Open access • Posted Content • DOI:10.1101/2021.07.28.454105

The contribution of uncharted RNA sequences to tumor identity in lung adenocarcinoma — Source link

Yunfeng Wang, Haoliang Xue, Marine Aglave, Antoine Laine ...+2 more authors

Institutions: Université Paris-Saclay

Published on: 28 Jul 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Intron, RNA, Alternative splicing and Gene

RESEARCH

The contribution of uncharted RNA sequences to tumor identity in lung adenocarcinoma

Yunfeng Wang[1,3], Haoliang Xue[1], Marine Aglave[2], Antoine Lainé[1], Mélina Gallopin[1] and Daniel Gautheret[1,2*]

*Correspondence: daniel.gautheret@universite-paris-saclay.fr
[1]Institute for Integrative Biology of the Cell (I2BC), Université Paris-Saclay, CNRS, CEA, 1 avenue de la Terrasse, 91190, Gif-sur-Yvette, France
[2]Gustave Roussy, 114 rue Edouard Vaillant, 94800, Villejuif, France
Full list of author information is available at the end of the article

Abstract

Background: Transcriptome analysis of cancer tissues has been instrumental in defining tumor subtypes, diagnostic signatures and cancer regulatory networks. Cancer transcriptomes are still predominantly analyzed at the level of gene expression. Few studies have addressed transcript-level variations, and most of these only looked at splice variants. Previously we introduced a k-mer based, reference-free method, DE-kupl, that performs differential analysis of RNA-seq data at the k-mer level, which enables distinguishing RNAs differing by a single nucleotide. Here we evaluate the significance of differential events discovered by this method in two independent lung adenocarcinoma RNA-seq datasets (N=583 and N=154).

Results: Focusing on differential events in a tumor vs normal setting, we found events in endogenous repeats, alternative splicing and polyadenylation sites, long non-coding RNAs, retained introns and unmapped RNAs. Replicability was highly significant for most event classes (assessed by comparing to events shared between unrelated tumors). Overall about 160,000 differential k-mer contigs were shared between datasets, including a large set of sequences from hypervariable genes such as immunoglobulins, $SFTP$ and mucin genes. Most interestingly, we identified a set of novel tumor-specific long non-coding RNAs in intergenic and intronic regions. We found that expressed endogenous transposons defined two major groups of patients (high/low repeat expression) with distinct clinical characteristic. A number of repeats, intronic RNAs and lincRNA achieved strong patient stratification in univariate or multivariate survival models. Finally, using antigen presentation prediction, we identified 55 contigs predicted to produce recurrent tumor-specific antigens.

Conclusions: K-mer based RNA-seq analysis enables description of cancer transcriptomes at nucleotide precision, independently of prior transcript annotation. Application to lung cancer data uncovered events stemming from a wide variety of transcriptional and postranscriptional mechanisms. Among those events, a significant subset was replicable between cohorts, thus constituting novel RNA hallmarks of cancer. The code is available at: https://github.com/Transipedia/dekupl-lung-cancer-inter-cohort.

Keywords: k-mers; contigs; repeats; LUAD; mapping-free; replicability

Background

Over a period of 20 years, cancer transcriptomics has transformed our understanding of tumor biology and led to improved tools for tumor typing and outcome prediction [1, 2]. While first generation transcriptome analysis was based on DNA microarrays with a focus on protein-coding genes, the current generation relies on RNA-seq

data, which promises to deliver a more comprehensive view of gene expression. However, in spite of its potential for transcript discovery, cancer RNA-seq data is still utilized mostly to quantify the expression of annotated genes listed in a reference transcriptome. This ignores a wide array of mRNA isoforms, non-coding RNAs, endogenous retroelements and transcripts from exogenous viruses and bacteria [3]. The quantity of information left unexploited in non-canonical transcripts remains unknown. A number of studies have started to address this question using publicly available cancer RNA-seq data, focusing on specific transcript classes such as splice variants [4, 5], lncRNAs [6], snoRNAs [7], repeats [8], bacterial RNA [9], or viral RNA [10]. Other neglected sources of RNA diversity are the so-called blacklisted regions of the genome that are too variable or repeated to be properly analyzed by conventional approaches [11]. To our knowledge, no attempt has been made to extract and evaluate at once all this non-standard RNA information from tumor RNA-seq data. We think this approach could be particularly valuable in cancer since every individual tumor harbors a unique transcriptome that departs from that of normal tissues in multiple, unpredictable ways.

Previously we introduced a computational method, DE-kupl [12], that performs differential analysis of RNA-seq data at the k-mer level. As this method is reference-free and mapping-free, it identifies any novel RNA or RNA isoform present in the data at nucleotide resolution, including poorly mapped transcripts such as RNAs from repeats and chimeric RNAs. Here we set ourselves to evaluate all non-reference events discovered by DE-kupl in a comparison of normal vs. tumor samples using lung adenocarcinoma as a test case. To mitigate false positives events inherent to any gene expression profiling [13, 14], we focused on events that were replicated in two independent datasets. This required the development of a dedicated protocol to identify shared events in unmapped RNA sequences. Results revealed a collection of novel tumor-specific unannotated lincRNAs, intron retentions, and splicing events. Most strikingly, a collection of endogenous retroelements form a major class of tumor defining transcripts and constitute potent survival signatures. We also identified a subset of events with no expression in normal tissues which could be potential neoantigens sources. We would like to suggest DE-kupl as a promising, comprehensive approach to cancer transcript profiling.

Methods

Datasets

LUAD-TCGA: 582 lung RNA-seq samples from the LUAD-TCGA project were downloaded from the dbgap repository with permission, including 524 lung adenocarcinoma (LUAD) tissues and 58 adjacent normal tissues [15]. LUAD-SEO: The LUAD RNA-seq dataset of Seo et al. [16] was downloaded from the SRA database (accession: ERP001058). This dataset contains fastq files of 87 LUAD and 77 adjacent normal tissues. Only the 77 paired normal and tumor samples were analyzed. PRAD-TCGA: For control, 557 PRAD-TCGA prostate RNA-seq datasets were downloaded from dbgap with permission, including 505 prostate adenocarcinoma (PRAD) and 52 normal controls [17]. Bam format files from the TCGA datasets were converted to fastq format using Picard tools version 2.18.16 (http://broadinstitute.github.io/picard).

DE-kupl pipeline

DE-kupl (version 5.3.0) was applied to the three datasets with the same parameters: in the filtering steps, k-mers with abundance fewer than 5 (min_recurrence_abundance) and present in no more than 10 samples (min_recurrence) were ruled out. In order to focus on non-canonical transcripts, we masked all k-mers pertaining to the main transcript of each Gencode gene as in [12]. Normalization factors for k-mer counts were computed by DE-kupl as medians of the ratios of sample counts by counts of a pseudo-reference obtained by taking the geometric mean of each k-mer across all samples. Herein we will use these counts as a proxy to represent the expression of the corresponding RNA fragment.

For differential expression analysis, the version of DESeq2 available at the time of the experiment was too slow for dealing with hundreds of samples and we found the faster "T-test" option to lack sensibility. Hence we used instead Limma [18], adapted to millions of k-mers using a chunk-based strategy (suppl. methods). This was found to perform 10 times faster than DESeq2. The performances of DESeq2, Limma and T-test for differential expression evaluation have been evaluated before [19]. Evaluations of k-mer counts were log-transformed and Limma was used to calculate log fold-changes and P-values. Retention thresholds for log2 fold changes and P-values were 1 and 0.05, respectively. All k-mers passing the filtering process above were merged into contigs and the contig table was saved as output. GC-contents in "up" and "down" contigs in the PRADtcga dataset were verified and did not present any bias (Additional file2: Table S1). High-quality contigs ("top contigs") were contigs with counts>10 in at least 15% of the smaller class (Normal or Tumor).

Gene-level expression was measured using Kallisto v0.43.02 [20] and Gencode v31 transcripts, followed by summing TPM values of transcripts from the same gene. Gene-level differential expression analysis was performed using Limma and the same normalization procedure as above. Downstream analyses were conducted using R version 3.5.2. Heatmaps were drawn using the ComplexHeatmap package (version 2.4.3) [21].

Shared event identification

Contigs from distinct DE-kupl analyses were decomposed into their constituent k-mer lists and a graph was constructed using the NetworkX Python package (version 2.3) [22], with k-mers as nodes and shared k-mers as edges. Contigs corresponding to the same local event are expected to form a fully connected subgraph or clique (Additional file 1: Fig. S1). We thus extracted all cliques to identify shared contigs. Hereafter we use the $\cap$ operator to represent contigs shared between two datasets.

Contig annotation

A uniform annotation procedure was applied to contigs from each independent analysis (LUADtcga, LUADseo, PRADtcga) and to shared contigs (LUADtcga $\cap$ LUADseo and LUADtcga $\cap$ PRADtcga). Initially, differential contigs were mapped and annotated with DE-kupl annotation (https://github.com/Transipedia/dekupl). Briefly, DE-kupl annotation maps contigs to the human genome and reports intronic, exonic or intergenic status, CIGAR string, IDs of mapped or neighboring

genes, differential usage status. A new repeat annotation field ("rep_type") was added based on Blast [23] alignments of contigs to the DFAM repeat database [24] (see Suppl. Methods). The results of DEkupl-annot were then loaded into R and submitted to further filtering and annotation. Firstly, a count filter was applied to retain only contigs with a count of 10 in at least 15% of the smaller class (Normal or Tumor). Contigs meeting this criterion were classified into event classes comprising SNV, intronic, splices, split, lincRNA, polyA, repeat and unmapped, as described in Additional file2: Table S3. Classes were non exclusive, meaning that a contig can belong to several classes. Since the TCGA datasets are unstranded, antisense events were not called. Differential usage (i.e. the relative change in expression of a local event relative to the expression of the host gene) was evaluated for each event mapped to an annotated gene. Intergenic contigs were further aligned with Blast against MiTranscriptome V2 [6] retrieved at http://mitranscriptome.org/ and converted to fasta using gffread (https://github.com/gpertea/gffread). Finally, we defined a new category called "neoRNAs", which includes contigs that are expressed in tumor tissues but silent in normal tissues.

