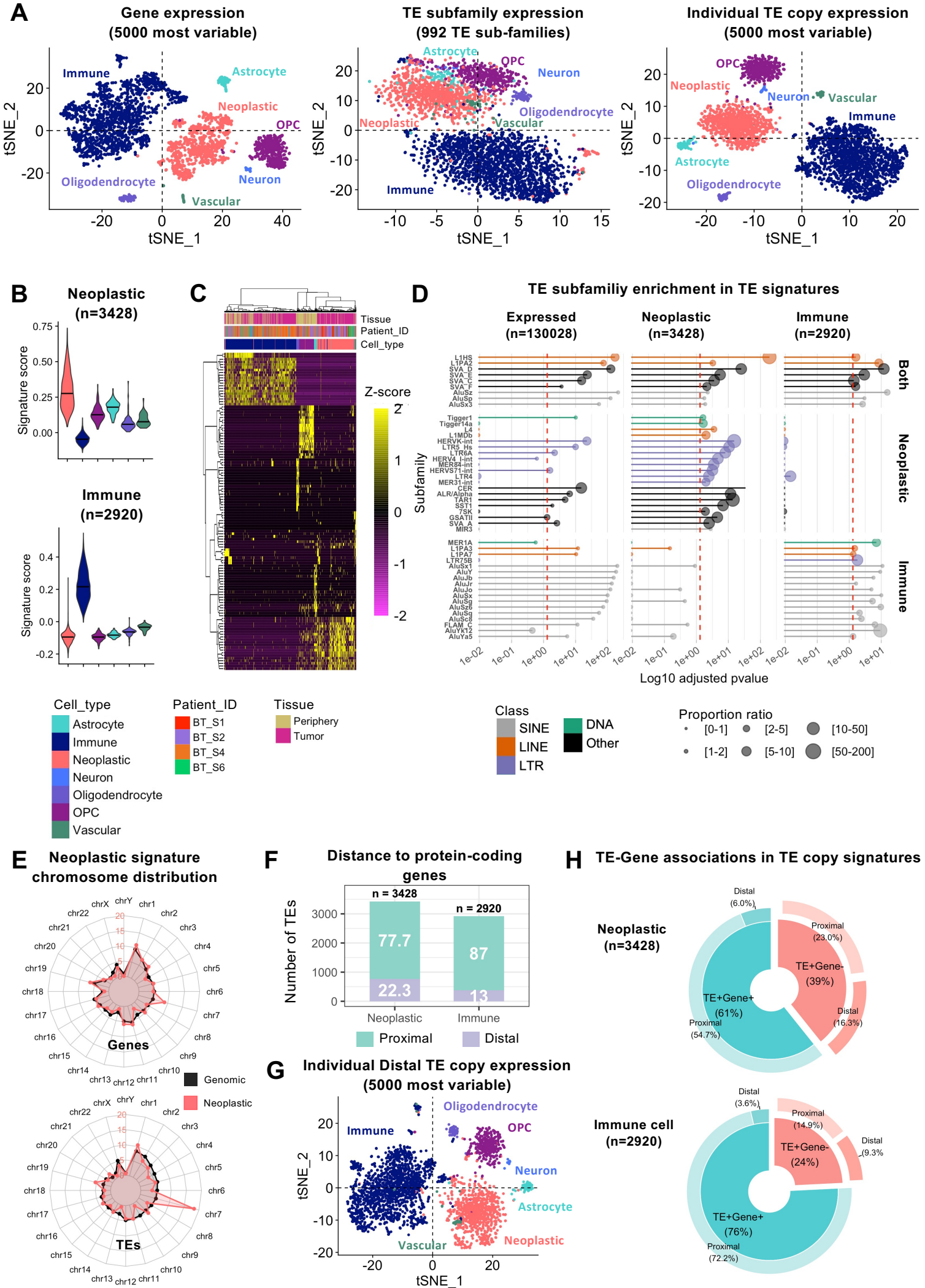

A
Gene expression (5000 most variable)
TE subfamily expression (992 TE sub-families)
Individual TE copy expression (5000 most variable)
tSNE_1
tSNE_2
Immune
Astrocyte
Neoplastic
OPC
Neuron
Oligodendrocyte
Vascular
B
Neoplastic (n=3428)
Immune (n=2920)
Signature score
C
Tissue
Patient_ID
Cell_type
Z-score
D
TE subfamiliy enrichment in TE signatures
Expressed (n=130028)
Neoplastic (n=3428)
Immune (n=2920)
Both
Neoplastic
Immune
Subfamily
Log10 adjusted pvalue
Cell_type
Astrocyte
Immune
Neoplastic
Neuron
Oligodendrocyte
OPC
Vascular
Patient_ID
BT_S1
BT_S2
BT_S4
BT_S6
Tissue
Periphery
Tumor
Class
SINE
LINE
LTR
DNA
Other
Proportion ratio
[0-1]
[1-2]
[2-5]
[5-10]
[10-50]
[50-200]
E
Neoplastic signature chromosome distribution
Genes
TEs
Genomic
Neoplastic
F
Distance to protein-coding genes
Number of TEs
n = 3428
n = 2920
77.7
22.3
87
13
Neoplastic
Immune
Proximal
Distal
G
Individual Distal TE copy expression (5000 most variable)
H
TE-Gene associations in TE copy signatures
Neoplastic (n=3428)
Distal (6.0%)
Proximal (23.0%)
TE+Gene- (39%)
TE+Gene+ (61%)
Distal (16.3%)
Proximal (54.7%)
Immune cell (n=2920)
Distal (3.6%)
Proximal (14.9%)
Distal (9.3%)
TE+Gene- (24%)
TE+Gene+ (76%)
Proximal (72.2%)

