Published online 1 February 2022

https://doi.org/10.1093/narcan/zcac001

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

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

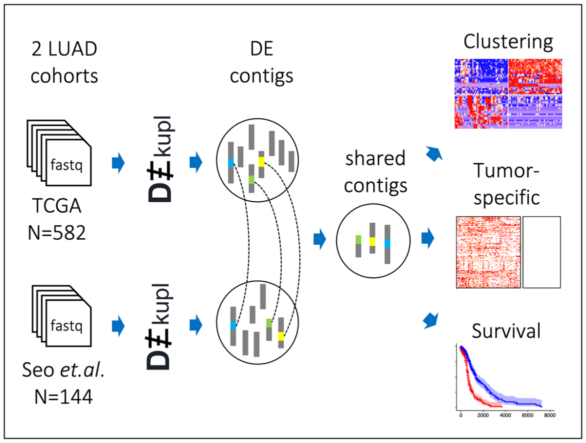
[1]Institute for Integrative Biology of the Cell (I2BC), Université Paris-Saclay, CNRS, CEA, 1 avenue de la Terrasse, 91190, Gif-sur-Yvette, France, [2]Annoroad Gene Technology Co., Ltd, 100176 Beijing, China and [3]Gustave Roussy, 114 rue Edouard Vaillant, 94800, Villejuif, France

Received July 31, 2021; Revised November 18, 2021; Editorial Decision December 17, 2021; Accepted January 10, 2022

ABSTRACT

The identity of cancer cells is defined by the interplay between genetic, epigenetic transcriptional and post-transcriptional variation. A lot of this variation is present in RNA-seq data and can be captured at once using reference-free, k-mer analysis. An important issue with k-mer analysis, however, is the difficulty of distinguishing signal from noise. Here, we use two independent lung adenocarcinoma datasets to identify all reproducible events at the k-mer level, in a tumor versus normal setting. We find reproducible events in many different locations (introns, intergenic, repeats) and forms (spliced, polyadenylated, chimeric etc.). We systematically analyze events that are ignored in conventional transcriptomics and assess their value as biomarkers and for tumor classification, survival prediction, neoantigen prediction and correlation with the immune microenvironment. We find that unannotated lincRNAs, novel splice variants, endogenous HERV, Line1 and Alu repeats and bacterial RNAs each contribute to different, important aspects of tumor identity. We argue that differential RNA-seq analysis of tumor/normal sample collections would benefit from this type k-mer analysis to cast a wider net on important cancer-related events. The code is available at https://github.com/Transipedia/dekupl-lung-cancer-inter-cohort.

GRAPHICAL ABSTRACT



INTRODUCTION

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 are still utilized mostly to quantify the expression of annotated genes listed in a reference transcriptome. This ignores a wide array of mRNA isoforms, noncoding 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

[*]To whom correspondence should be addressed. Tel: +33 0467339488; Email: daniel.gautheret@i2bc.paris-saclay.fr

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 directly from the raw 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 versus 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. 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.

MATERIALS AND 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. This was found to perform 10 times faster than DESeq2. The performances of DESeq2, Limma and *T*-test for differential expression analysis have been evaluated before (19). K-mer counts were log-transformed and Limma was used to calculate differential expression *P*-values, adjusted for multiple testing using the Benjamini–Hochberg procedure (20). Retention thresholds for log2 fold changes and adjusted *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 File 2: Supplementary Table S1). 95.9% of LUADtcga contigs (98.2% of LUADseo) mapped to the human genome and, among mapped contigs, 86.1% were aligned over >95% of their length, indicating a low rate of misassembly (Additional File 2: Supplementary Table S2). The ratio of correct assembly is certainly higher than this since a contig can be partly aligned and yet be correctly assembled. Also, this ratio increases with contig length (Additional File 1: Supplementary Figure 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 (21) 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) (22).

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) (23), with contigs 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: Supplementary Figure S2). As any k-mer is present in only one contig per dataset and we compared two datasets, all cliques only involved two contigs. However the method can generalize to three or more datasets allowing cliques of three or more. We thus extracted all cliques to identify shared contigs. Hereafter we use the ∩ 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 ∩ LUADseo and LUADtcga ∩ 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 (24) alignments of contigs to the DFAM repeat database (25). 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 File 2: Supplementary 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 DE-kupl's differential usage adjusted *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) (26).

Sample clustering based on repeats

We used the K-means algorithm 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 were 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 (22).

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 (https://CRAN.R-project.org/package=survival; https://CRAN.R-project.org/package=survminer). Hazard ratios (HR) and adjusted *P*-values were calculated for each contig. Contigs with HR > 1 and adjusted *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) (33) 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 (34).

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) (https://CRAN.R-project.org/package=factoextra). Heatmap views were obtained using ComplexHeatmap (22).

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) (35) with default parameters. BAM files were visualized using Integrative Genomics Viewer (IGV 2.6.2) (36).

RESULTS

Gene-level versus contig-level differential events

We performed tumor versus 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 (Figure 1A). For simplicity, we shall hereafter use term ‘expression’ when referring to either gene expression or k-mer/contig counts. While the number of DE genes in the three comparisons ranged from 6000 to 9000, 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 (Figure 1B). Comparison of DE contigs and DE genes identified by a conventional protocol showed that most DE genes were covered by DE contigs (74% in the LUADtcga cohort), and this in spite of the removal of k-mers from annotated genes. Indeed, as noted before (12), SNPs and intronic reads generate k-mers that are not present in references and thus are not removed. However, DE contigs were also found in 22% of non-DE genes and 11% of intergenic regions (Additional File 2: Supplementary Table S4).

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 (Figure 1A, Additional File 1: Supplementary Figure S2). 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 File 2: Supplementary Table S3 and Materials 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 (Figure 1C). Over these, 120 822 contigs were considered of sufficient quality based on counts and occurrence in a minimal number of samples (see Materials and Methods). 83% of shared contigs were overexpressed in tumors versus only 17% underexpressed (Figure 1C).

