Kmerator Suite: design of specific k -mer signatures and automatic metadata discovery in large RNA-seq datasets

Sébastien Riquier, Chloé Bessiere, Benoit Guibert, Anne-Laure Bouge, Anthony Boureux, Florence Ruffle, Jérôme Audoux, Nicolas Gilbert, Haoliang Xue, Daniel Gautheret, et al.

Published online 23 June 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 3 **1**
https://doi.org/10.1093/nargab/lqab058

Kmerator Suite: design of specific *k*-mer signatures and automatic metadata discovery in large RNA-seq datasets

Sébastien Riquier[1,†], Chloé Bessiere[1,†], Benoit Guibert[1], Anne-Laure Bouge[2], Anthony Boureux[1], Florence Ruffle[1], Jérôme Audoux[2], Nicolas Gilbert[1], Haoliang Xue[3], Daniel Gautheret[3] and Thérèse Commes[1,*]

[1]IRMB, University of Montpellier, INSERM, 80 rue Augustin Fliche, 34295, Montpellier, France, [2]SeqOne, 34000, Montpellier, France and [3]Institute for Integrative Biology of the Cell, CEA, CNRS, Université Paris-Saclay, 91198, Gif sur Yvette, France

Received November 17, 2020; Revised May 10, 2021; Editorial Decision May 31, 2021; Accepted June 17, 2021

ABSTRACT

The huge body of publicly available RNA-sequencing (RNA-seq) libraries is a treasure of functional information allowing to quantify the expression of known or novel transcripts in tissues. However, transcript quantification commonly relies on alignment methods requiring a lot of computational resources and processing time, which does not scale easily to large datasets. *K*-mer decomposition constitutes a new way to process RNA-seq data for the identification of transcriptional signatures, as *k*-mers can be used to quantify accurately gene expression in a less resource-consuming way. We present the Kmerator Suite, a set of three tools designed to extract specific *k*-mer signatures, quantify these *k*-mers into RNA-seq datasets and quickly visualize large dataset characteristics. The core tool, Kmerator, produces specific *k*-mers for 97% of human genes, enabling the measure of gene expression with high accuracy in simulated datasets. KmerExploR, a direct application of Kmerator, uses a set of predictor gene-specific *k*-mers to infer metadata including library protocol, sample features or contaminations from RNA-seq datasets. KmerExploR results are visualized through a user-friendly interface. Moreover, we demonstrate that the Kmerator Suite can be used for advanced queries targeting known or new biomarkers such as mutations, gene fusions or long non-coding RNAs for human health applications.

INTRODUCTION

Publicly available human RNA-sequencing (RNA-seq) datasets are precious resources for biomedical research. RNA-seq data are widely used to identify actively transcribed genes, quantify gene or transcript expression, identify new fusion transcripts or identify alternative splicing or mutation events. The search for specific transcriptional events or RNAs across large-scale data has become essential in precision medicine. Advanced tools such as recount2 (1) have achieved transcript counts in large datasets, available in an online resource. However, these tools are reference based and only provide counts for precomputed transcripts. An increasing number of studies attempt to analyze in a retrospective fashion the vast repository of RNA-seq data, including normal and pathological conditions, to discover or validate RNA biomarkers for disease diagnosis (2,3).

For this purpose, it is important to select relevant RNA-seq datasets with homogeneous characteristics and sufficient samples among thousands of publicly available files. The reanalysis of RNA-seq datasets poses two major challenges. The first challenge is to filter data series and select the most homogeneous and reliable set of libraries for exploration in the context of incomplete metadata (4). The second challenge is to perform RNA biomarker quantification in reasonable time and with sufficient accuracy to extract biological information in such datasets. Alignment-based methods like STAR (5) and CRAC (6) require significant computational resources, making them inadequate for querying datasets on the order of 100–1000 files for a specific biomarker. Pseudo-alignment algorithms like Kallisto (7) and Salmon (8) are much faster but most commonly use a reference transcriptome far from the real complex biological RNA diversity. This highlights the need for tools enabling fast and specific quantification of candidate se-

*To whom correspondence should be addressed. Tel: +33 4 67330190; Email: therese.commes@inserm.fr
†These authors contributed equally to this work.

quences in a large set of RNA-seq data. Recently, approaches relying on *k*-mers from raw sequence files have emerged and are used for the query of transcriptomic data. These methods require less time and computational resources than common ones and are suited to various biological questions, including the analysis of unannotated and atypical RNA transcriptional events. For instance, Okamura and Kinoshita proposed an ultrafast mRNA quantification method, based on unique *k*-mers, that outperforms conventional approaches (9). Yu *et al*. (10) investigated gene fusion queries of all tumor samples from The Cancer Genome Atlas project using *k*-mer sets. The DEkupl pipeline developed by Audoux *et al*. (11) finds differential events between two groups of RNA-seq data at the *k*-mer level.