Functional enrichment of intronic events

Candidate intronic events were identified based on the DE-kupl differential usage P-value (computed by comparing the expression or the contig with that of the host gene). Gene Ontology biological process enrichment of host genes was assessed using the clusterProfiler R package (version 3.16.0) [25].

Sample clustering based on repeats

We used the K-means algorithm [26] to cluster LUAD patients into two main subgroups based on the expression of contigs matching AluSx, L1P1_orf2 and L1P3_orf2 repeats. Clusters were then analyzed for enrichment in clinical features, immune infiltration, tumor mutational burden and copy number variants. LUAD driver genes were retrieved from the COSMIC Cancer Gene Census (CGC) list [27]. Oncoplots were drawn using the maftools R package (version 2.4.10) [28]. The estimated tumor mutational burden (TMB) for each patient was computed using the total number of non-synonymous mutations from the Mutation Annotation Format (MAF) file, divided by the estimated size of the whole exome. Copy number variation (CNV) data was downloaded by the TCGAbiolinks R package (version 2.16.3) [29], which provides a mean copy number estimate of segments covering the whole genome (inferred from Affy SNP 6.0). The ratio of gain and loss for each patient was estimated by the fraction of segments indicating CNVs. Heatmap representations were produced with ComplexHeatmap [21].

Correlation with immune infiltration

Immune infiltration analysis was performed on the LUADseo dataset. Relative proportions of infiltrating immune cells were determined using CIBERSORT [30]. Relationships between immune cell types and shared contigs (grouped by annotation category) were computed as the Spearman correlation between the contig expression and the relative proportion of the cell type in all samples. Any contig with an absolute Spearman correlation coefficient above 0.5 with at least one immune cell type was retained.

Neoantigen prediction

For prediction of recurrent tumor-specific antigen, we selected contigs absent in all normal tissues but present in at least 15% of tumor tissues. We translated contig sequences using EMBOSS transeq over 6 frames [31]. Sequences with stop codons were ruled out and candidate peptides were submitted to netMHCpan 4.0 [32] to predict binding affinity to MHC-class-I molecules. Peptide–MHC Class I interactions with strong binding levels (by default 0.5%) were reported.

Survival analysis based on event classes

Since the LUADseo dataset does not include survival information, we only performed the survival analysis on the LUADtcga dataset. Overall survival time and status was downloaded from the GDC portal (https://portal.gdc.cancer.gov/projects/TCGA-LUAD). We performed both univariate Cox regression and multivariate Cox regression on each event class to assess the prognosis value of the differential events. Survival analysis was performed using the survival (version 3.2.3) and survminer (version 0.4.7) R packages [33, 34]. Hazard ratios (HR) and P-values were calculated for each contig. Contigs with HR>1 and P-value<0.05 were considered as potential risk factors. For multivariate Cox regression, contigs were initially selected by cox-lasso regression using the glmnet R package (version 4.0.2) [35] applied independently to each contig class. The multivariate model was then constructed using selected contigs. Patients were divided into high and low-risk groups based on the median value of all risk scores for representation in Kaplan–Meier (KM) curves [36].

Unsupervised clustering analysis

We applied Principal Component Analysis (PCA) and hierarchical clustering to each event class. PCA analysis was performed with the factoextra R package (version 1.0.7) [34]. Heatmap views were obtained using ComplexHeatmap [21].

Sequence alignment views

We created "metabam" alignment files for tumor and normal tissues from each cohort. To this aim, we randomly sampled 1M reads from each fastq file of each subcohort using seqtk (https://github.com/lh3/seqtk) and aligned the aggregated reads to the genome (GRCh38) using STAR (version 2.7.0f) [37] with default parameters. BAM files were visualized using Integrative Genomics Viewer (IGV 2.6.2) [38].

Results

Gene-level *vs.* contig-level differential events

We performed tumor *vs.* normal differential expression (DE) analysis on two independent Lung adenocarcinoma RNA-seq datasets from TCGA (LUADtcga) and Seo et al. (LUADseo) and on a prostate adenocarcinoma dataset from TCGA (PRADtcga) as a control. Each dataset was submitted to a conventional, gene-level, differential expression analysis and a k-mer level differential expression analysis where all k-mers from annotated genes were first removed and the resulting differential k-mers were assembled into contigs (Fig 1A). For simplification, we shall hereafter use term "expression" when referring to either gene expression or contig

k-mer counts. While the number of DE genes in the three comparisons ranged from 6,000 to 9,000, the number of DE k-mers was about a thousand times larger (2 to 12 millions). Assembly of k-mers into contigs reduced this number to about 400,000 DE contigs in each analysis (Fig 1B).

We next compared the DE genes and contigs discovered in independent datasets to identify shared DE events. While this process is trivial for genes, it is not for contigs, since contigs found in each dataset have no standard identifier that could be used to relate them. We thus implemented a graph analysis procedure that identified shared contigs based on their common k-mers (Fig 1A, Additional file 1: Fig. S1). A final annotation step assigned contigs to non exclusive categories based on their mapping characteristics or expression (repeats, lincRNAs, splice variant, polyadenylation variants, split RNAs, tumor-specific RNAs) as described in Additional file2: Table S3 and Methods. The numbers of shared elements slightly differ between LUADtcga and LUADseo because a minority of elements are in a 2-to-1 or 1-to-2 relationship in the contig graph. If not otherwise specified, numbers of elements are given for the LUADtcga cohort.

Overall 160,610 differential contigs were shared between the two LUAD analyses (Fig 1C). Over these, 120,822 contigs were considered of sufficient quality based on counts and occurrence in a minimal number of samples (see Methods). 83% of shared contigs were overexpressed in tumors vs. only 17% underexpressed (Fig 1C).

Event replicability

The replicability of differential events was generally lower for k-mer or contigs than for genes. Fig 1D shows the number of differential expression genes and contigs shared by the two independent LUAD analyzes, with contigs binned by annotation class. About 41% of differential expression genes (3032 genes) were shared by the two LUAD analyses, compared to an average of 14% for differential expression contigs (repeats: 3.7%, unmapped RNAs: 10%, alternative polyAs: 13%, lincRNAs: 14%, alternative splices: 20%, retained introns: 20%). Although the ratio of shared events was relatively low for k-mer analysis, it was considerably higher than when comparing two unrelated pathologies (LUADtcga $\cap$ PRADtcga, Fig 1D), and this applied to all event classes except repeats. This indicates that, although k-mer based differential expression events are noisy, a significant subset is replicable in independent studies. Furthermore, we observed a strong correlation between the fold-change value of differential expression contigs and the likelihood to be shared between cohorts (Additional file 1: Fig. S2), demonstrating the non-randomness of high scoring, non-reference events.

DE contig localization, hypervariable genes

The majority of shared contigs are genic (83%), 45% are intronic and 32% carry SNVs or indels (Fig 2A). These characteristics are induced by the initial filter that removed all k-mers matching reference transcripts, retaining any intronic or SNV-carrying k-mer. Therefore a large number of SNV and intronic contigs are just "passenger" events of DE genes. We confirmed this by analyzing the correlation between numbers of DE contigs and host gene expression. We found a significant correlation (Pearson CC=0.45), but this correlation was reduced (Pearson CC=0.28) in shared

DE contigs, indicating shared contigs contain fewer passenger events (Additional file 3).

More than 400 genes were matched by 35 or more contigs. We classified these genes into two categories: for 296 genes, most contigs matched introns and were up-regulated in tumors (Fig 2A, B, Additional file 2: Table S5). These mostly correspond to the aforementioned passenger events. The second category is composed of 107 genes we refer to as "hypervariable" as they tend to yield a large number of contigs carrying SNVs, indels and larger rearrangements (Fig 2A, C, Additional file 2: Table S5). The largest sets of hypervariable genes are IGK, IGL and IGH immunoglobulin genes. This is not surprising given immunoglobulins (i) are highly variable due to V(D)J segment recombination and (ii) are expressed by plasma B-cells which are abundant in the tumor immune infiltrate [39], hence these genes are seen as up-regulated in tumors. Interestingly, those IG sequence variants are found expressed in different patients and across the two cohorts, suggesting our approach can be used to profile immunoglobulin repertoires, as performed recently with other RNA-seq datasets [40]. To evaluate the accuracy of DE-kupl contigs assembled from IG genes, we selected all contigs mapped to one arbitrary IG gene (IGHV: 100 contigs) and aligned them to IGHV contigs from the IMGT database [41]. Ninety out of 100 contigs had significant matches in the corresponding IMGT category extending over 90% of the contig length (Additional file2: Table S6).