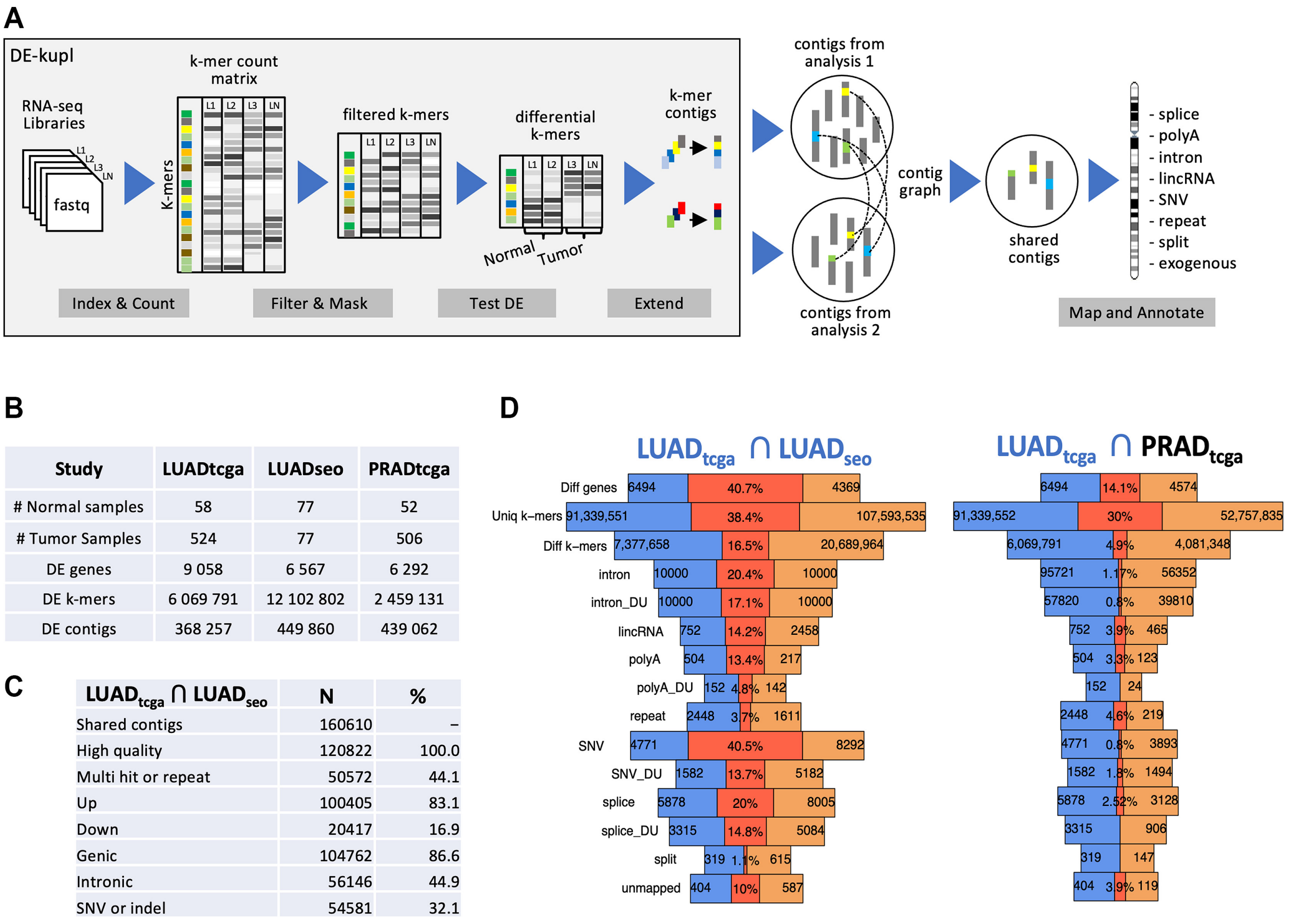
Event replicability

The replicability of differential events was generally lower for k-mer or contigs than for genes. Figure 1D shows the number of DE genes and contigs shared by the two independent LUAD analyses, with contigs binned by annotation class. About 41% of DE genes (3032 genes) were shared by the two LUAD analyses, compared to an average of 14% for DE 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 ∩ PRADtcga, Figure 1D), and this applied to all event classes except repeats. This indicates that, although k-mer based DE events are noisy, a significant subset is replicable in independent studies. The likelihood to be shared between cohorts was strongly correlated with the fold-change value of DE contigs (Additional File 1: Supplementary Figure S3), demonstrating the non-randomness of high scoring, non-reference events. Likelihood to be shared also increased with contig size (Additional File 1: Supplementary Figure S1). More than 65% of contigs over 500 nt in size were shared between cohorts, thus reaching a higher level of replicability than DE genes.

DE contig localization, hypervariable genes

The majority of shared contigs are genic (83%), 45% are intronic and 32% carry SNVs or indels (Figure 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 (Figure 2A, B, Additional File 2: Supplementary 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 (Figure 2A, C, Additional File 2: Supplementary 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 (37), hence these genes are seen as upregulated 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 (38). To evaluate the accuracy of DE-kupl contigs assembled from IG genes, we se-



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

$\text{LUAD}_{\text{tcga}} \cap \text{LUAD}_{\text{seo}}$	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

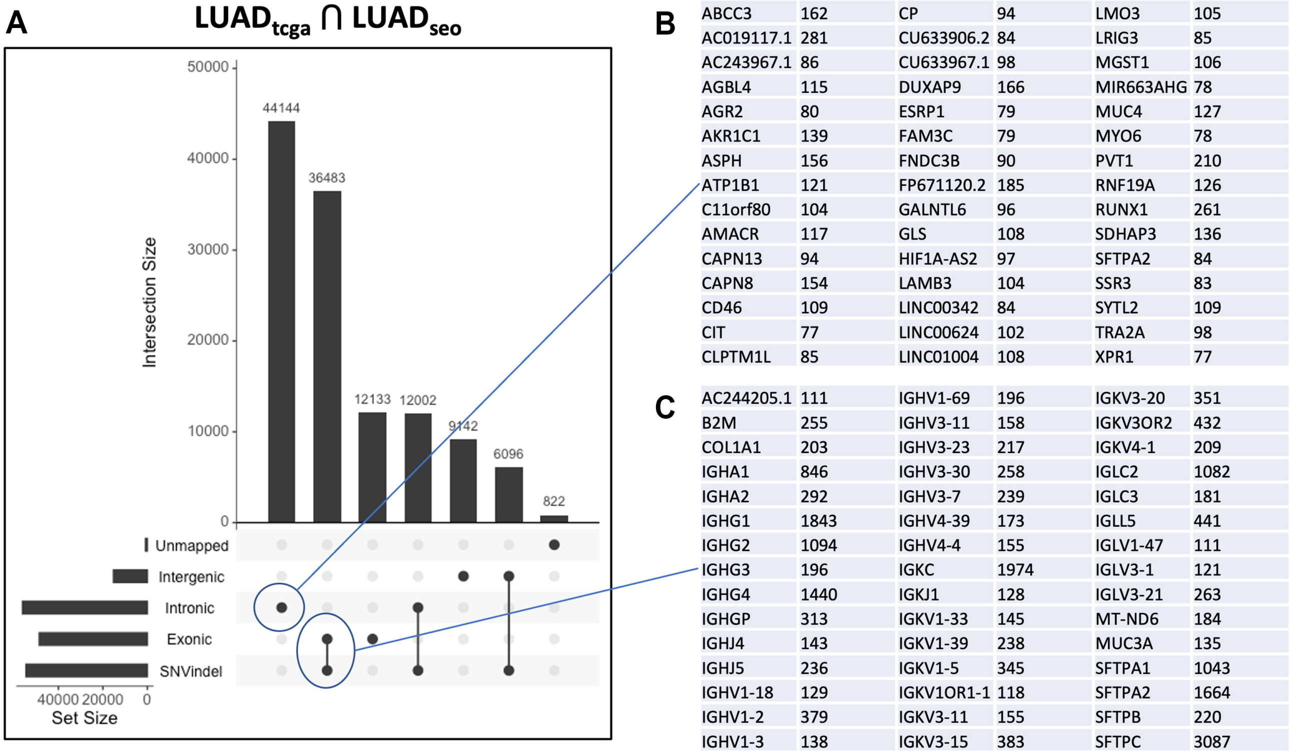
Figure 1. Overall analysis procedure and properties of identified contigs. (**A**) Computational pipeline for inferring differential contigs in each tumor/normal cohort, extraction of shared contigs and annotation. (**B**) Sizes of RNA-seq cohorts analyzed and numbers of differential events observed. (**C**) Summary statistics of differential contigs identified as shared between the LUADtcga and LUADseo analyses. (**D**) Number of differential genes, k-mers and contigs in each independent analysis and shared between analyses. 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.