Moreover, classical methods fail to interrogate the whole transcriptome complexity as each RNA is the result of a complex chain of events that combines genetic variation, transcription regulation and RNA processing combined with pathological alterations (12). The *k*-mer approach we propose is not an equivalent method compared to the above-mentioned ones, but a new way to explore RNA-seq data that could also be used for in-depth exploration outside the reference.

Although any transcript sequence can be decomposed into *k*-mers, only a subset of these *k*-mers is specific for the transcript. We call this subset the *k*-mer signature. These specific *k*-mers can then be quantified in RNA-seq raw data, making it quick and easy to measure the candidate transcript expression level in a wide range of RNA-seq datasets.

In this paper, we present the Kmerator Suite, a set of three tools designed to (i) extract *k*-mer signatures from transcripts, (ii) quantify these *k*-mers into RNA-seq datasets and (iii) visualize large RNA-seq dataset characteristics using precomputed signatures. The core of this suite is Kmerator, which generates *k*-mer signatures specific for genes or transcripts. The second tool, countTags, is used to quantify selected *k*-mers across raw RNA-seq files. We first tested the performance of Kmerator + countTags over the whole transcriptome and showed that *k*-mer signature quantification results were close to simulated count data. The third tool, KmerExploR, demonstrates the capacity of the Kmerator + countTags pipeline combined to a set of predefined *k*-mer signatures, to perform metadata extraction from raw RNA-seq data. KmerExploR extracts sample characteristics related to the sequencing protocol (ribosomal depletion, polyA+, strand-specific protocol, 5′/3′ bias, etc.), tissue origin (sex) and possible contaminations (mycoplasma, virus, other species or cell lines). Such high-level quality control procedures are valuable as a screening tool before analyzing datasets of uncertain quality, such as public datasets. KmerExploR can also be used in advanced applications to look for user-defined transcripts resulting from mutated alleles or gene fusions in RNA-seq datasets.

MATERIALS AND METHODS

Kmerator: *k*-mer signature identification

An overview of the Kmerator Suite is provided in Figure 1A. Kmerator is a tool designed for the prediction of specific *k*-mers from input sequences, considering a reference genome and an Ensembl-like fasta transcriptome (see Figure 1A and Supplementary Figure S1A). It is implemented in Julia programming language (https://julialang.org) and distributed with GitHub (https://github.com/Transipedia/kmerator). Kmerator strictly depends on a reference genome [fasta or Jellyfish (13) index format] and on an Ensembl fasta format transcriptome, to define a *k*-mer as specific or not, depending on the number of occurrences on each reference. The reference genome and transcriptome fasta, used in this paper, have been downloaded here: https://www.ensembl.org/info/data/ftp/index.html. The procedure also needs a list of gene/transcript Ensembl IDs (or gene symbols) or sequences in fasta format from which Kmerator will extract specific *k*-mers. As shown in Supplementary Figure S1A, Kmerator first uses the Jellyfish software to index and count *k*-mers from the reference genome and transcriptome. For both genome and transcriptome fasta files, Jellyfish produces a hash table including all possible *k*-mers and their number of occurrences. These hash tables are stored for further querying. Second, using Jellyfish query, Kmerator generates, for each input gene/transcript, the list of *k*-mers derived from this sequence and their corresponding genome and transcriptome counts. These *k*-mers are then filtered according to the following criteria: (i) only *k*-mers associated with a biological event (transcript or gene, splice variant, chimeric RNA, circular RNA, etc.) are retained and (ii) *k*-mers must be specific according to Kmerator rules (see Figure 1C and Supplementary Figure S1A). Indeed, Kmerator includes three different levels of specificity (–level option), ‘gene’, ‘transcript’ and ‘chimera’, detailed below:

- Gene level specific *k*-mers are found zero (to include *k*-mers containing splicing junctions) or one time in the reference genome. They are also present in the reference transcriptome in at least one isoform transcript sequence. If we want to select only *k*-mers matching at least *n* isoforms on a total of *N*, a threshold can be set to the proportion of isoforms n/N the *k*-mer has to be specific to, using the –threshold option.
- Transcript level specific *k*-mers are found zero or one time in the reference genome. They also match the reference transcriptome only once (transcript specificity). If the candidate transcript is not annotated, the –unannotated option must be added. In this case, *k*-mers found zero or one time in the reference genome and that do not map to the reference transcriptome are retained.
- Chimera level specific *k*-mers are found neither in the reference genome nor in the reference transcriptome. This level must be combined to the –unannotated option. Kmerator outputs the list of specific *k*-mers (also called *k*-mer gene/transcript signature) according to the chosen parameters in fasta format, for each input sequence.

Kmerator command line options. The *k*-mer length can be set using the –length option. In the present study, we used the default 31 nt *k*-mer length according to the literature (11). The level of specificity is chosen among ‘gene’, ‘transcript’ and ‘chimera’ with the –level option. When using the gene level, the APPRIS database (http://appris.bioinfo.cnio.es) can be queried to identify the ‘PRINCIPAL’ transcript,