Other hypervariable loci were found in surfactant protein ($SFTP$) and Mucin genes which are known to harbor a high level of polymorphism [42, 43]. We observed polymorphism not only in the form of SNPs, but also in the form of splicing variations. Five $SFTP$ genes alone combine over 9000 SNVs and 800 splice sites contigs, while 12 Mucin genes harbour 1324 contigs including 42 splice variants (Additional file 1: Fig. S3A-B, Additional file 2: Table S5). While $SFTP$ contigs were all underexpressed in tumors, Mucin contigs were mostly overexpressed (Additional file 2: Table S5). Mucins are immunogenic [43] and are important biomarkers for prognosis [44] and drug resistance [45]. The existence of recurrent mucin variants overexpressed in tumors may be relevant for these therapeutic and biomarker developments. We also observed hypervariability in $CEACAM5$ and $KR19$, two other prognostic biomarkers and/or immunotherapy targets [46, 47] (Additional file 1: Fig. S3C, Additional file 2: Table S5).

Intron retention and other intronic events

We found intronic contigs with differential usage (DU) in 313 host genes, 290 (93%) of which were up-regulated in tumors (Additional file 2: Table S4). 70% of the host genes were also up-regulated, thus the apparent overexpression of these intronic sequences may have been confounded by overexpression of host genes. However, 30% of host genes were not overexpressed, and in 103 cases, intron and host gene expressions varied in opposite directions (93 introns up and 10 introns down). Our annotation pipeline did not differentiate intron retentions (as shown for example in Additional file 1: Fig. S4A) from transcription units occurring within introns (example in Additional file 1: Fig. S4B). We observed intron retention events in lung cancer drivers $EGFR$ and MET (Additional file 1: Fig. S4C and Additional file 1: Fig. S4D). In $EGFR$, the retained intron was located between exons 18 and

19, just upstream of the principal oncogenic *EGFR* mutations located in exons 19-21. Intron retention before exon 19 would likely produce a truncated form of *EGFR* compatible with oncogenic activation.

Additional file 1: Fig. S5A shows the 20 intronic events with the most significant differential usage P-values. All show opposite directions of intron and gene expression. Gene Ontology enrichment analysis indicates host genes are enriched for inflammation and immune response pathways involving neutrophil and T cells (additional file 1: Fig. S5B), suggesting these events may come from regulations in the tumor microenvironment rather than in the tumor itself.

Novel lincRNAs

Contigs that do not map any Gencode annotated gene are of particular interest as they potentially represent novel lincRNA biomarkers of lung tumors. Overall we identified shared DE contigs in 885 intergenic regions, which we labelled as lincRNAs. As genic regions already included annotated lncRNAs and pseudogenes from Gencode, the actual number of DE contigs in lncRNAs and pseudogenes was much higher (N=2892) but we focus here on unannotated regions. lincRNA contigs were mostly overexpressed in tumors (83% of contigs) and often contained a known repeat element (73% of contigs). Their average length was 137 nt, however actual transcription units were generally longer as most units were composed of multiple contigs, as shown in examples in Additional file 1: Fig. S6. Most intergenic contigs (793 out of 823) were already annotated in the independent Mitranscriptome lncRNA database [6], which was expected since this database was also produced from TCGA RNA-seq data. Less than one third of the flanking genes of intergenic contigs were differentially expressed, indicating that novel lincRNA expression was most often independent from that of flanking genes.

Expressed repeats delineate patient subgroups with distinct clinical properties

The dominant model for endogenous retroelements (EREs) expression is that EREs are mainly expressed in germline and embryonic stem cells while they are repressed in differentiated somatic cells. However recent studies have shown expression of EREs in somatic cells is more common and heterogeneous than expected[48]. Repeat-containing reads are difficult to analyze by RNA-seq standard pipelines due to ambiguity in the alignment process. We thus questioned whether our alignment-free procedure could help reveal these events. From the initial set of 50572 contigs annotated as repeats (Fig 1C), we selected a high quality subset of 10341 contigs over 60 bp in size and with expression above a set threshold (see Methods). Of these, 87.7% were overexpressed in tumors (Additional file 2: Table S4).

Fig 3A shows the distribution of contigs per repeat family. Most repeats correspond to Line 1 and Alu family sequences. The most frequent repeat overall is L1P1, a Line 1 of the L1Hs family which is the only retrotransposition-competent EREs in the human genome [49]. L1P1/L1Hs elements, as well as human endogenous retrovirus (HERV), were almost exclusively over-expressed in tumors, suggesting tumor-specific activation of these elements. In contrast, Alu elements, which are often expressed as part of protein coding genes, were either over- or under-expressed in

tumors. Fig 3A shows the top 20 repeat types that contribute more contigs. Fig 3B-C shows the expression heatmap of the 60 repeats contributing more contigs. For each type of repeats, we selected the contig with the highest absolute fold-change.

Repeat contigs also included a group annotated as "simple repeats", containing microsatellites and other low complexity elements. Contrarily to EREs, these do not have the capacity to be expressed independently. Indeed, in over 70% of cases, these contigs were uniquely mapped to genic sequences. In addition to annotated repeats and simple repeats, DE-kupl identified 4762 contigs (4497 up, 265 down) with multiple genome hits but no match in the DFAM repeat database (Additional file 2: Table S4). Many of these repeats were from Mucins, immunoglobulins and multicopy gene families such as $NBPF$ and $TBC1$. These repeats are shared between two cohorts and thus represent robust events of (mostly) overexpressed RNA fragments in tumors that would hardly be noticed in regular RNA-seq analysis due to their low mappability.

To investigate repeat-based patient subgroups, we performed clustering of tumors based on the most frequent repeat elements in Fig 3A: AluSx, L1P1_orf2, and L1P3_orf2 (as FLAM repeats are a family of Alu-like monomers that give birth to the left arms of the Alu elements, we did not account for FLAM_C_1_143). K-means clustering with k varying from 2 to 4 groups consistently found two major subgroups: subgroup 1 ("repeat-low") displayed generally low expression of Alu and L1 repeats compared to subgroup 2 ("repeat-high") (Fig 4A).

We then related the two repeat subgroups with somatic alterations observed in TCGA patients. Patients in the repeat-high group were more frequently mutated in LUAD drivers $CSMD3$, $TP53$, $PTPRD$, $PTPRT$, $GRIN2A$, $EPHA3$, and $MB21D2$ (Fig 4B, Fisher P<0.05). Patients in the repeat-high group had a significantly higher TMB (Wilcoxon P=1.5e-07) and a higher ratio of CNVs than other patients (Wilcoxon P=5.5e-05 for gain; P=0.019 for loss) (Fig 4C).

We observed no difference between subgroups in terms of age, gender, tumor stage, overall survival (OS), and vital status, but found more smokers in the repeat-high group (Wilcoxon P=0.02). We then assessed the immune cell contents of samples estimated by gene expression deconvolution. The repeat-high subgroup had lower proportions of dendritic cells, M2 macrophages, mast cells, monocytes and CD4+ T cells and overall immune content than the repeat-low subgroup (Fig 4D). In summary, "repeat-high" tumors associate with higher genome instability, more frequent smoking and lower immune infiltration.

Immune cell-associated contigs

We sought which contigs best correlated with tumor immune cell contents estimated by gene expression deconvolution. Sixty five contigs were found correlated with at least one type of immune cell (Additional file 1: Fig. S7). Most of these were uniquely mapped to genic introns or exons and underexpressed in tumors. Positive correlations were mostly observed with M2/M0 macrophages or resting CD4+ T cells, *i.e.* with a generally repressive or quiescent immune environment. However, a few contigs were associated to immune active M1 macrophages, including two contigs matching $GBP5$ (a marker of activated macrophages) and $CXCR2P1$ (a pseudogene expressed in an intron of $RUFY4$, a gene expressed in dendritic cells).

Overall, immune cell-associated contigs mapped leukocyte-specific or immunity-related genes, suggesting most contigs originated from the immune cell themselves (Additional file 2: Table S11).

Perhaps the most intriguing set of immune cell-associated contigs was that correlated to naive CD4+ T-cells. These cells are not especially enriched in tumor or normal samples, yet they correlate with six DE contigs. One contig was strongly repressed in tumors and corresponded to *Klebsiella pneumoniae* large subunit rRNA. Indeed, *Klebsiella* is a common lung bacterium against which cross-reactive T-cells are present in the naive CD4+ T-cell repertoire [50]. Our results thus suggest the joint occurrence of *Klebsiella* and matching CD4+ T-cell in normal lungs, and their disappearance in tumors. Of note, this *Klebsiella* contig also correlates positively with multiple contigs in the *SFTP* gene (Additional file 2: Table S12), in line with *SFTP* roles in defense against respiratory pathogens [51].

The other five contigs associated with naive CD4+ T-cells were all overexpressed in tumors. These included two intergenic repeats related to HERV (human endogenous retrovirus): HERV-E and MER9. The HERV-E contig was expressed from the *env* gene of a near full-length retroelement. One may hypothesize that expression and antigen production by the *env* gene trigger recruitment of CD4+ T-cells, as observed already in breast cancer [52]. Alternatively, reactivation of HERV elements could be an intrinsic feature of the CD4+ T-cells [53]. This analysis illustrates how non-reference RNA quantification can illuminate the interplay between cell types and specific RNA elements including exogenous elements in a bulk tissue.