lected all contigs mapped to one arbitrary IG gene (IGHV: 100 contigs) and aligned them to IGHV contigs from the IMGT database (39). Ninety out of 100 contigs had significant matches in the corresponding IMGT category extending over 90% of the contig length (Additional File 2: Supplementary Table S6).

Other hypervariable loci were found in surfactant protein (*SFTP*) and Mucin genes which are known to harbor a high level of polymorphism (40,41). 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 harbor 1324 contigs including 42 splice variants (Additional File 1: Supplementary Figure S4A,B, Additional File 2: Supplementary Table S5). While *SFTP* contigs were all underexpressed in tumors, Mucin contigs were mostly overexpressed (Additional File 2: Supplementary Table S5). Mucins are immunogenic (41) and are important biomarkers for prognosis (42) and drug resistance (43). The existence of recurrent mucin variants overexpressed in tumors may be relevant for these therapeutic and biomarker developments. We also observed hypervariability in *CEACAM*5 and *KR*19, two other prognostic biomarkers and/or immunotherapy targets (44,45) (Additional File 1: Supplementary Figure S4C, Additional File 2: Supplementary 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: Supplementary Table S7). About 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: Supplementary Figure S5A) from transcription units occurring within introns (example in Additional File 1: Supplementary Figure S5B). We observed intron retention events in lung cancer drivers *EGFR* and *MET* (Additional File 1: Supplementary Figure S5C and Additional File 1: Supplementary Figure S5D). In *EGFR*, the retained intron was located between exons 18 and 19, just upstream of the principal oncogenic *EGFR* mu-



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

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.

tations located in exons 19–21. Intron retention before exon 19 would likely produce a truncated form of *EGFR* compatible with oncogenic activation.

The 20 intronic events with the most significant differential usage (Additional File 1: Supplementary Figure S6A) 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: Supplementary Figure S6B), 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 labeled 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: Supplementary Figure S7. 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

Transposable elements (TEs) are reactivated in cancer (46,47), and their expression correlates with changes in the tumor microenvironment (8,48,,49. Those repeated elements are difficult to analyze by standard RNA-seq pipelines due to ambiguity in the alignment process. We questioned whether the alignment-free procedure could help reveal these events. From the initial set of 50 572 contigs annotated as repeats (Figure 1C), we selected a high quality subset of 10 341 contigs over 60 bp in size and with expression above a set threshold (see Materials and Methods). Of these, 87.7% were overexpressed in tumors (Additional File 2: Supplementary Table S7).

Figure 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 TEs in the human genome (50). L1P1/L1Hs elements, as well as human endogenous retrovirus (HERV), were almost exclusively over-expressed in tumors, consistent with tumor-specific activation of these elements. In

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

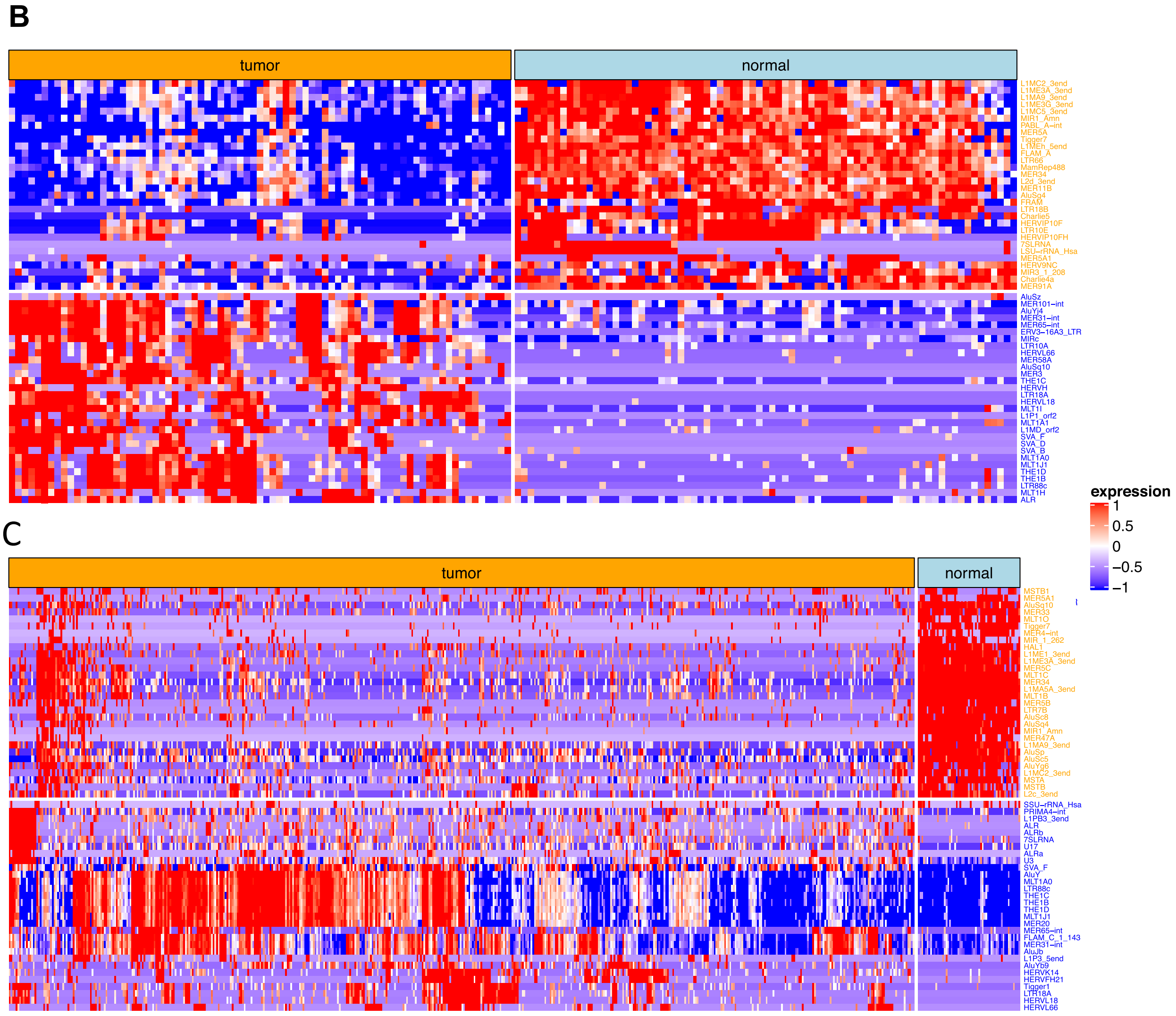


Figure 3. Analysis of differential repeats. (**A**) Top 20 repeat types with the most contigs in the LUADtcga dataset. (**B**/**C**) Expression heatmaps of top repeat-containing contigs (ranked by fold change) in the LUADseo (A) and LUADtcga (B) datasets. Contig expression level is represented from blue (lowest) to red (highest). For each type of repeat, the contig with the highest absolute fold-change is shown.