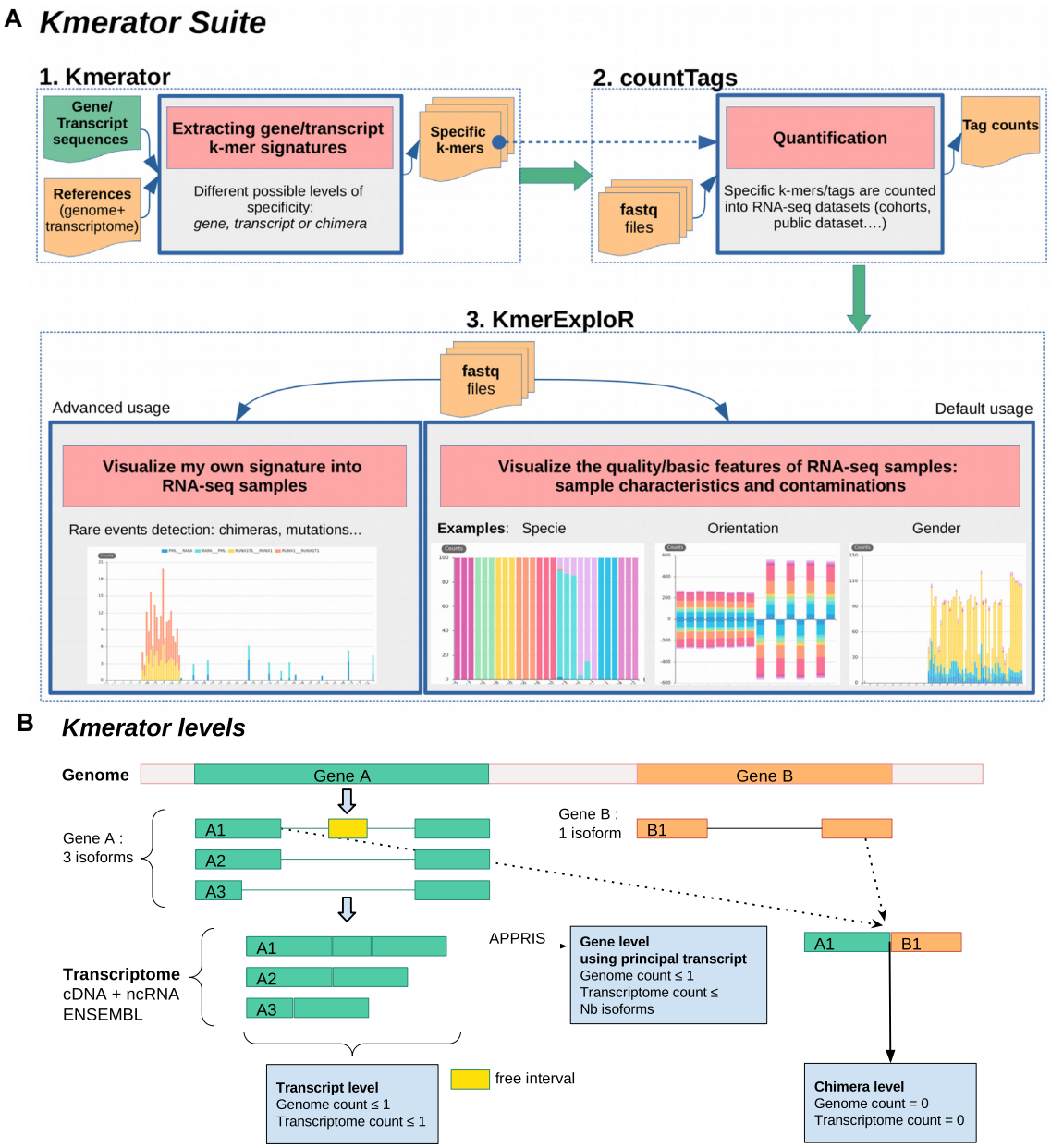

Figure 1. Kmerator Suite and Kmerator levels: definitions. **(A)** The Kmerator Suite is a set of three tools: (1) Kmerator extracts gene/transcript *k*-mer signatures. It takes as input a reference genome and a reference transcriptome + a list of gene or transcript sequences to extract specific *k*-mers from. The output is a set of fasta files (one per input gene/transcript sequence) with the specific *k*-mers. (2) countTags quantifies input *k*-mers in a set of input sequencing raw files (fastq files) and outputs a count table. (3) KmerExploR is a particular application of Kmerator/countTags to visualize input RNA-seq dataset (set of fastq files) characteristics. The default usage includes characteristics related to the sequencing protocol (ribosomal depletion, polyA+, strand-specific protocol, 5′/3′ bias), tissue origin (sex) and possible contaminations (mycoplasma, virus, other species or HeLa cell line). Users can also visualize their own signatures with the advanced usage. Details are given in the text and Supplementary Figure S1. **(B)** Kmerator extracts gene/transcript *k*-mer signatures with three possible levels of stringency. This figure describes how the different levels are defined (transcript, gene or chimera) for two example genes A and B. Example gene A has three isoforms: A1, A2 and A3. A1 is the only one with a free interval, i.e. a region not covered by other isoforms, and is defined as the principal transcript (APPRIS database). Therefore, at the transcript level, each transcript has its own specific *k*-mer set, depending on its coverage with other isoforms. At the gene level, the principal transcript defined with the APPRIS database is used, and specific *k*-mers can be common to several isoforms. At the chimera level (example of A1–B1 fusion), the *k*-mer is not described in annotations.

using the –appris option. APPRIS defines as the 'PRINCIPAL' isoform a CDS (coding sequence) variant for each gene, based on the range of protein features. When this option is not used or no principal sequence is given by APPRIS [i.e. for long non-coding RNA (lncRNA)], the isoform with the longest sequence is kept. In this study, we always used the gene level in combination with the –appris option.

Kmerator usage on the entire transcriptome for performance assessment. Kmerator was tested to extract *k*-mer signatures from the whole human Ensembl transcriptome (combination of cDNA and ncRNA fasta files, version 91). The Ensembl reference transcriptome was filtered to remove any transcript with alternate loci (labels with '_alt') and have been processed by Kmerator at both transcript (i.e. 199 181 transcripts) and gene (54 874 genes) levels with the –appris option previously described. At the transcript level, 62 transcripts have been ignored due to their length inferior to the *k*-mer length (31 nt). The processing to generate the specific *k*-mers on the whole transcriptome has been completed in <3 days at the gene level (88 003 855 *k*-mers) and 24 h at the transcript level (69 760 957 *k*-mers), using a LINUX server with 30 computing cores and 20 GB hard disk space. This step has to be done only one time for one chosen reference transcriptome. Once we have all the annotated transcript *k*-mer signatures, we can rapidly quantify them in any RNA-seq data.