Novel sources of shared neoantigens enriched in lincRNAs

Tumors express a large diversity of transcripts that are not usually expressed in normal tissues. When translated, these transcripts can produce peptides recognized as non-self by the epitope presentation machinery, triggering antitumor immune response [54]. These tumor-specific antigens or neoantigens are the object of active investigation for immunotherapy and tumor vaccine development. Protocols for neoantigen discovery usually start from a list of nonsynonymous somatic mutations identified from WES or WGS libraries and whose expression is confirmed by RNA-seq. Candidate mutated peptides are then submitted to an epitope presentation prediction pipeline [55]. This protocol predicts potential neoantigens from annotated and mappable regions. However, neoantigens can be produced from any transcript, including repeats and supposedly non-coding lncRNAs [56, 57]. Therefore we thought our reference-free approach could be a good source for such elements.

We considered contigs with no expression in normal tissues as potential neoantigen sources. To focus on shared neoantigens, we further requested contigs to be expressed in at least 15% of tumor samples. This selected 2375 contigs in the LUADtcga dataset (Fig 5.A). About 20% of these contigs (N=472) where also silent in normal tissues of the LUADseo cohort (Fig 5.B). We evaluated the potential of these "strictly tumoral" contigs for neoantigen presentation. Fifty five strictly tumoral contigs produced peptides predicted to be strong MHC-class-I binders by netMHCpan (Additional file 2: Table S10). Although potential neoantigen-producing contigs were found in several categories and locations, intergenic location was the most

significantly enriched category (Additional file 1: Fig. S8). Overall, contigs from intergenic regions, non-coding RNAs and pseudogenes contributed 58% of predicted neoantigens (Additional file 2: Table S10), consistent with previous reports of abundant neoantigen production from non-coding regions in other cancers [57].

Repeats, intronic RNAs and lincRNA as survival predictors

To identify RNA elements associated with outcome, we retrieved overall survival (OS) data for the TCGA cohort and performed univariate Cox regression with the different classes of contigs. Thirty nine contigs were significantly related to OS after multiple testing correction (Additional file 2: Table S7). Outcome-related contigs are mostly enriched in repeats (Additional file 2: Table S8), especially HERV elements (4 out of the 10 top repeats) and Alu/L1 family elements (AluSx and L1P3_orf2). While HERV elements expression was always negatively related to OS, the trend for other repeats was variable, with different Line1 and Alu elements having either positive or negative relation to OS (Additional file 2: Table S7). Another interesting OS-related element was a novel splice variant in ELF1, a transcription factor of the ETS family involved in multiple cancers (Additional file 2: Table S7)[58].

We then performed multivariate Cox regression using sets of contigs selected by lasso regression within each contig category and using differentially expressed genes (Additional file 2: Table S9). Models based on annotated and simple repeats had the best prognostic power (log-rank P=2e-16, 2e-13, respectively, Fig 6). The "annotated repeat" model was based on 12 contigs, including six L1 and three HERV elements, reinforcing the relevance of these repeats for prognosis. The "simple repeat" model included 12 contigs with microsatellite-like repeats, of which 11 were uniquely mapped to the genome (Additional file2: Table S9). Other strong outcome predictors were obtained using lincRNA, intronic and unmapped contigs, all of which achieved a better patient stratification than a model based on DE genes (Fig 6).

Unsupervised sample clustering based on non-reference RNAs

To investigate the capacity of non-reference RNAs to distinguish tumor and normal tissues in an unsupervised fashion, we performed PCA clustering of samples using contigs from each class (Fig 7). Tumor and normal tissues can be distinguished based on SNV, splice, intron, and lincRNA event classes as clearly as based on differentially expressed genes ("DEG" in Fig 7). This capacity is consistently observed in both cohorts. However, while many repeats are important with respect to tumor subclasses and survival, repeats altogether do not permit a clear separation of tumor and normal tissues in unsupervised clustering. Classes "polyA", "split" and "unmapped" did not achieve clear separation either, which was more expected as these sets were much smaller in size.

Discussion

Using reference-free analysis of LUAD RNA-seq data, we identified a large set of differential RNA elements that were present in two independent LUAD cohorts. We classified these elements based on their genomic location, mapping characteristics and repeat contents. We did not analyze in detail all contig classes but focused

instead on contigs mapping to hypervariable genes, repeats, lincRNAs and intronic elements. Besides these, a number of splice variants, chimeras, exogenous (non-human) sequences were found differentially expressed and could be pursued further.

A defining class of differential events involved endogenous repeats. The expression of L1 and Alu repeats defined two major tumor subgroups. The subgroup with higher L1/Alu expression was associated with more frequent mutations in $P53$, a higher mutational and copy number burden and a reduced immune cell infiltrate. This is consistent with previous observations that retrotransposition events can be controlled by $P53$ [59], correlate with a repressed immune environment [59, 60] and can lead to genome instability [61]. Expressed repeats also had significant prognostic power. Multivariate signatures composed of HERV and L1 elements, or simple repeats, stratified patients into distinct survival groups. Of note, HERV expression has been sporadically involved in various cancer types [62] and has recently been associated with poor prognosis in colorectal cancer [63].

A limitation of k-mer approaches for TE analysis is that transcripts are not fully assembled and thus the nature of repeats, whether expressed as functional retroelements or as part of mRNA or lncRNAs cannot be systematically established. Nonetheless, the majority of DE contigs are long enough to enable unambiguous mapping on the human genome, hence their origin could be further explored, including when coming from novel insertion events.

An attractive aspect of reference-free RNA-seq analysis is the capacity to identify novel forms of known cancer drivers or biomarkers. Indeed, we identified novel intron retention events in $EGFR$ and MET and multiple new variants of $CEACAM5$ and $KR19$. Perhaps even more interesting is the ability to detect potential neoantigen sources in variant transcripts. Tumor-specific neoantigens have previously been identified from repeats and non-coding regions using mapping-based strategies [54, 57]. However, our approach casts a wider net as it collects all events independently of their origin, including when arising from unmappable or profoundly rearranged regions. Indeed we identified about 500 strictly tumoral contigs shared by patients from the two independent cohorts, 55 of which were predicted to produce MHC-class-I neoantigens. These shared neoantigen candidates are of particular interest since their targeting by antitumor therapy would potentially benefit groups of multiple patients.

The wealth of information uncovered in the present study is a strong incentive to explore other applications of reference-free transcriptomics. One such application is the identification of patient-specific abnormal transcripts under a 1 *vs* n experimental design, which is addressed by the Mintie software [64]. Reference-free strategies can also be used for building predictive models. We [65] and others [66, 67] are exploring this kind of approach to classify cancer RNA-seq samples with promising results. Finally, reference-free differential analysis of the type used in this study could be of particular interest in meta-transcriptomics projects where RNAs are sequenced from an environment containing unknown bacterial, archaeal or eukaryotic species. Our protocol guarantees that any RNA that is specific to a sample subset will be captured independently of its origin. We hope the present analysis will encourage others to explore other data sources in a reference-free manner.

List of abbreviations
SNV: Single-Nucleotide Variants
CNV: Copy Number Variant
SV: Structural Variant
AS: Alternative Splicing
TCGA: The Cancer Genome Atlas
LUAD: Lung Adenocarcinoma
PRAD: Prostate Adenocarcinoma
EREs: endogenous retroelements

Declarations
Ethics approval and consent to participate
Not applicable

Consent to publish
Not applicable.

Availability of data and materials
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Funding
This work was funded in part by Agence Nationale de la Recherche grant ANR-18-CE45-0020 and by a PhD studentship to YW by Annoroad Gene Technology, Beijing.

Authors' contributions
YW and DG designed the workflow and analyzed the results, YW downloaded and processed the datasets, YW and DG wrote the manuscript, MA and MG assisted in statistical analysis, HX assisted in coding scripts. AL annotated the repeat types.

Acknowledgements
The results shown in this work are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

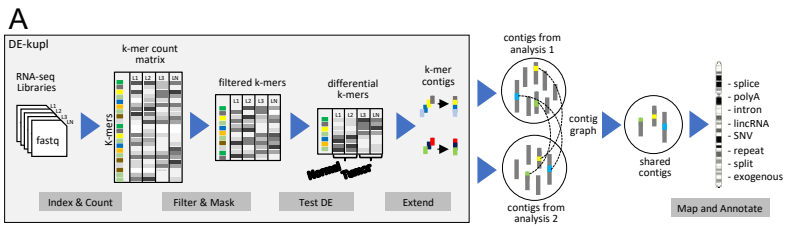
Author details
[1]Institute for Integrative Biology of the Cell (I2BC), Université Paris-Saclay, CNRS, CEA, 1 avenue de la Terrasse, 91190, Gif-sur-Yvette, France. [2]Gustave Roussy, 114 rue Edouard Vaillant, 94800, Villejuif, France. [3]Annoroad Gene Technology Co., Ltd,, 100176, Beijing, China.