Figure 2

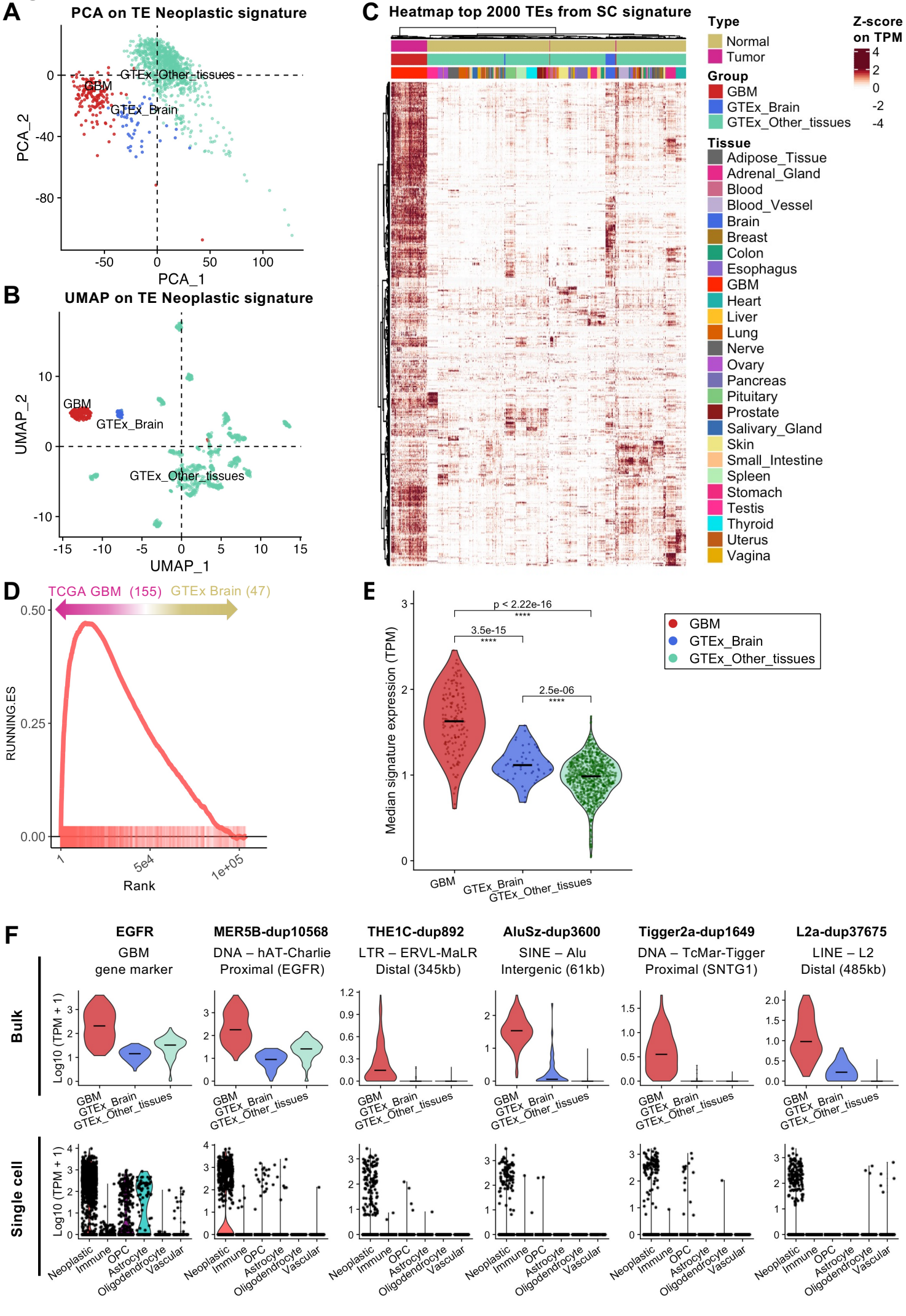

A
PCA on TE Neoplastic signature
GTEx_Other_tissues
GBM
GTEx_Brain
PCA_2
PCA_1
B
UMAP on TE Neoplastic signature
GBM
GTEx_Brain
GTEx_Other_tissues
UMAP_2
UMAP_1
C
Heatmap top 2000 TEs from SC signature
Type
Normal
Tumor
Group
GBM
GTEx_Brain
GTEx_Other_tissues
Tissue
Adipose_Tissue
Adrenal_Gland
Blood
Blood_Vessel
Brain
Breast
Colon
Esophagus
GBM
Heart
Liver
Lung
Nerve
Ovary
Pancreas
Pituitary
Prostate
Salivary_Gland
Skin
Small_Intestine
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Z-score on TPM
D
TCGA GBM (155)
GTEx Brain (47)
RUNNING.ES
Rank
E
p < 2.22e-16
3.5e-15
2.5e-06
Median signature expression (TPM)
GBM
GTEx_Brain
GTEx_Other_tissues
F
EGFR
GBM gene marker
MER5B-dup10568
DNA – hAT-Charlie
Proximal (EGFR)
THE1C-dup892
LTR – ERVL-MaLR
Distal (345kb)
AluSz-dup3600
SINE – Alu
Intergenic (61kb)
Tigger2a-dup1649
DNA – TcMar-Tigger
Proximal (SNTG1)
L2a-dup37675
LINE – L2
Distal (485kb)
Bulk
Single cell
Log10 (TPM + 1)
Neoplastic
Immune
OPC
Astrocyte
Oligodendrocyte
Vascular