contrast, Alu elements, which are often expressed as part of protein coding genes, were either over- or under-expressed in tumors. Figure 3A shows the top 20 repeat types that contribute more contigs. Figure 3B,C shows the expression heatmaps of the 60 repeats contributing more contigs. Comparison with DE repeat families identified with the Salmon software (51) on the LUADtcga dataset (albeit restricted to matched normal/tumor samples) (47) shows our approach identifies a higher number of DE repeat families (732 versus 61, Supplementary File S2: Supplementary Table S8). However, one should keep in mind that Salmon assigns reads to a predefined list of full-length transcripts, which is much more stringent than reporting DE contigs.

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: Supplementary Table S7). Many of these repeats were from Mucins, immunoglobulins and multicopy gene families such as *NBPF* and *TBC*1. 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 Figure 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') (Figure 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 *CSMD*3, *TP*53, *EGFR*, *PTPRD*, *PTPRT*, *GRIN*2*A*, *EPHA*3 and *MB*21*D*2 (Figure 4B, Fisher $P < 0.05$). Patients in the repeat-high group had a significantly higher TMB (Wilcoxon $P = 1.5\text{e-}07$) and a higher ratio of CNVs than other patients (Wilcoxon $P = 5.5\text{e-}05$ for gain; $P = 0.019$ for loss) (Figure 4C). We finally compared repeat subgroups for immune cell contents predicted by gene expression deconvolution. The repeat-high subgroup had a lower overall immune content than the repeat-low subgroup (Figure 4D).

In summary, properties of 'repeat-high' tumors are consistent with previous observations that derepression of TEs can be controlled by *P*53 (48,52), correlate with a repressed immune environment (8,47,52,49) and can lead to genome instability (46).

Immune cell-associated contigs

We extracted DE contigs correlating with predicted immune cell contents in both LUADtcga and LUADseo cohorts (Additional File 1: Supplementary Figure S8A, S8B). Most contigs were uniquely mapped to genic sequences and underexpressed in tumors. Positive correlations were consistently observed in both cohorts with M1/M2 macrophages and resting cells, i.e. with a generally repressive or quiescent immune environment. In both cohorts, most immune cell-associated contigs were from leukocyte-specific or immunity related genes (e.g. *MSA*4*A*7, *MSR*1, *OSCAR*, *TLR*8, etc.), suggesting these contigs originated from the immune cells themselves.

In the LUADseo cohort, two contigs positively correlated with naive CD4+ T-cells (Additional File 1: Supplementary Figure S8B). One of them was strongly repressed in tumors and matched a *Klebsiella pneumoniae* rRNA fragment. *Klebsiella* is a common lung bacterium against which cross-reactive T-cells are present in the naive CD4+ T-cell repertoire (53). This result thus suggests 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: Supplementary Table S9), in line with *SFTP* roles in defense against respiratory pathogens (54).

We noticed among immune cell-correlated contigs the presence of HERV (human endogenous retrovirus) family repeats overexpressed in tumors in both cohorts. An HERV-E contig in the LUADseo cohort correlated with CD4+ T-cells and was expressed from the *env* gene of a near full-length retroelement. Relaxing correlation thresholds on HERV contigs revealed several HERV-E elements overexpressed in tumors and correlating with various types of CD4+ T-cells, but also with CD8+ T-cells and NK-cells in both cohorts (Additional File 1: Supplementary Figure S9).

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 nonself by the epitope presentation machinery, triggering antitumor immune response (55). 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 (56). 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 (57,58). Therefore we thought our reference-free approach could be a valuable 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 (Figure 5A). About 20% of these contigs ($N = 472$) were also silent in normal tissues of the