References
1. Gollub, M.J., Prowda, J.C.: Primary melanoma of the esophagus: radiologic and clinical findings in six patients. Radiology **213**(1), 97–100 (1999)
2. Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., *et al.*: Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology **27**(8), 1160 (2009)
3. Morillon, A., Gautheret, D.: Bridging the gap between reference and real transcriptomes. Genome biology **20**(1), 1–7 (2019)
4. Kahles, A., Lehmann, K.-V., Toussaint, N.C., Hüser, M., Stark, S.G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., Caesar-Johnson, S.J., *et al.*: Comprehensive analysis of alternative splicing across tumors from 8,705 patients. Cancer cell **34**(2), 211–224 (2018)
5. Vitting-Seerup, K., Sandelin, A.: Isoformswitchanalyzer: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. Bioinformatics **35**(21), 4469–4471 (2019)
6. Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., *et al.*: The landscape of long noncoding rnas in the human transcriptome. Nature genetics **47**(3), 199–208 (2015)
7. Gong, J., Li, Y., Liu, C.-j., Xiang, Y., Li, C., Ye, Y., Zhang, Z., Hawke, D.H., Park, P.K., Diao, L., *et al.*: A pan-cancer analysis of the expression and clinical relevance of small nucleolar rnas in human cancer. Cell reports **21**(7), 1968–1981 (2017)
8. Solovyov, A., Vabret, N., Arora, K.S., Snyder, A., Funt, S.A., Bajorin, D.F., Rosenberg, J.E., Bhardwaj, N., Ting, D.T., Greenbaum, B.D.: Global cancer transcriptome quantifies repeat element polarization between immunotherapy responsive and t cell suppressive classes. Cell reports **23**(2), 512–521 (2018)
9. Ouchenir, L., Renaud, C., Khan, S., Bitnun, A., Boisvert, A.-A., McDonald, J., Bowes, J., Brophy, J., Barton, M., Ting, J., et al.: The epidemiology, management, and outcomes of bacterial meningitis in infants. Pediatrics **140**(1) (2017)
10. Zapatka, M., Borozan, I., Brewer, D.S., Iskar, M., Grundhoff, A., Alawi, M., Desai, N., Sültmann, H., Moch, H., Cooper, C.S., *et al.*: The landscape of viral associations in human cancers. Nature genetics **52**(3), 320–330 (2020)

11. Amemiya, H.M., Kundaje, A., Boyle, A.P.: The encode blacklist: identification of problematic regions of the genome. Scientific reports **9**(1), 1–5 (2019)
12. Audoux, J., Philippe, N., Chikhi, R., Salson, M., Gallopin, M., Gabriel, M., Le Coz, J., Drouineau, E., Commes, T., Gautheret, D.: De-kupl: exhaustive capture of biological variation in rna-seq data through k-mer decomposition. Genome biology **18**(1), 1–15 (2017)
13. Ioannidis, J.P.: Microarrays and molecular research: noise discovery? Lancet (London, England) **365**(9458), 454–455 (2005)
14. Michiels, S., Koscielny, S., Boulet, T., Hill, C.: Gene expression profiling in cancer research. Bulletin du cancer **94**(11), 976–980 (2007)
15. Network, C.G.A.R., *et al.*: Comprehensive molecular profiling of lung adenocarcinoma. Nature **511**(7511), 543 (2014)
16. Seo, J.-S., Ju, Y.S., Lee, W.-C., Shin, J.-Y., Lee, J.K., Bleazard, T., Lee, J., Jung, Y.J., Kim, J.-O., Shin, J.-Y., *et al.*: The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome research **22**(11), 2109–2119 (2012)
17. Abeshouse, A., Ahn, J., Akbani, R., Ally, A., Amin, S., Andry, C.D., Annala, M., Aprikian, A., Armenia, J., Arora, A., *et al.*: The molecular taxonomy of primary prostate cancer. Cell **163**(4), 1011–1025 (2015)
18. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K.: limma powers differential expression analyses for rna-sequencing and microarray studies. Nucleic acids research **43**(7), 47–47 (2015)
19. De Paepe, K.: Comparison of methods for differential gene expression using rna-seq data (2015)
20. Bray, N.L., Pimentel, H., Melsted, P., Pachter, L.: Near-optimal probabilistic rna-seq quantification. Nature biotechnology **34**(5), 525–527 (2016)
21. Gu, Z., Eils, R., Schlesner, M.: Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics **32**(18), 2847–2849 (2016)
22. Hagberg, A., Swart, P., S Chult, D.: Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States) (2008)
23. Madden, T.: The blast sequence analysis tool. In: The NCBI Handbook [Internet]. 2nd Edition. National Center for Biotechnology Information (US), ??? (2013)
24. Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F., Wheeler, T.J.: The dfam database of repetitive dna families. Nucleic acids research **44**(D1), 81–89 (2016)
25. Yu, G., Wang, L.-G., Han, Y., He, Q.-Y.: clusterprofiler: an r package for comparing biological themes among gene clusters. Omics: a journal of integrative biology **16**(5), 284–287 (2012)
26. MacQueen, J., *et al.*: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967). Oakland, CA, USA
27. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., Forbes, S.A.: The cosmic cancer gene census: describing genetic dysfunction across all human cancers. Nature Reviews Cancer **18**(11), 696–705 (2018)
28. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., Koeffler, H.P.: Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome research **28**(11), 1747–1756 (2018)
29. Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I., *et al.*: Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. Nucleic acids research **44**(8), 71–71 (2016)
30. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., Alizadeh, A.A.: Robust enumeration of cell subsets from tissue expression profiles. Nature methods **12**(5), 453–457 (2015)
31. Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R., Potter, S.C., Finn, R.D., *et al.*: The embl-ebi search and sequence analysis tools apis in 2019. Nucleic acids research **47**(W1), 636–641 (2019)
32. Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., Nielsen, M.: Netmhcpan-4.0: improved peptide–mhc class i interaction predictions integrating eluted ligand and peptide binding affinity data. The Journal of Immunology **199**(9), 3360–3368 (2017)
33. Therneau, T.M., Lumley, T.: Package 'survival'. R Top Doc **128**(10), 28–33 (2015)
34. Kassambara, A., Kosinski, M., Biecek, P., Fabian, S.: Package 'survminer'. Drawing Survival Curves using 'ggplot2'.(R package version 0.3. 1.) (2017)
35. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of statistical software **33**(1), 1 (2010)
36. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American statistical association **53**(282), 457–481 (1958)
37. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: Star: ultrafast universal rna-seq aligner. Bioinformatics **29**(1), 15–21 (2013)
38. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P.: Integrative genomics viewer. Nature biotechnology **29**(1), 24–26 (2011)
39. Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Yang, T.-H.O., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., *et al.*: The immune landscape of cancer. Immunity **48**(4), 812–830 (2018)
40. Mandric, I., Rotman, J., Yang, H.T., Strauli, N., Montoya, D.J., Van Der Wey, W., Ronas, J.R., Statz, B., Yao, D., Petrova, V., *et al.*: Profiling immunoglobulin repertoires across multiple human tissues using rna sequencing. Nature communications **11**(1), 1–14 (2020)
41. Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., *et al.*: Imgt®, the international immunogenetics information system®. Nucleic acids research **37**(suppl_1), 1006–1012 (2009)
42. Imielinski, M., Guo, G., Meyerson, M.: Insertions and deletions target lineage-defining genes in human cancers. Cell **168**(3), 460–472 (2017)
43. Swallow, D.M., Gendler, S., Griffiths, B., Corney, G., Taylor-Papadimitriou, J., Bramwell, M.E.: The human