Figure 3

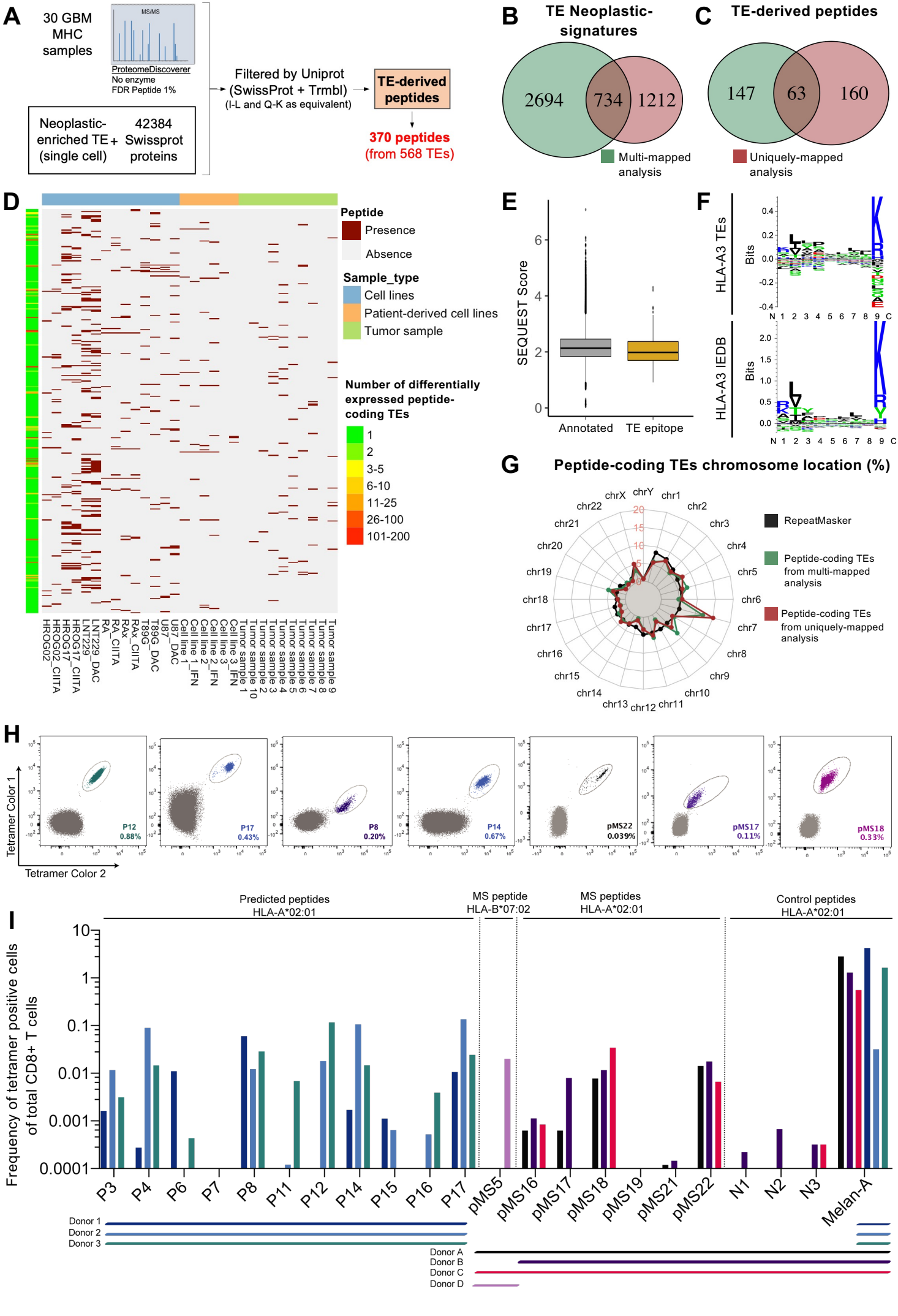

A
30 GBM MHC samples
MS/MS
ProteomeDiscoverer
No enzyme
FDR Peptide 1%
Neoplastic-enriched TE (single cell)
42384 Swissprot proteins
Filtered by Uniprot (SwissProt + Trmbl)
(I-L and Q-K as equivalent)
TE-derived peptides
370 peptides (from 568 TEs)
B
TE Neoplastic-signatures
2694
734
1212
Multi-mapped analysis
C
TE-derived peptides
147
63
160
Uniquely-mapped analysis
D
Peptide
Presence
Absence
Sample_type
Cell lines
Patient-derived cell lines
Tumor sample
Number of differentially expressed peptide-coding TEs
1
2
3-5
6-10
11-25
26-100
101-200
HROG02_CIITA
HROG17_CIITA
LNT229_DAC
RA_CIITA
RAx_CIITA
T89G_DAC
U87_DAC
Cell line 1_IFN
Cell line 2_IFN
Cell line 3_IFN
Tumor sample 1
Tumor sample 2
Tumor sample 3
Tumor sample 4
Tumor sample 5
Tumor sample 6
Tumor sample 7
Tumor sample 8
Tumor sample 9
E
SEQUEST Score
Annotated
TE epitope
F
HLA-A3 TEs
HLA-A3 IEDB
Bits
G
Peptide-coding TEs chromosome location (%)
RepeatMasker
Peptide-coding TEs from multi-mapped analysis
Peptide-coding TEs from uniquely-mapped analysis
H
Tetramer Color 1
Tetramer Color 2
P12 0.88%
P17 0.43%
P8 0.20%
P14 0.67%
pMS22 0.039%
pMS17 0.11%
pMS18 0.33%
I
Predicted peptides HLA-A*02:01
MS peptide HLA-B*07:02
MS peptides HLA-A*02:01
Control peptides HLA-A*02:01
Frequency of tetramer positive cells of total CD8+ T cells
P3
P4
P6
P7
P8
P11
P12
P14
P15
P16
P17
pMS5
pMS16
pMS17
pMS18
pMS19
pMS21
pMS22
N1
N2
N3
Melan-A
Donor 1
Donor 2
Donor 3
Donor A
Donor B
Donor C
Donor D