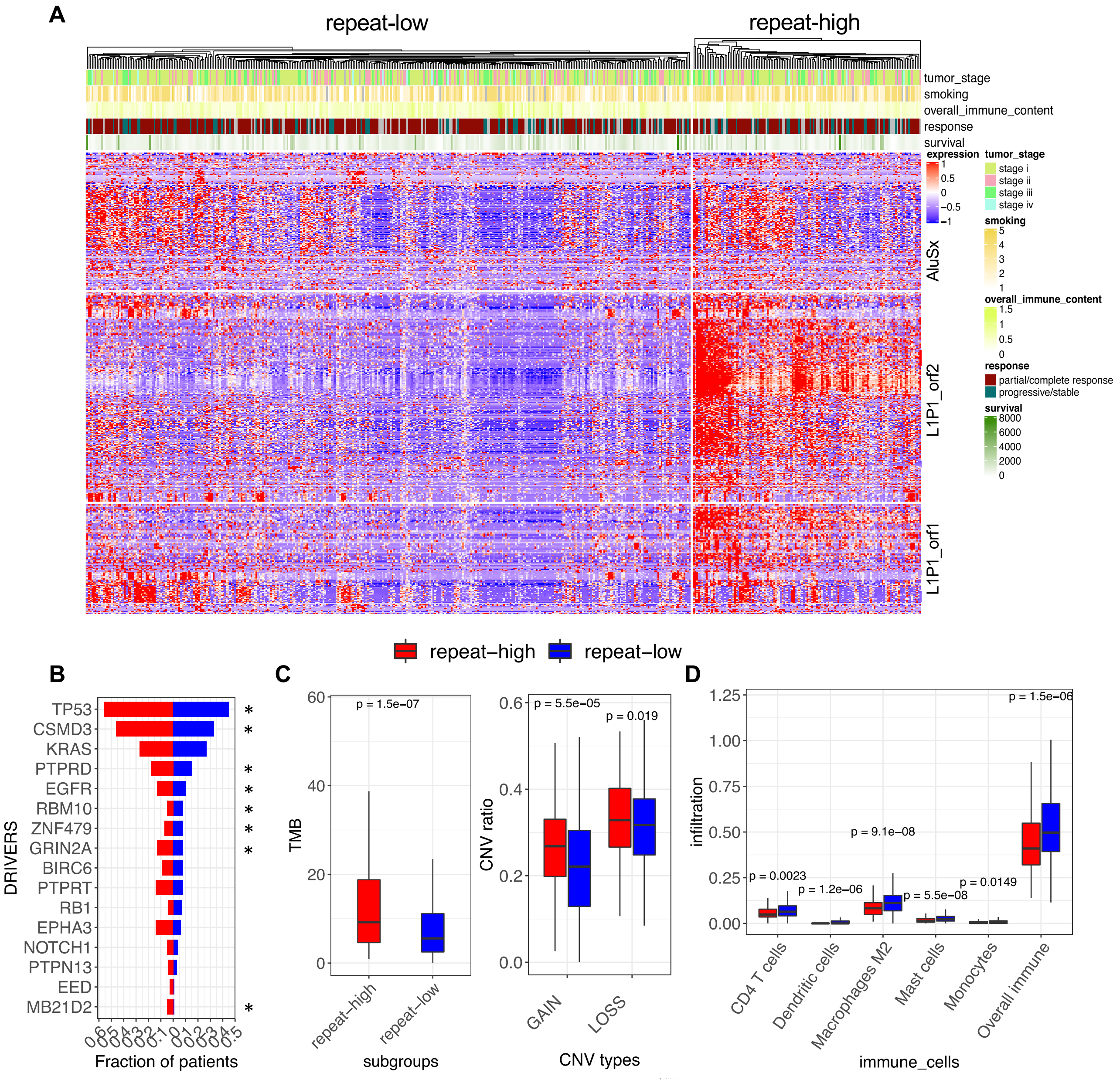


Figure 4. Characterization of patient subgroups based on repeat-containing contigs. (**A**) Clustering of LUADtcga patients into two subgroups based on Alu and L1P1 repeat expression. Subgroups were defined by K-means. (**B**) Fraction of patients with driver mutations for 16 COSMIC LUAD drivers. Drivers 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.

LUADseo cohort (Figure 5B). 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: Supplementary 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: Supplementary Figure S10). Overall, contigs from intergenic regions, non-coding RNAs and pseudogenes contributed 58% of predicted neoantigens (Additional File 2: Supplementary Table S10), consistent with previous reports of abundant neoantigen production from non-coding regions in other cancers (58).

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

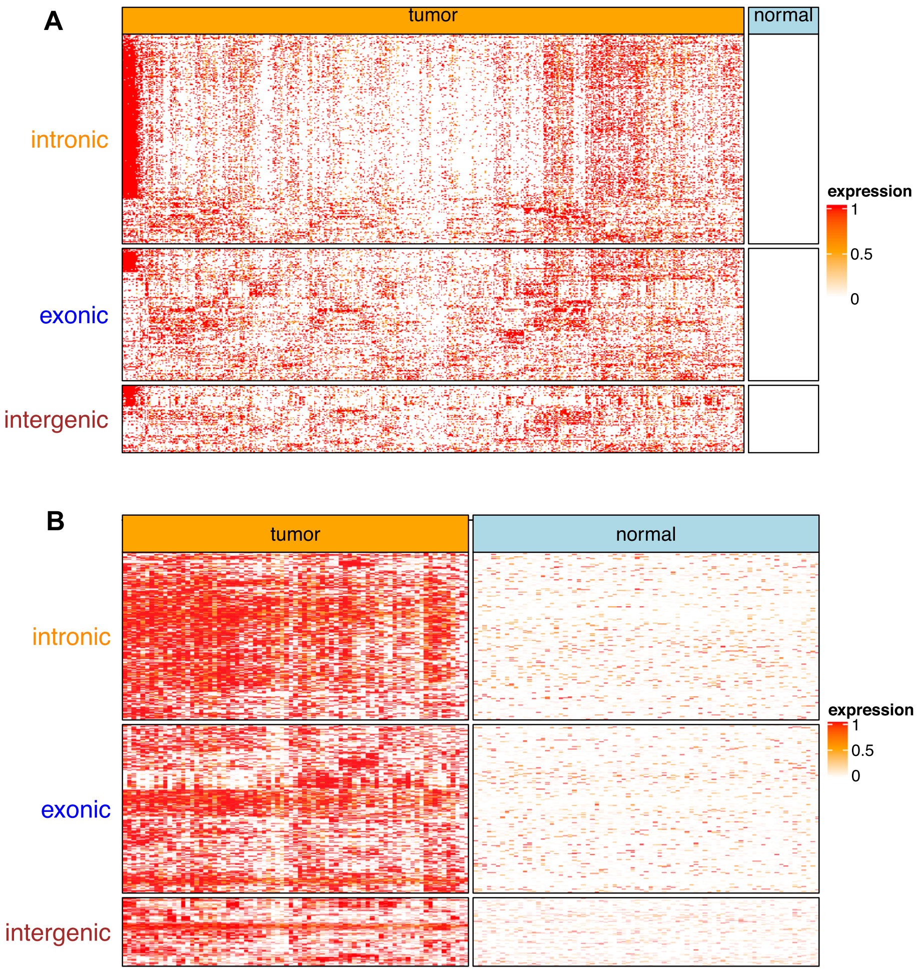


Figure 5. Expression heatmap of potential neoantigen sources in the two LUAD datasets. Tumor-specific contigs were first selected in LUADtcga (**A**) and validated in the LUADseo dataset (**B**).

File 2: Supplementary Table S11). Outcome-related contigs are mostly enriched in repeats (Additional File 2: Supplementary Table S12), 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: Supplementary Table S11). 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: Supplementary Table S11) (59).

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: Supplementary Table S13). Models based on annotated and simple repeats had the best prognostic power (log-rank $P = 2e\text{-}16$, 2e-13, respectively, Figure 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 File 2: Supplementary Table S13). 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 (Figure 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 (Figure 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 Figure 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.