tumour-associated epithelial mucins are coded by an expressed hypervariable gene locus pum. Nature **328**(6125), 82–84 (1987)
44. Ning, Y., Zheng, H., Zhan, Y., Liu, S., Zang, H., Luo, J., Wen, Q., Fan, S., *et al.*: Comprehensive analysis of the mechanism and treatment significance of mucins in lung cancer. Journal of Experimental & Clinical Cancer Research **39**(1), 1–10 (2020)
45. Aithal, A., Rauth, S., Kshirsagar, P., Shah, A., Lakshmanan, I., Junker, W.M., Jain, M., Ponnusamy, M.P., Batra, S.K.: Muc16 as a novel target for cancer therapy. Expert opinion on therapeutic targets **22**(8), 675–686 (2018)
46. Wang, X.-M., Zhang, Z., Pan, L.-H., Cao, X.-C., Xiao, C.: Krt19 and ceacam5 mrna-marked circulated tumor cells indicate unfavorable prognosis of breast cancer patients. Breast cancer research and treatment **174**(2), 375–385 (2019)
47. Thistlethwaite, F.C., Gilham, D.E., Guest, R.D., Rothwell, D.G., Pillai, M., Burt, D.J., Byatte, A.J., Kirillova, N., Valle, J.W., Sharma, S.K., *et al.*: The clinical efficacy of first-generation carcinoembryonic antigen (ceacam5)-specific car t cells is limited by poor persistence and transient pre-conditioning-dependent respiratory toxicity. Cancer Immunology, Immunotherapy **66**(11), 1425–1436 (2017)
48. Larouche, J.-D., Trofimov, A., Hesnard, L., Ehx, G., Zhao, Q., Vincent, K., Durette, C., Gendron, P., Laverdure, J.-P., Bonneil, É., *et al.*: Widespread and tissue-specific expression of endogenous retroelements in human somatic tissues. Genome medicine **12**, 1–16 (2020)
49. Rangwala, S.H., Zhang, L., Kazazian, H.H.: Many line1 elements contribute to the transcriptome of human somatic cells. Genome biology **10**(9), 1–18 (2009)
50. Cassotta, A., Goldstein, J.D., Durini, G., Jarrossay, D., Baggi Menozzi, F., Venditti, M., Russo, A., Falcone, M., Lanzavecchia, A., Gagliardi, M.C., *et al.*: Broadly reactive human cd4+ t cells against enterobacteriaceae are found in the naïve repertoire and are clonally expanded in the memory repertoire. European journal of immunology **51**(3), 648–661 (2021)
51. Wright, J.R.: Host defense functions of pulmonary surfactant. Neonatology **85**(4), 326–332 (2004)
52. Wang-Johanning, F., Radvanyi, L., Rycaj, K., Plummer, J.B., Yan, P., Sastry, K.J., Piyathilake, C.J., Hunt, K.K., Johanning, G.L.: Human endogenous retrovirus k triggers an antigen-specific immune response in breast cancer patients. Cancer research **68**(14), 5869–5877 (2008)
53. White, C.H., Beliakova-Bethell, N., Lada, S.M., Breen, M.S., Hurst, T.P., Spina, C.A., Richman, D.D., Frater, J., Magiorkinis, G., Woelk, C.H.: Transcriptional modulation of human endogenous retroviruses in primary cd4+ t cells following vorinostat treatment. Frontiers in immunology **9**, 603 (2018)
54. Smith, C.C., Selitsky, S.R., Chai, S., Armistead, P.M., Vincent, B.G., Serody, J.S.: Alternative tumour-specific antigens. Nature Reviews Cancer **19**(8), 465–478 (2019)
55. Gopanenko, A.V., Kosobokova, E.N., Kosorukov, V.S.: Main strategies for the identification of neoantigens. Cancers **12**(10), 2879 (2020)
56. Ouspenskaia, T., Law, T., Clauser, K.R., Klaeger, S., Sarkizova, S., Aguet, F., Li, B., Christian, E., Knisbacher, B.A., Le, P.M., et al.: Thousands of novel unannotated proteins expand the mhc i immunopeptidome in cancer. bioRxiv (2020)
57. Laumont, C.M., Vincent, K., Hesnard, L., Audemard, É., Bonneil, É., Laverdure, J.-P., Gendron, P., Courcelles, M., Hardy, M.-P., Côté, C., et al.: Noncoding regions are the main source of targetable tumor-specific antigens. Science translational medicine **10**(470) (2018)
58. Sizemore, G.M., Pitarresi, J.R., Balakrishnan, S., Ostrowski, M.C.: The ets family of oncogenic transcription factors in solid tumours. Nature Reviews Cancer **17**(6), 337–351 (2017)
59. Jung, H., Choi, J.K., Lee, E.A.: Immune signatures correlate with l1 retrotransposition in gastrointestinal cancers. Genome research **28**(8), 1136–1146 (2018)
60. Zhang, X., Zhang, R., Yu, J.: New understanding of the relevant role of line-1 retrotransposition in human disease and immune modulation. Frontiers in Cell and Developmental Biology **8**, 657 (2020)
61. Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L.J., Lohr, J.G., Harris, C.C., Ding, L., Wilson, R.K., *et al.*: Landscape of somatic retrotransposition in human cancers. Science **337**(6097), 967–971 (2012)
62. Bannert, N., Hofmann, H., Block, A., Hohn, O.: Hervs new role in cancer: from accused perpetrators to cheerful protectors. Frontiers in microbiology **9**, 178 (2018)
63. Golkaram, M., Salmans, M.L., Kaplan, S., Vijayaraghavan, R., Martins, M., Khan, N., Garbutt, C., Wise, A., Yao, J., Casimiro, S., *et al.*: Hervs establish a distinct molecular subtype in stage ii/iii colorectal cancer with poor outcome. NPJ genomic medicine **6**(1), 1–11 (2021)
64. Cmero, M., Schmidt, B., Majewski, I.J., Ekert, P.G., Oshlack, A., Davidson, N.M.: Mintie: identifying novel structural and splice variants in transcriptomes using rna-seq data. bioRxiv (2020)
65. Nguyen, H.T., Xue, H., Firlej, V., Ponty, Y., Gallopin, M., Gautheret, D.: Reference-free transcriptome signatures for prostate cancer prognosis. BMC cancer **21**(1), 1–12 (2021)
66. Lorenzi, C., Barriere, S., Villemin, J.-P., Bretones, L.D., Mancheron, A., Ritchie, W.: imoka: k-mer based software to analyze large collections of sequencing data. Genome Biology **21**(1), 1–19 (2020)
67. Thomas, A., Barriere, S., Broseus, L., Brooke, J., Lorenzi, C., Villemin, J.-P., Beurier, G., Sabatier, R., Reynes, C., Mancheron, A., *et al.*: Gecko is a genetic algorithm to classify and explore high throughput sequencing data. Communications biology **2**(1), 1–8 (2019)

Figures



B

Study	LUADtcga	LUADseo	PRADtcga
# Normal samples	58	77	52
# Tumor Samples	524	77	506
DE genes	9 058	6 567	6 292
DE k-mers	6 069 791	12 102 802	2 459 131
DE contigs	368 257	449 860	439 062

C

$LUAD_{tcga} \cap LUAD_{seo}$	N	%
Shared contigs	160610	–
High quality	120822	100.0
Multi hit or repeat	50572	44.1
Up	100405	83.1
Down	20417	16.9
Genic	104762	86.6
Intronic	56146	44.9
SNV or indel	54581	32.1

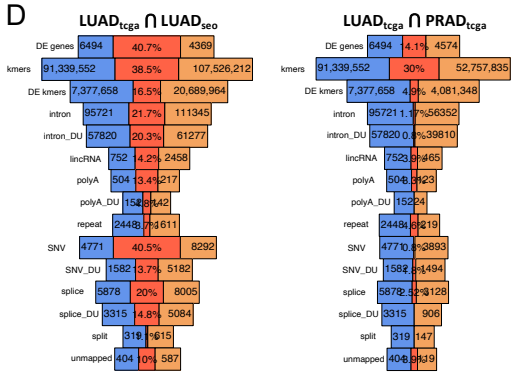
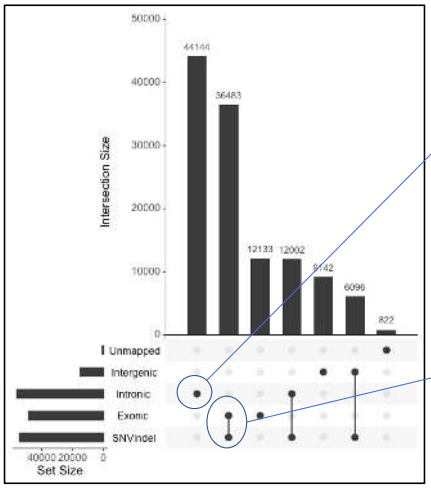


Figure 1 Overall analysis procedure and properties of identified contigs. (A) Computational pipeline for inferring differential contigs in each tumor/normal cohort, extraction of shared contigs and annotation. (B) Sizes of RNA-seq cohorts analyzed and numbers of differential events observed. (C) Summary statistics of differential contigs identified as shared between the LUADtcga and LUADseo analyzes. (D) Number of differential genes, k-mers and contigs in each independent analysis and shared between analyzes. On each row, lateral areas represent differential genes/k-mers/contigs found in each independent analysis and the central area represents shared differential genes/k-mers/contigs. Contigs are classified into different annotation groups.

A



B

ABCC3	162	CP	94	LMO3	105
AC019117.1	281	CU633906.2	84	LRIG3	85
AC243967.1	86	CU633967.1	98	MGST1	106
AGBL4	115	DUXAP9	166	MIR663AHG	78
AGR2	80	ESRP1	79	MUC4	127
AKR1C1	139	FAM3C	79	MYO6	78
ASPH	156	FNDC3B	90	PVT1	210
ATP1B1	121	FP671120.2	185	RNF19A	126
C11orf80	104	GALNTL6	96	RUNX1	261
AMACR	117	GLS	108	SDHAP3	136
CAPN13	94	HIF1A-AS2	97	SFTPA2	84
CAPN8	154	LAMB3	104	SSR3	83
CD46	109	LINC00342	84	SYTL2	109
CIT	77	LINC00624	102	TRA2A	98
CLPTM1L	85	LINC01004	108	XPR1	77

C

AC244205.1	111	IGHV1-69	196	IGKV3-20	351
B2M	255	IGHV3-11	158	IGKV3OR2	432
COL1A1	203	IGHV3-23	217	IGKV4-1	209
IGHA1	846	IGHV3-30	258	IGLC2	1082
IGHA2	292	IGHV3-7	239	IGLC3	181
IGHG1	1843	IGHV4-39	173	IGLL5	441
IGHG2	1094	IGHV4-4	155	IGLV1-47	111
IGHG3	196	IGKC	1974	IGLV3-1	121
IGHG4	1440	IGKJ1	128	IGLV3-21	263
IGHGP	313	IGKV1-33	145	MT-ND6	184
IGHJ4	143	IGKV1-39	238	MUC3A	135
IGHJ5	236	IGKV1-5	345	SFTPA1	1043
IGHV1-18	129	IGKV1OR1-1	118	SFTPA2	1664
IGHV1-2	379	IGKV3-11	155	SFTPB	220
IGHV1-3	138	IGKV3-15	383	SFTPC	3087

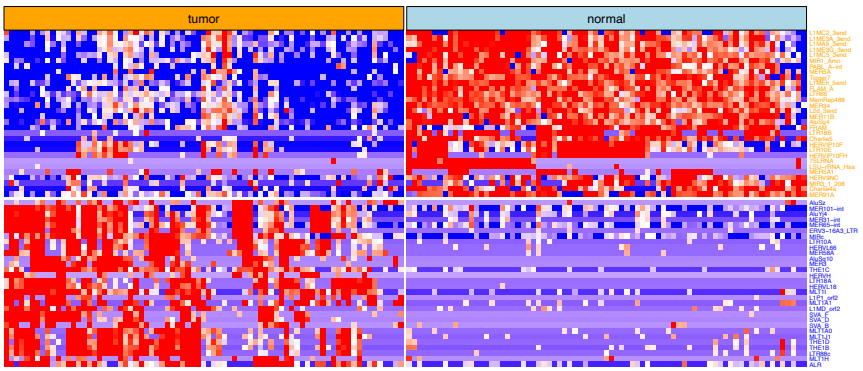
Figure 2 General properties of shared differential expression contigs in LUAD. (A) UpsetR plot of major contig categories based on mapping location and presence of SNV or indels. (B) 45 top genes by number of mapped contigs in the circled intronic category. (C) 45 top genes by number of mapped contigs in the circled exonic+SNVindel category. Numbers of contigs mapped to each gene are indicated.