Figure 4

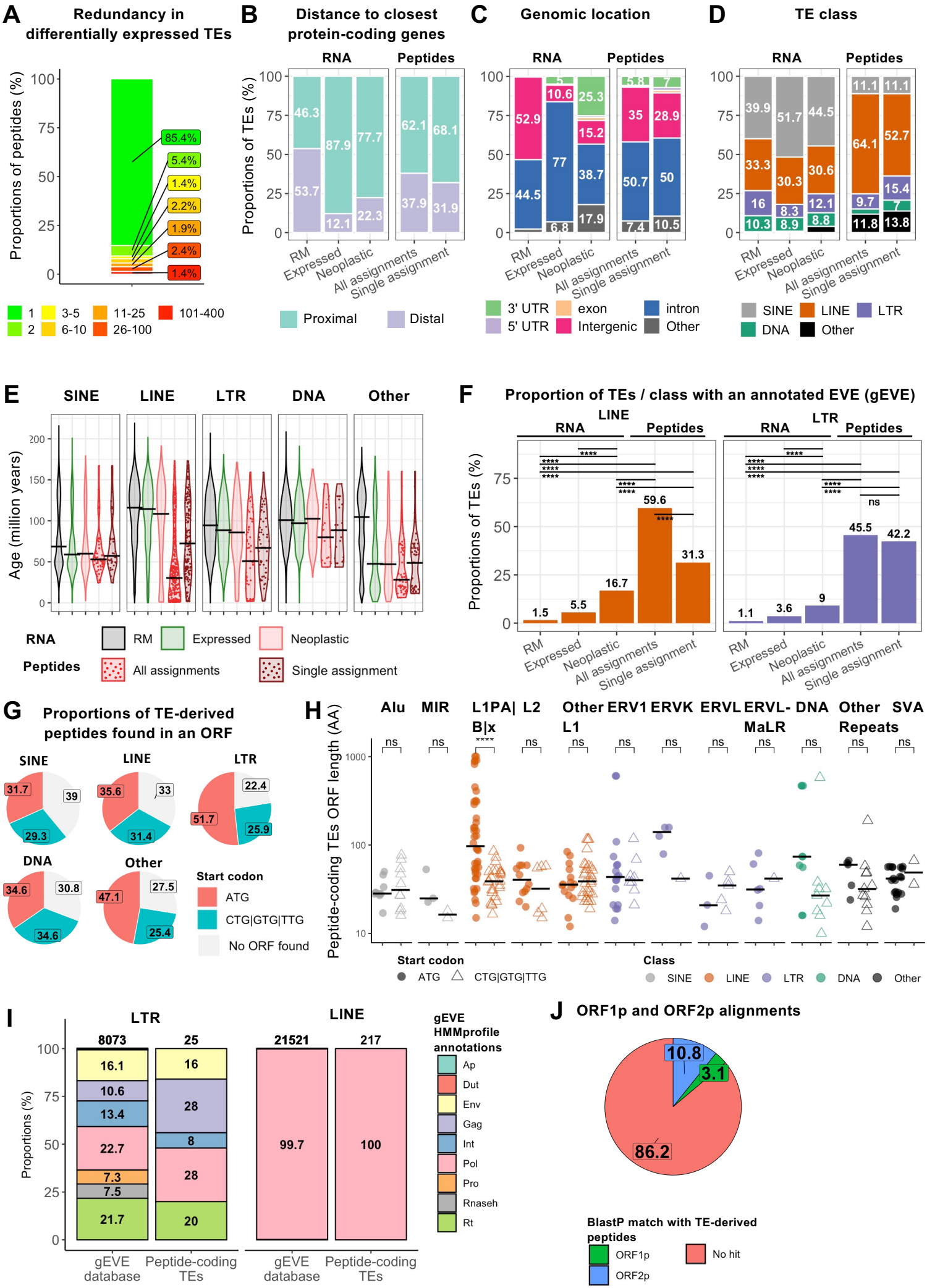

A
Redundancy in differentially expressed TEs
Proportions of peptides (%)
85.4%
5.4%
1.4%
2.2%
1.9%
2.4%
1.4%
1
2
3-5
6-10
11-25
26-100
101-400
B
Distance to closest protein-coding genes
RNA
Peptides
Proportions of TEs (%)
46.3
53.7
87.9
12.1
77.7
22.3
62.1
37.9
68.1
31.9
RM
Expressed
Neoplastic
All assignments
Single assignment
Proximal
Distal
C
Genomic location
52.9
44.5
5
10.6
77
6.8
25.3
15.2
38.7
17.9
5.8
35
50.7
7.4
7
28.9
50
10.5
3' UTR
5' UTR
exon
Intergenic
intron
Other
D
TE class
39.9
33.3
16
10.3
51.7
30.3
8.3
8.9
44.5
30.6
12.1
8.8
11.1
64.1
9.7
11.8
11.1
52.7
15.4
7
13.8
SINE
LINE
LTR
DNA
Other
E
SINE
LINE
LTR
DNA
Other
Age (million years)
RNA
RM
Expressed
Neoplastic
Peptides
All assignments
Single assignment
F