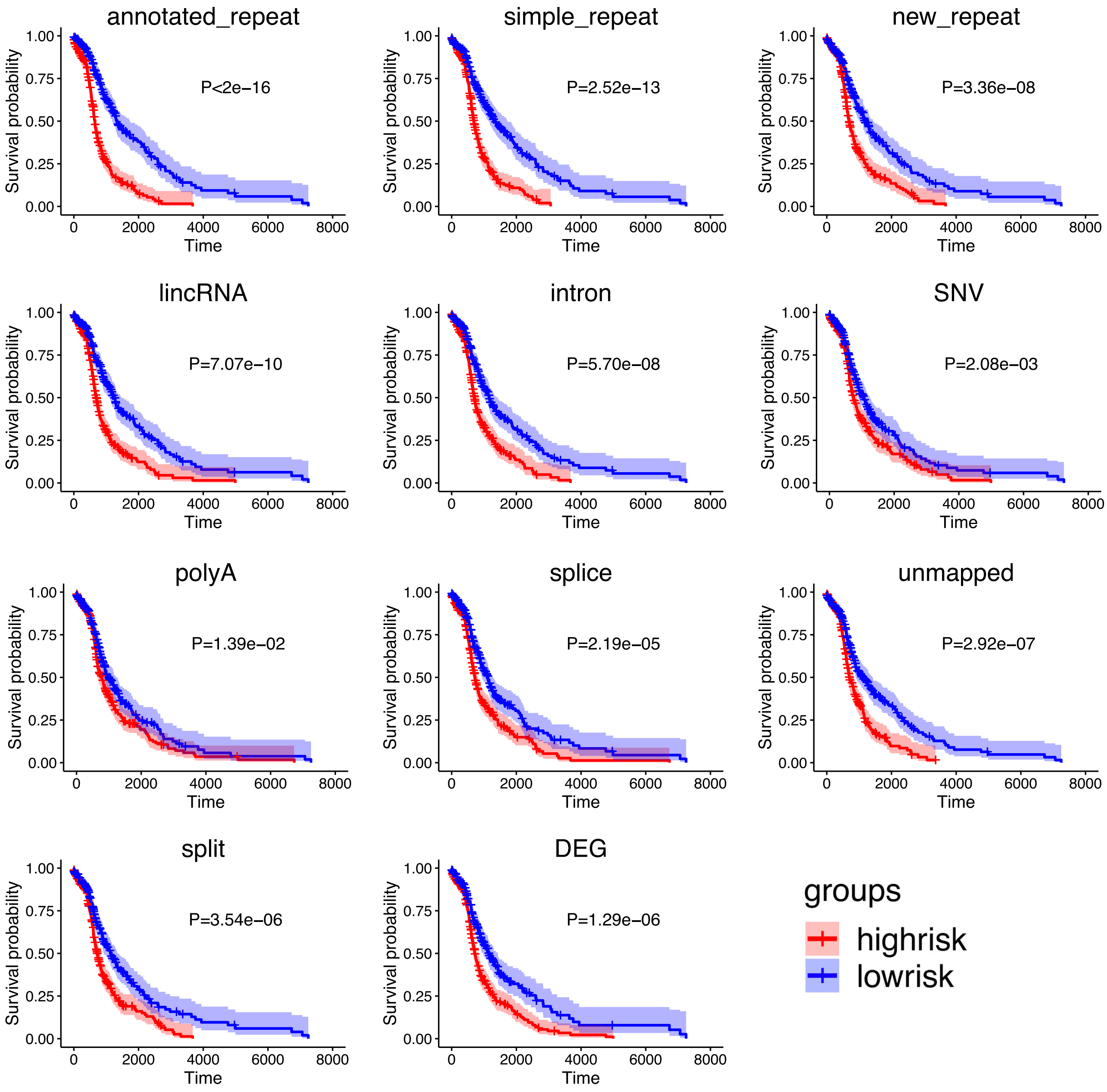


Figure 6. Kaplan–Meier curves of multivariate survival models per class of event. Patients in high and low-risk groups are shown in red and blue, respectively. Repeat events were separated into annotated, new and simple repeats. The other categories with more lasso-selected contigs were also included (Additional File 2: Supplementary Table S12).

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 may 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 *P*53, a higher mutational and copy number burden and a reduced immune cell infiltrate, recapitulating prior knowledge on the effects of TE reactivation (8,47–49,52,52).

Expressed repeats also had significant prognostic power. Multivariate signatures composed of HERV and L1 elements, or simple repeats, stratified patients into distinct

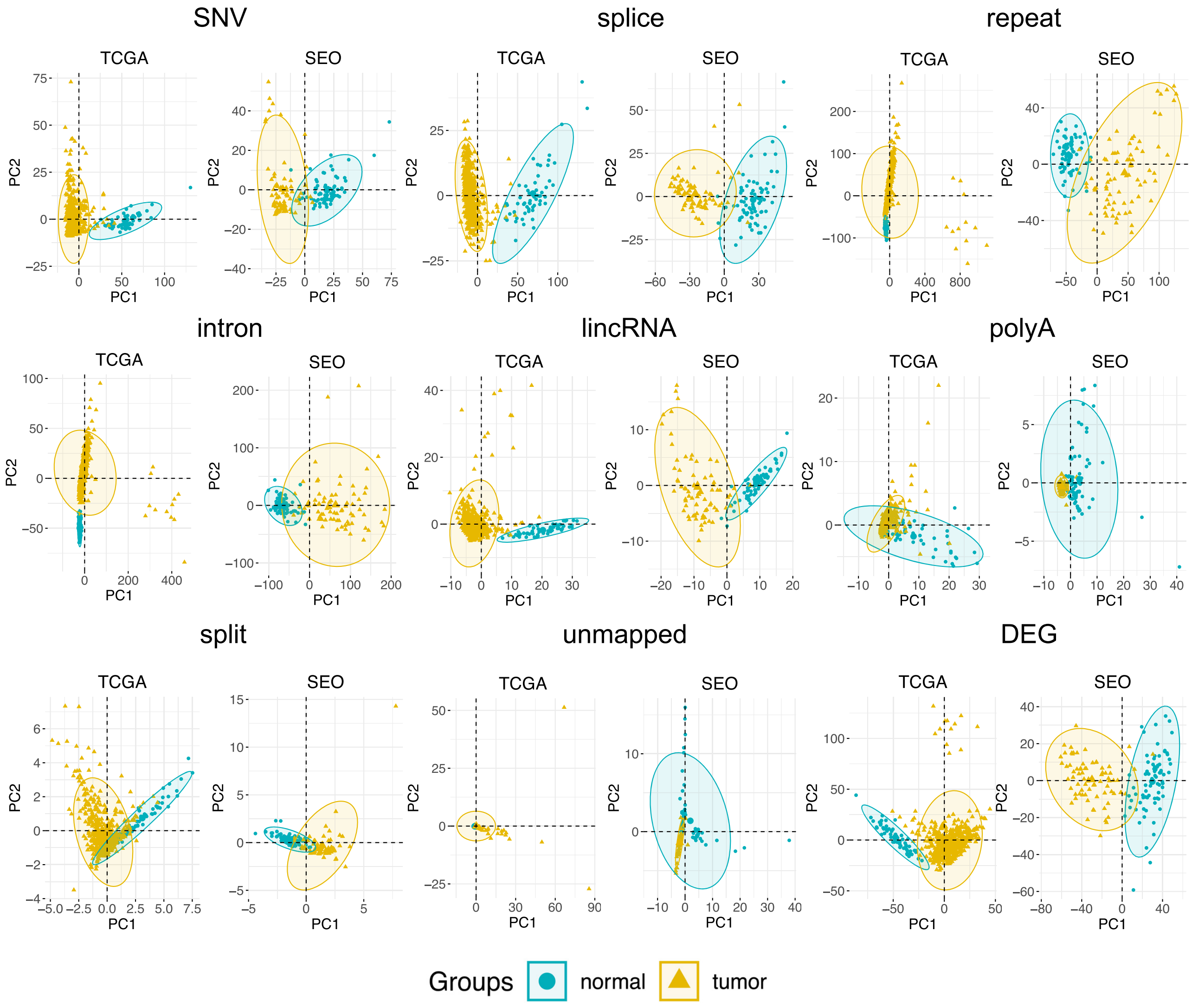


Figure 7. Principal component analysis of samples based on DE contigs and genes. Each panel represents PCA performed with one class of contigs and with differentially expressed genes (DEG), for the LUADtcga (TCGA) and LUADseo (SEO) datasets. Normal and tumor samples are marked using circles and triangles,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

survival groups. HERV contigs were among the strongest outcome predictors, with least favorable outcome for higher HERV expression, in line with previous observations of HERV expression associated with poor prognosis in melanoma, breast and colorectal cancer (60–62).