A

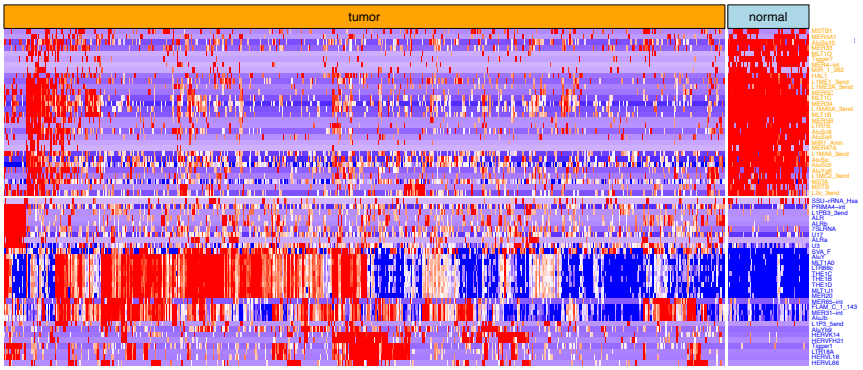
Top 20 repeat types with the most contigs from LUADtcga

rep.type	contigs	Up in tumor	Down in tumor	SNVs	protein.coding	lncRNA
L1P1_orf2	755	754	1	568	233	166
AluSx	455	369	86	250	299	60
FLAM_C_1_143	316	252	64	128	216	38
L1P3_orf2	302	288	14	189	109	82
AluJb	276	227	49	119	189	28
LSU-rRNA_Hsa	264	259	5	249	51	185
AluSz6	174	143	31	74	115	25
AluSp	147	105	42	84	85	18
PRIMA4-int	123	92	31	38	92	8
L1PB_orf2	118	116	2	44	52	26
L1P1_5end	115	115	0	70	39	32
MIR_1_262	111	98	13	10	90	15
AluJr4	110	90	20	47	69	20
AluY	110	79	31	68	59	12
L1HS_5end	109	109	0	65	31	19
L1PREC2_orf2	106	103	3	37	29	32
HERVH	105	96	9	30	19	52
AluSz	104	91	13	61	62	13
L1M3_orf2	89	88	1	5	49	16
AluJr	86	66	20	30	48	14

B



C



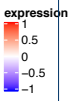


Figure 3 Tumor identity based on top repeats (A) Top 20 repeat types with the most contigs in LUADtcga dataset. (B/C) Expression heatmap of top repeat-containing contigs (ranked by fold change) in the LUADseo (A) and LUADtcga (B) datasets. Contig expression level in heatmaps is represented from blue (lowest) to red (highest).

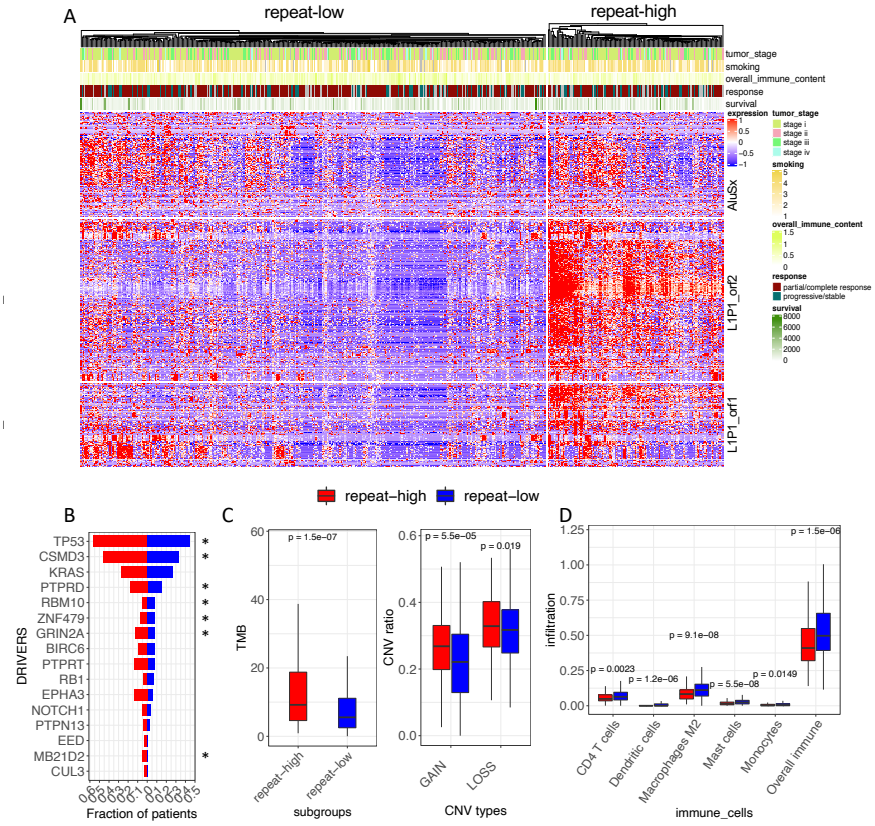


Figure 4 Characterization of patient subgroups based on repeat-containing contigs. (A) Clustering of LUADtcga patients into two subgroups based on Alu and L1P1 repeat expression. Subgroups were defined by K-means. (B) Fraction of patients with driver mutations for 16 COSMIC LUAD drivers. Drivers with Fisher P value < 0.05 were marked with star. (C) Mutational burden and CNV frequency distribution between two subgroups. (D) Variation of immune features between subgroups. The red and blue represent the repeat-high and repeat-low subgroups, respectively. P-values are computed by Wilcoxon test.

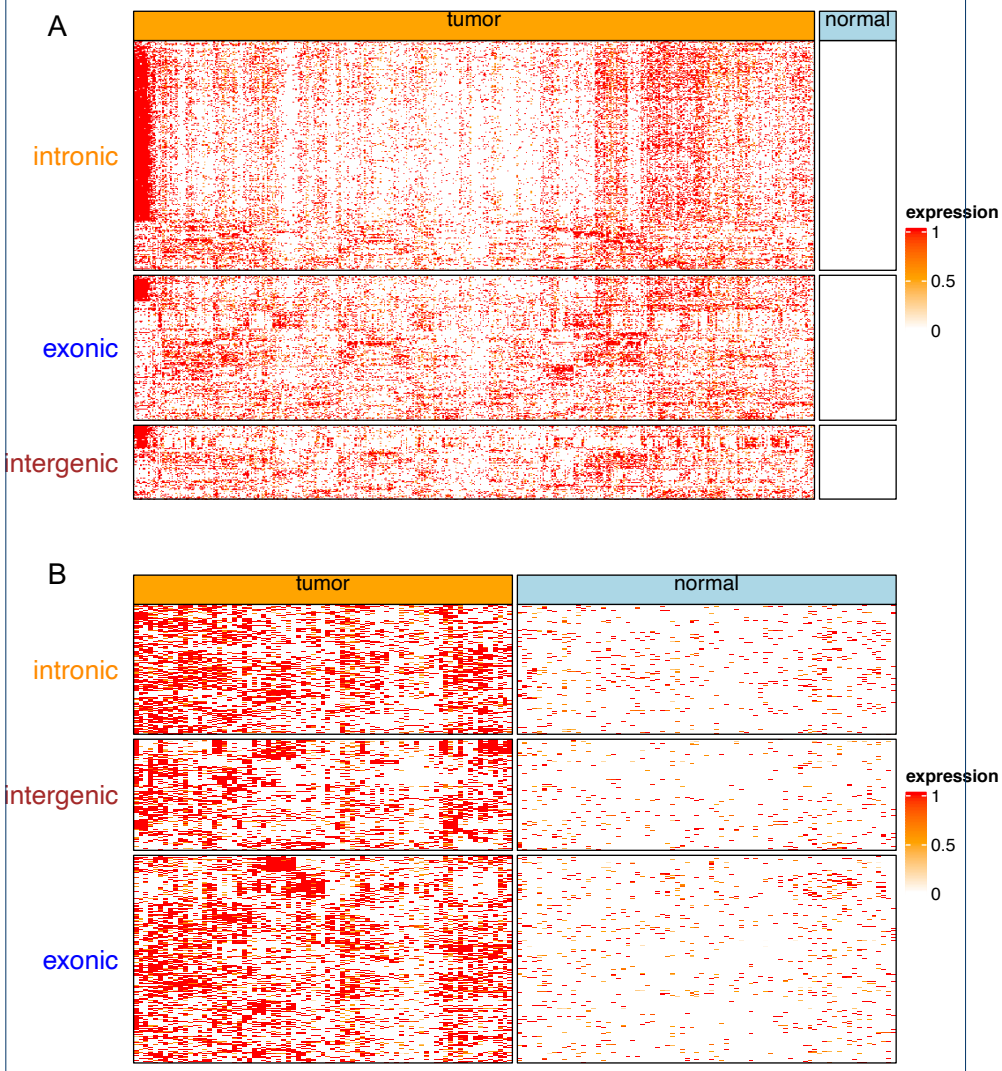


Figure 5 Expression heatmap of potential neoantigen sources in the two LUAD datasets. Tumor-specific contigs were first selected in LUADtcga (A) and validated in the LUADseo dataset (B).