Proportion of TEs / class with an annotated EVE (gEVE)
LINE
LTR
RNA
Peptides
Proportions of TEs (%)
1.5
5.5
16.7
59.6
31.3
1.1
3.6
9
45.5
42.2
ns
G
Proportions of TE-derived peptides found in an ORF
SINE
LINE
LTR
DNA
Other
31.7
39
29.3
35.6
33
31.4
22.4
51.7
25.9
34.6
30.8
34.6
47.1
27.5
25.4
Start codon
ATG
CTG|GTG|TTG
No ORF found
H
Alu
MIR
L1PA|B|x
L2
Other L1
ERV1
ERVK
ERVL
ERVL-MaLR
DNA
Other Repeats
SVA
Peptide-coding TEs ORF length (AA)
ns
****
Start codon
ATG
CTG|GTG|TTG
Class
SINE
LINE
LTR
DNA
Other
I
LTR
LINE
8073
25
21521
217
Proportions (%)
16.1
10.6
13.4
22.7
7.3
7.5
21.7
16
28
8
28
20
99.7
100
gEVE database
Peptide-coding TEs
gEVE HMMprofile annotations
Ap
Dut
Env
Gag
Int
Pol
Pro
Rnaseh
Rt
J
ORF1p and ORF2p alignments
10.8
3.1
86.2
BlastP match with TE-derived peptides
ORF1p
ORF2p
No hit