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 may be further explored, including when coming from novel insertion events. Furthermore, it should be noted that conventional read mapping does not guarantee either that full-length TEs are expressed.

An alternative strategy for reference-free RNA-seq analysis is to assemble reads first using *de novo* assembly software. However, this process is error-prone (63,64) and computationally intensive. Kazemian *et al*. achieved *de novo* assembly of over a thousand TCGA samples (65), but this was restricted to unmapped reads, therefore missing most RNA variations. More recently, an assembly-first strategy was introduced to identify patient-specific abnormal transcripts under a ‘1 to *n*’ experimental design (66). *De novo* read assembly produces longer contigs, which facilitates biological interpretation and is better suited to large structural variants. With insertions longer than the size of a k-mer, *de novo* assembly can recover the whole insertion while k-mer assembly may end up with two contigs. However our ‘DE-first’ strategy retains many advantages. First, by removing condition-irrelevant k-mers at the initial stage, it significantly reduces computing costs and limits risks of mis-assembly. Second, as our assembly process stops when ambiguities are met, every variation is singled out; thus every SNV, splice junction or other type of local event is independently quantified and its association to disease can be assessed.

An attractive aspect of reference-free RNA-seq analysis is the capacity to identify novel forms of known cancer drivers or biomarkers. We identified novel intron retention events in *EGFR* and *MET* and multiple new variants of *CEACAM*5 and *KR*19. 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, mostly based on mapping strategies (55) but also reference-free using k-mers (58). The reference-free 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 large patient groups.

The wealth of information uncovered in the present study is a strong incentive to explore other applications of reference-free transcriptomics, notably for building predictive models. We (67) and others (68,69) 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.

SUPPLEMENTARY DATA

Supplementary data are available at NAR Cancer online.

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. Y.W. and D.G. designed the workflow and analyzed the results; Y.W. downloaded and processed the datasets; Y.W. and D.G. wrote the manuscript; M.A. and M.G. assisted in statistical analysis; H.X. participated in algorithm design. A.L. annotated the repeat types.

FUNDING

Agence Nationale de la Recherche [ANR-18-CE45-0020]; Annoroad Gene Technology, Beijing (to Y.W.).

Conflict of interest statement. None declared.

REFERENCES

1. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
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.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Onco.*, **27**, 1160.
3. Morillon,A. and Gautheret,D. (2019) Bridging the gap between reference and real transcriptomes. *Genome Biol.*, **20**, 112.
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.* (2018) Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell*, **34**, 211–224.
5. Vitting-Seerup,K. and Sandelin,A. (2019) IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics*, **35**, 4469–4471.
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.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.*, **47**, 199–208.
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.* (2017) A pan-cancer analysis of the expression and clinical relevance of small nucleolar RNAs in human cancer. *Cell Rep.*, **21**, 1968–1981.
8. Solovyov,A., Vabret,N., Arora,K.S., Snyder,A., Funt,S.A., Bajorin,D.F., Rosenberg,J.E., Bhardwaj,N., Ting,D.T. and Greenbaum,B.D. (2018) Global cancer transcriptome quantifies repeat element polarization between immunotherapy responsive and T cell suppressive classes. *Cell Rep.*, **23**, 512–521.
9. Robinson,K.M., Crabtree,J., Mattick,J.S., Anderson,K.E. and Hotopp,J. C.D. (2017) Distinguishing potential bacteria-tumor associations from contamination in a secondary data analysis of public cancer genome sequence data. *Microbiome*, **5**, 9.
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.* (2020) The landscape of viral associations in human cancers. *Nat. Genet.*, **52**, 320–330.
11. Amemiya,H.M., Kundaje,A. and Boyle,A.P. (2019) The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.-UK*, **9**, 9354.
12. Audoux,J., Philippe,N., Chikhi,R., Salson,M., Gallopin,M., Gabriel,M., Le Coz,J., Drouineau,E., Commes,T. and Gautheret,D. (2017) DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. *Genome Biol.*, **18**, 243.
13. Ioannidis,J.P. (2005) Microarrays and molecular research: noise discovery? *Lancet (London, England)*, **365**, 454–455.
14. Michiels,S., Koscielny,S., Boulet,T. and Hill,C. (2007) Gene expression profiling in cancer research. *Bull. du Cancer*, **94**, 976–980.
15. Network,C. G. A.R. *et al.* (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**, 543.
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.* (2012) The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res.*, **22**, 2109–2119.
17. Abeshouse,A., Ahn,J., Akbani,R., Ally,A., Amin,S., Andry,C.D., Annala,M., Aprikian,A., Armenia,J., Arora,A. *et al.* (2015) The molecular taxonomy of primary prostate cancer. *Cell*, **163**, 1011–1025.
18. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47–e47.
19. Soneson,C. and Delorenzi,M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform.*, **14**, 91.
20. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. (Methodological)*, **57**, 289–300.
21. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
22. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.
23. Hagberg,A., Swart,P. and S Chult,D. (2008) In: *Exploring network structure, dynamics, and function using NetworkX*. Technical report,

Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

24. Madden,T. (2013) The BLAST sequence analysis tool. In: *The NCBI Handbook [Internet]*. 2nd edn, National Center for Biotechnology Information (US). Bethesda, MD (United States)
25. Hubley,R., Finn,R.D., Clements,J., Eddy,S.R., Jones,T.A., Bao,W., Smit,A.F. and Wheeler,T.J. (2016) The Dfam database of repetitive DNA families. *Nucleic Acids Res.*, **44**, D81–D89.
26. Yu,G., Wang,L.-G., Han,Y. and He,Q.-Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics*, **16**, 284–287.
27. Sondka,Z., Bamford,S., Cole,C.G., Ward,S.A., Dunham,I. and Forbes,S.A. (2018) The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer*, **18**, 696–705.
28. Mayakonda,A., Lin,D.-C., Assenov,Y., Plass,C. and Koeffler,H.P. (2018) Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.*, **28**, 1747–1756.
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.* (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71.
30. Newman,A.M., Liu,C.L., Green,M.R., Gentles,A.J., Feng,W., Xu,Y., Hoang,C.D., Diehn,M. and Alizadeh,A.A. (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, **12**, 453–457.
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.* (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.*, **47**, W636–W641.
32. Jurtz,V., Paul,S., Andreatta,M., Marcatili,P., Peters,B. and Nielsen,M. (2017) NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.*, **199**, 3360–3368.
33. Friedman,J., Hastie,T. and Tibshirani,R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1.
34. Kaplan,E.L. and Meier,P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, **53**, 457–481.
35. 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.
36. Robinson,J.T., Thorvaldsdóttir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat Biotechnol.*, **29**, 24–26.
37. 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.* (2018) The immune landscape of cancer. *Immunity*, **48**, 812–830.
38. 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.* (2020) Profiling immunoglobulin repertoires across multiple human tissues using RNA sequencing. *Nat. Commun.*, **11**, 3126.
39. Lefranc,M.-P., Giudicelli,V., Ginestoux,C., Jabado-Michaloud,J., Folch,G., Bellahcene,F., Wu,Y., Gemrot,E., Brochet,X., Lane,J. *et al.* (2009) IMGT®, the international ImMunoGeneTics information system®. *Nucleic Acids Res.*, **37**, D1006–D1012.
40. Imielinski,M., Guo,G. and Meyerson,M. (2017) Insertions and deletions target lineage-defining genes in human cancers. *Cell*, **168**, 460–472.
41. Swallow,D.M., Gendler,S., Griffiths,B., Corney,G., Taylor-Papadimitriou,J. and Bramwell,M.E. (1987) The human tumour-associated epithelial mucins are coded by an expressed hypervariable gene locus PUM. *Nature*, **328**, 82–84.
42. Ning,Y., Zheng,H., Zhan,Y., Liu,S., Zang,H., Luo,J., Wen,Q., Fan,S. and et,al. (2020) Comprehensive analysis of the mechanism and treatment significance of Mucins in lung cancer. *J. Exp. Clin. Cancer Res.*, **39**, 162.
43. Aithal,A., Rauth,S., Kshirsagar,P., Shah,A., Lakshmanan,I., Junker,W.M., Jain,M., Ponnusamy,M.P. and Batra,S.K. (2018) MUC16 as a novel target for cancer therapy. *Exp. Opin. Ther. Pat.*, **22**, 675–686.
44. Wang,X.-M., Zhang,Z., Pan,L.-H., Cao,X.-C. and Xiao,C. (2019) KRT19 and CEACAM5 mRNA-marked circulated tumor cells indicate unfavorable prognosis of breast cancer patients. *Breast Cancer Res. Tr.*, **174**, 375–385.
45. 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.* (2017) 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 Immunol. Immun.*, **66**, 1425–1436.
46. 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.* (2012) Landscape of somatic retrotransposition in human cancers. *Science*, **337**, 967–971.
47. 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.
48. Levine,A.J., Ting,D.T. and Greenbaum,B.D. (2016) P53 and the defenses against genome instability caused by transposons and repetitive elements. *Bioessays*, **38**, 508–513.
49. Zhang,X., Zhang,R. and Yu,J. (2020) New Understanding of the Relevant Role of LINE-1 Retrotransposition in Human Disease and Immune Modulation. *Front. Cell Dev. Biol.*, **8**, 657.
50. Rangwala,S.H., Zhang,L. and Kazazian,H.H. (2009) Many LINE1 elements contribute to the transcriptome of human somatic cells. *Genome Biol.*, **10**, R100.
51. Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.
52. Jung,H., Choi,J.K. and Lee,E.A. (2018) Immune signatures correlate with L1 retrotransposition in gastrointestinal cancers. *Genome Res.*, **28**, 1136–1146.
53. Cassotta,A., Goldstein,J.D., Durini,G., Jarrossay,D., Baggi Menozzi,F., Venditti,M., Russo,A., Falcone,M., Lanzavecchia,A., Gagliardi,M.C. and et,al. (2021) Broadly reactive human CD4+ T cells against Enterobacteriaceae are found in the naïve repertoire and are clonally expanded in the memory repertoire. *Eur. J. Immunol.*, **51**, 648–661.
54. Wright,J.R. (2004) Host defense functions of pulmonary surfactant. *Neonatology*, **85**, 326–332.
55. Smith,C.C., Selitsky,S.R., Chai,S., Armistead,P.M., Vincent,B.G. and Serody,J.S. (2019) Alternative tumour-specific antigens. *Nat. Rev. Cancer*, **19**, 465–478.
56. Gopanenko,A.V., Kosobokova,E.N. and Kosorukov,V.S. (2020) Main strategies for the identification of neoantigens. *Cancers*, **12**, 2879.
57. 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.* (2021) Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-021-01021-3.
58. 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.* (2018) Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.*, **10**, eaau5516.
59. Sizemore,G.M., Pitarresi,J.R., Balakrishnan,S. and Ostrowski,M.C. (2017) The ETS family of oncogenic transcription factors in solid tumours. *Nat. Rev. Cancer*, **17**, 337–351.
60. Hahn,S., Ugurel,S., Hanschmann,K.-M., Strobel,H., Tondera,C., Schadendorf,D., Löwer,J. and Löwer,R. (2008) Serological response to human endogenous retrovirus K in melanoma patients correlates with survival probability. *AIDS Res. Hum. Retrov.*, **24**, 717–723.
61. Zhao,J., Rycaj,K., Geng,S., Li,M., Plummer,J.B., Yin,B., Liu,H., Xu,X., Zhang,Y., Yan,Y. *et al.* (2011) Expression of human endogenous retrovirus type K envelope protein is a novel candidate prognostic marker for human breast cancer. *Genes Cancer*, **2**, 914–922.
62. Golkaram,M., Salmans,M.L., Kaplan,S., Vijayaraghavan,R., Martins,M., Khan,N., Garbutt,C., Wise,A., Yao,J., Casimiro,S. *et al.* (2021) HERVs establish a distinct molecular subtype in stage II/III colorectal cancer with poor outcome. *NPJ Genom. Med.*, **6**, 13.
63. Hayer,K.E., Pizarro,A., Lahens,N.F., Hogenesch,J.B. and Grant,G.R. (2015) Benchmark analysis of algorithms for

determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics*, **31**, 3938–3945.
64. Steijger,T., Abril,J.F., Engström,P.G., Kokocinski,F., Hubbard,T.J., Guigó,R., Harrow,J. and Bertone,P. (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, **10**, 1177–1184.
65. Kazemian,M., Ren,M., Lin,J.-X., Liao,W., Spolski,R. and Leonard,W.J. (2015) Comprehensive assembly of novel transcripts from unmapped human RNA-Seq data and their association with cancer. *Mol. Syst. Biol.*, **11**, 826.
66. Cmero,M., Schmidt,B., Majewski,I.J., Ekert,P.G., Oshlack,A. and Davidson,N.M. (2021) MINTIE: identifying novel structural and splice variants in transcriptomes using RNA-seq data. *Genome Biol.*, **22**, 296.
67. Nguyen,H.T., Xue,H., Firlej,V., Ponty,Y., Gallopin,M. and Gautheret,D. (2021) Reference-free transcriptome signatures for prostate cancer prognosis. *BMC Cancer*, **21**, 394.
68. Lorenzi,C., Barriere,S., Villemin,J.-P., Bretones,L.D., Mancheron,A. and Ritchie,W. (2020) iMOKA: k-mer based software to analyze large collections of sequencing data. *Genome Biol.*, **21**, 261.
69. Thomas,A., Barriere,S., Broseus,L., Brooke,J., Lorenzi,C., Villemin,J.-P., Beurier,G., Sabatier,R., Reynes,C., Mancheron,A. *et al.* (2019) GECKO is a genetic algorithm to classify and explore high throughput sequencing data. *Commun. Biol.*, **2**, 222.