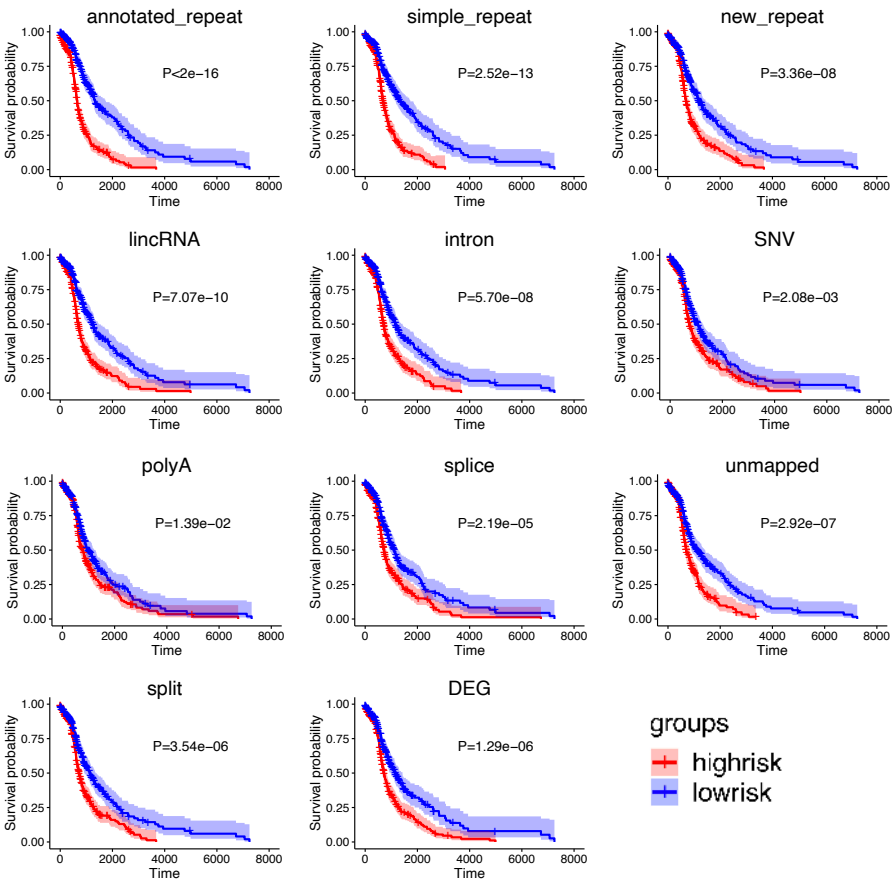


Figure 6 Kaplan-Meyer curves of multivariate survival models per class of event. Patients in high and low-risk groups are shown in red and blue, respectively. Repeat events were separated into annotated, new and simple repeats. The other categories with more lasso-selected contigs were also included (Additional file 3: Table S8).

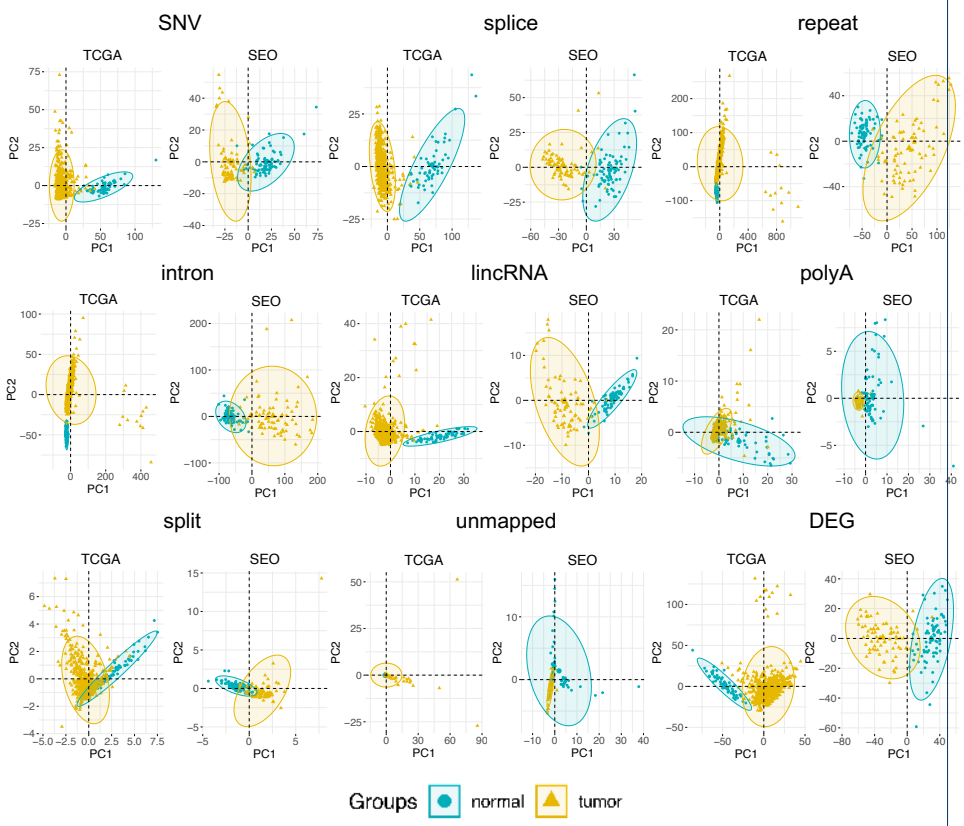


Figure 7 Principal component analysis of samples based on DE contigs and genes. Each panel represents PCA performed with one class of contigs and with differentially expressed genes (DEG), for the LUADtcga (TCGA) and LUADseo (SEO) datasets. Normal and tumor samples are marked using blue and yellow, respectively. Confidence ellipses are drawn with package factoextra for each group.

Additional files

Additional file 1 — Figure S1
The graph-based protocol detecting shared contigs between TCGA and SEO datasets. (A) Contigs from each dataset. The bars marked with the same color represent the same k-mers. (B) Cliques construction based on the common k-mers. (C) Shared contigs identification based on the cliques.

Additional file 1 — Figure S2
Enrichment analysis of shared DEGs and contigs between TCGA and SEO datasets. The x axis represents the ranked DEGs or contigs based on log2FC in ascending order. The red vertical dotted line represents the position of log2FC cutoff.

Additional file 1 — Figure S3
Hypervariable genes in our analysis. Upset graph shows the overlap between different categories, including intronic, exonic, spliced and SNV or indel.

Additional file 1 — Figure S4
IGV views of intronic events. Each frame shows a metabam file composed of randomly sampled reads corresponding to the subcohort indicated on the left panel. The lower panel shows DE contigs and Gencode annotation. A: multiple intron retention in CEACAM5; B: lncRNA element expressed in an intron of MBD5; C: intron retention in EGFR; D: intron retention in MET.

Additional file 1 — Figure S5
Intronic event analysis. (A) Log2FC values of the top 20 intronic events (DU). Red and blue colors represent the expression fold change of intronic contigs and host genes, respectively. (B) Gene Ontology functional enrichment. Color represents the P-values and size represents the ratio of genes.

Additional file 1 — Figure S6
IGV views of lincRNA elements overexpressed in tumors. Each frame shows a metabam file composed of randomly sampled reads corresponding to the subcohort indicated on the left panel. The lower panel shows DE contigs and Gencode annotation.

Additional file 1 — Figure S7
Heatmap of Spearman correlation coefficient (CC) of contig counts and abundance of immune cell types evaluated by CIBERSORT. All contigs with a CC>0.5 with at least one immune cell type are shown. Immune cells not correlated with at least one contig are not shown. Row names show gene symbols and repeat types of contigs, whenever applicable. Row name colors indicate different contig categories. The log2FC sidebar shows expression fold change of contigs between normal and tumor samples.

Additional file 1 — Figure S8
Fractions of event types in strictly tumoral contigs predicted to produce neoantigens ("neo", N=472) and total shared DE contigs ("all", N=2375). Intergenic contigs are significantly over-represented in "neo" contigs (Fisher's exact P=1.2e-20).

Additional file 2 — Table S1-S12
Table S1: Nucleotide contents of DE-kupl contigs for the TCGA LUAD dataset. Table S2: Description of event categories extracted from DE-kupl-annot tables. Table S3: General characteristics of contigs shared between LUADtcga and LUADseo. Table S4: Summary statistics for all event categories in contigs shared between LUADtcga and LUADseo. Table S5: Genes with more than 35 mapped contigs (shared LUAD contigs. Colored columns indicate ratio of contigs in said categories). Table S6: Blast results of 100 contigs mapped to IGHV genes. Table S7: Univariate Cox regression results of all categories. Table S8: Enrichment of OS-related events. Table S9: Multivariate Cox regression results of all categories. Table S10: Peptides of strong binding levels predicted by netMHCpan 4.0 from "neoRNA" contigs. Table S11: GO enrichment using host genes of immune related contigs. Table S12: Contigs correlated with the $klebsiella$ contig.

Additional file 3
Correlation analysis of number of contigs and host gene expression.