Figure 5

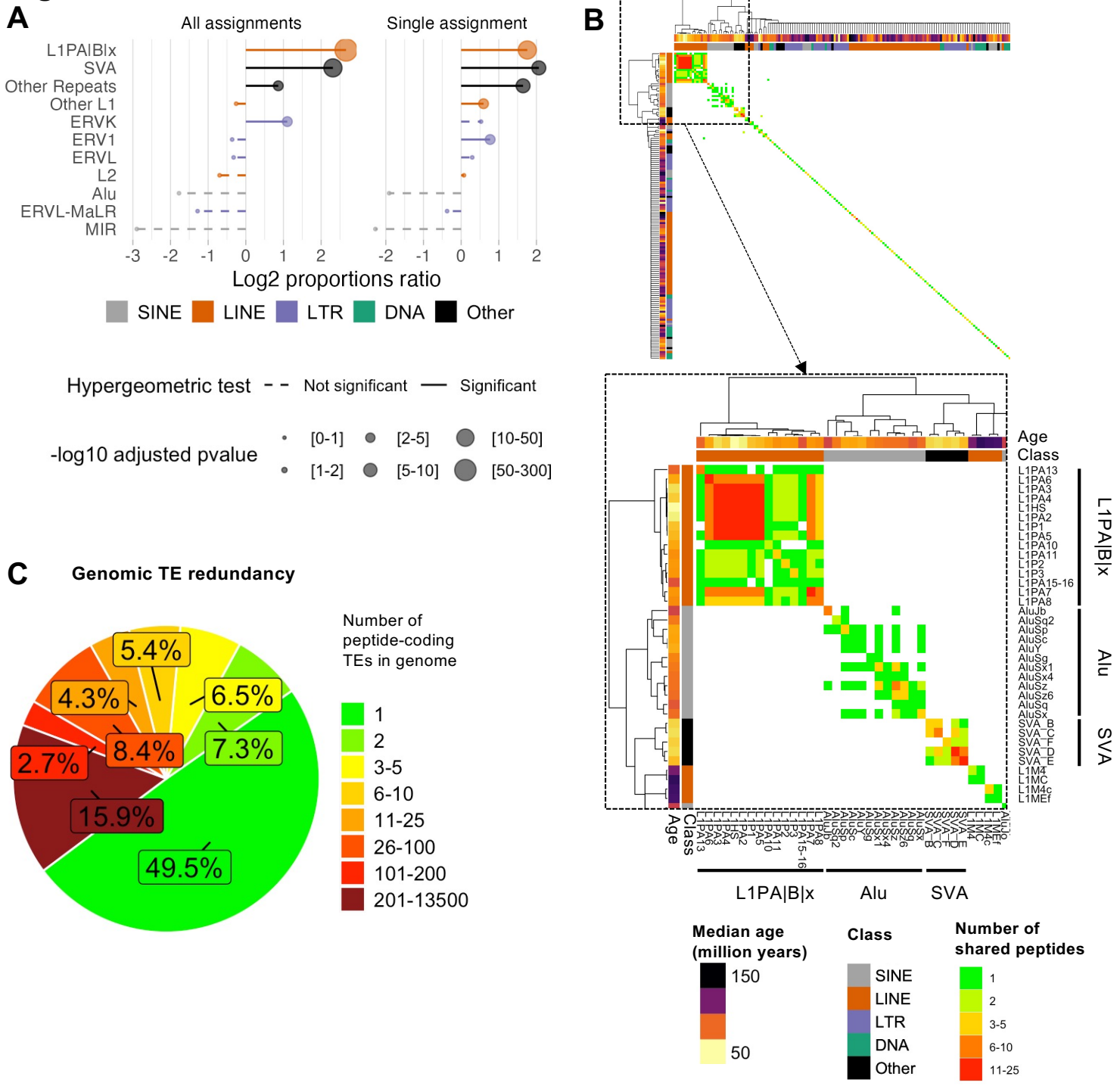
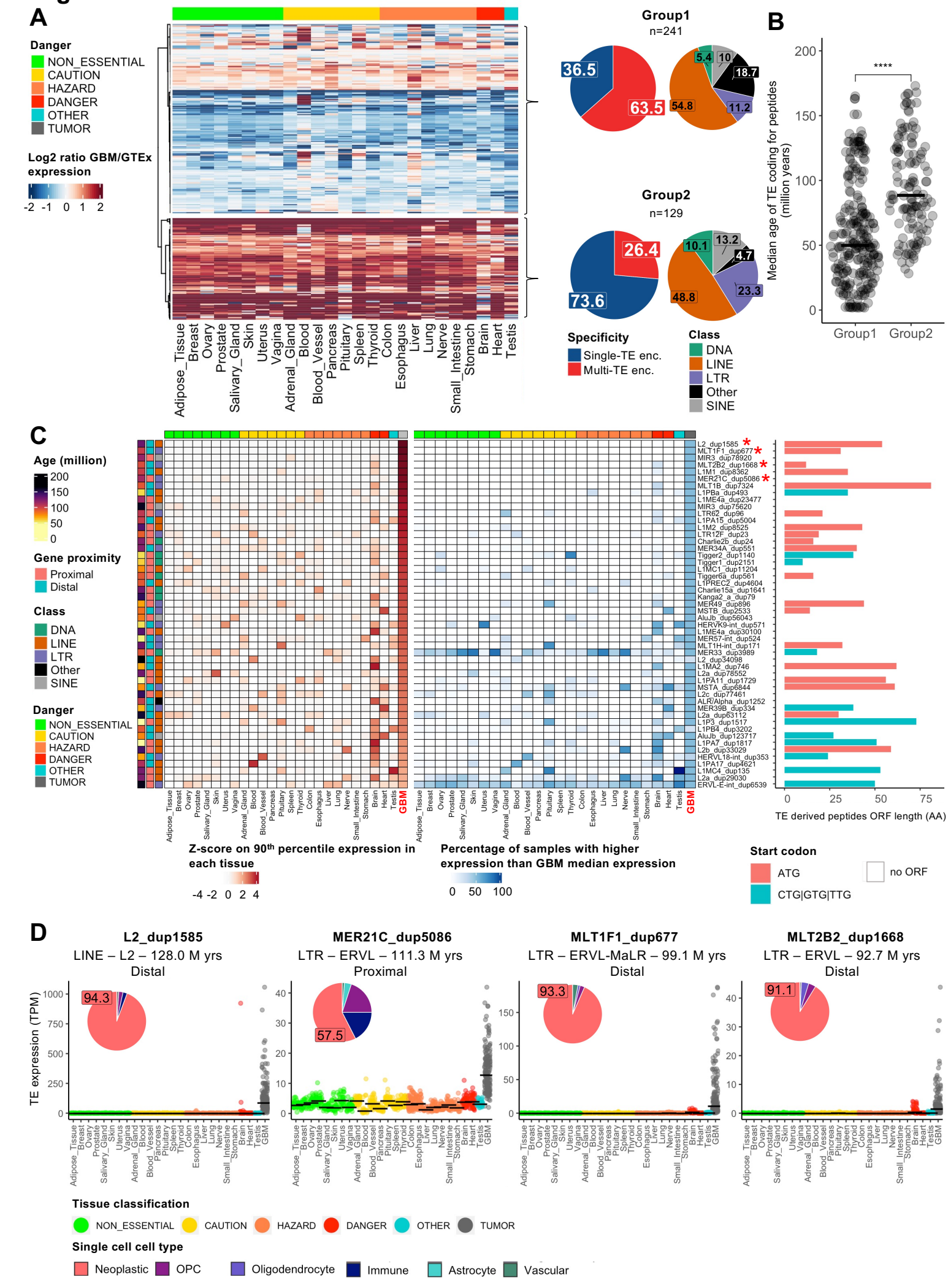

A
All assignments
Single assignment
L1PA|B|x
SVA
Other Repeats
Other L1
ERVK
ERV1
ERVL
L2
Alu
ERVL-MaLR
MIR
-3 -2 -1 0 1 2
Log2 proportions ratio
SINE LINE LTR DNA Other
Hypergeometric test
Not significant
Significant
-log10 adjusted pvalue
[0-1] [2-5] [10-50]
[1-2] [5-10] [50-300]
B
Age
Class
L1PA13
L1PA6
L1PA3
L1PA4
L1HS
L1PA2
L1P1
L1PA5
L1PA10
L1PA11
L1P2
L1P3
L1PA15-16
L1PA7
L1PA8
AluJb
AluSq2
AluSp
AluSc
AluY
AluSg
AluSx1
AluSx4
AluSz
AluSz6
AluSq
AluSx
SVA_B
SVA_C
SVA_F
SVA_D
SVA_E
L1M4
L1MC
L1M4c
L1MEf
L1PA|B|x
Alu
SVA
Median age (million years)
150
50
Class
SINE
LINE
LTR
DNA
Other
Number of shared peptides
1
2
3-5
6-10
11-25
C
Genomic TE redundancy
5.4%
6.5%
4.3%
8.4%
7.3%
2.7%
15.9%
49.5%
Number of peptide-coding TEs in genome
1
2
3-5
6-10
11-25
26-100
101-200
201-13500
D
SINE
LINE
LTR
DNA
Other
Median age (million years)
150
100
50
0
Alu
MIR
L1PA|B|x
L2
Other L1
RTE
ERV1
ERVK
ERVL
ERVL-MaLR
DNA
Other Repeats
SVA
E
Median age (million years)
150
100
50
0
R = −0.61, p < 2.2e-16
1
10
100
1000
10000
Number of TEs generating a unique peptide
Class
DNA
LINE
LTR
Other
SINE

Figure 6


A
Danger
NON_ESSENTIAL
CAUTION
HAZARD
DANGER
OTHER
TUMOR
Log2 ratio GBM/GTEx expression
-2 -1 0 1 2
Adipose_Tissue
Breast
Ovary
Prostate
Salivary_Gland
Skin
Uterus
Vagina
Adrenal_Gland
Blood
Blood_Vessel
Pancreas
Pituitary
Spleen
Thyroid
Colon
Esophagus
Liver
Lung
Nerve
Small_Intestine
Stomach
Brain
Heart
Testis
Group1
n=241
36.5
63.5
5.4
10
18.7
54.8
11.2
Group2
n=129
26.4
73.6
13.2
10.1
4.7
48.8
23.3
Specificity
Single-TE enc.
Multi-TE enc.
Class
DNA
LINE
LTR
Other
SINE
B
Median age of TE coding for peptides (million years)
200
150
100
50
0
****
Group1
Group2
C
Age (million)
200
150
100
50
0
Gene proximity
Proximal
Distal
Class
DNA
LINE
LTR
Other
SINE
Danger
NON_ESSENTIAL
CAUTION
HAZARD
DANGER
OTHER
TUMOR
GBM
L2_dup1585
MLT1F1_dup677
MIR3_dup78920
MLT2B2_dup1668
L1M1_dup8362
MER21C_dup5086
MLT1B_dup7324
L1PBa_dup493
L1ME4a_dup23477
MIR3_dup75620
LTR62_dup96
L1PA15_dup5004
L1M2_dup8525
LTR12F_dup23
Charlie2b_dup24
MER34A_dup551
Tigger2_dup1140
Tigger1_dup2151
L1MC1_dup11204
Tigger6a_dup561
L1PREC2_dup4604
Charlie15a_dup1641
Kanga2_a_dup79
MER49_dup896
MSTB_dup2533
AluJb_dup56043
HERVK9-int_dup571
L1ME4a_dup30100
MER57-int_dup524
MLT1H-int_dup171
MER33_dup3989
L2_dup34098
L1MA2_dup746
L2a_dup78552
L1PA11_dup1729
MSTA_dup6844
L2c_dup77461
ALR/Alpha_dup1252
MER39B_dup334
L2a_dup63112
L1P3_dup1517
L1PB4_dup3202
AluJb_dup123717
L1PA7_dup1817
L2b_dup33029
HERVL18-int_dup353
L1PA17_dup4621
L1MC4_dup135
L2a_dup29030
ERVL-E-int_dup6539
0 25 50 75
TE derived peptides ORF length (AA)
Z-score on 90th percentile expression in each tissue
-4 -2 0 2 4
Percentage of samples with higher expression than GBM median expression
0 50 100
Start codon
ATG
CTG|GTG|TTG
no ORF
D
L2_dup1585
LINE – L2 – 128.0 M yrs
Distal
94.3
MER21C_dup5086
LTR – ERVL – 111.3 M yrs
Proximal
57.5
MLT1F1_dup677
LTR – ERVL-MaLR – 99.1 M yrs
Distal
93.3
MLT2B2_dup1668
LTR – ERVL – 92.7 M yrs
Distal
91.1
TE expression (TPM)
Tissue classification
NON_ESSENTIAL
CAUTION
HAZARD
DANGER
OTHER
TUMOR
Single cell cell type
Neoplastic
OPC
Oligodendrocyte
Immune
Astrocyte
Vascular