K-mer counting and expression quantification

Simulated data. To test the precision of *k*-mer quantification, we created a set of 10 simulated RNA-seq data for which we have the exact counts. We first used the R *compcodeR* package (14) and the 'generateSyntheticData' function to simulate a count matrix with two conditions with five samples in each (samples.per.cond = 5). Each line of this matrix corresponds to a transcript of the Ensembl v91 annotation. Counts of transcripts with a length equal or inferior to 200 nt were not simulated. To highlight the quantification process, we increased the number of differentially expressed genes (n.diffexp = 10 000) with balanced over- and underexpressed fractions (fraction.upregulated = 0.5) and with authorized different dispersions between the conditions (between.group.diffdisp = TRUE, fraction.non.overdispersed = 0). Besides, we set the sequencing depth by RNA-seq file to 100 million reads (seq depth = 100 000 000) and we did not filter low counts (filter.threshold.total = 0). Providing this data frame and the Ensembl reference transcriptome, we used the 'simulate_experiment_countmat' function, from *polyester* R package (15), to generate paired-end and strand-specific (fr fashion) RNA-seq reads in fasta format. Finally, the fasta files have been converted to fastq.gz format using seqtk (https://github.com/lh3/seqtk).

countTags. *K*-mers designed by Kmerator on the whole transcriptome were counted into the 10 simulated RNA-seq data. For this purpose, the list of *k*-mers was submitted to countTags (https://github.com/Transipedia/countTags), a tool written in C language (see Figure 1A). countTags searches for short sequences (<32 nt) and their reverse complement with an exact match in fastq files and counts their occurrences. We used a *k*-mer length of 31 nt (–*k* 31) and the paired-end option (–paired), and we also used the countTags normalization option to normalize *k*-mer counts per billion of *k*-mers present in the dataset, using the –kbp option. As many specific *k*-mers are associated with one single transcript/gene, we computed the mean *k*-mer count by transcript/gene.

Comparison with Kallisto. We compared the Kmerator + countTags pipeline with Kallisto regarding the performances in transcript/gene expression quantification on simulated data detailed above. As our pipeline cannot quantify genes/transcripts without specific *k*-mers, we limited Kallisto quantification to the genes/transcripts having specific *k*-mers. Kallisto 0.43.1 (7) was run using the –fr-stranded option with the Ensembl v91 annotation file. For each pipeline, TPM (transcripts per million) counts were compared to true normalized TPM using the Spearman's correlation, either at the transcript level or at the gene level. Counts estimated by Kallisto were merged at the gene level by summing normalized transcript counts.

KmerExploR: exploring large RNA-seq datasets

KmerExploR is a command line tool powered by the backend pipeline Kmerator + countTags. KmerExploR provides *k*-mer quantification results in RNA-seq samples as a graphical and user-friendly html interface (see Figure 1A). To deal with data heterogeneity and the weaknesses of RNA-seq technology, we developed a turnkey application using KmerExploR. Characterization of a requested RNA-seq dataset can be improved with the quantification of selected genes (predictor genes) via the Kmerator + countTags pipeline. Predictor genes and their corresponding specific *k*-mers are included in KmerExploR and have been selected based on the literature to answer specific biological questions:

- Are my RNA-seq data based on polyA selection protocol or ribo-depletion?
- Are my RNA-seq libraries stranded or not?
- What is/are the sex corresponding to my samples?
- Is there a read coverage bias from 5′ to 3′ end along my dataset transcripts?
- Are my RNA-seq data contaminated by HeLa (presence of HeLa-derived human papillomavirus 18), mycoplasmas or other viruses such as hepatitis B virus?
- What is/are the species present in my samples?

Implementation. KmerExploR is a command line tool written in python 3. It can be installed on a server or on a personal computer from GitHub or with pip command (see https://github.com/Transipedia/kmerexplor). No additional modules are required. KmerExploR does not need a lot of memory and can be launched from a laptop. Indeed, for a common analysis of 36 paired-end samples (80 GB of fastq files), it takes 250 MB of memory (RAM per core) and 24 min. In comparison, the popular useful and complementary QC tool fastQC (https://qubeshub.org/resources/fastqc) takes 3300 MB of memory (RAM per core) and 15 min. KmerExploR includes countTags, described above.

From input fastq files, KmerExploR runs countTags, with a multithreading option, to quantify built-in *k*-mer selection associated with each predictor gene. The detailed diagram is shown in Supplementary Figure S1B. KmerExploR can also directly take countTags output files, as for large datasets it could be useful to separately run countTags on a cluster, for example. KmerExploR outputs an html file with css and javascript in separate files, using the echartsjs library to display user-friendly and graphical information (https://echarts.apache.org/en/index.html). Categories to show are described either in the built-in config file or in the user personal config file. KmerExploR also produces a tabulated text file with mean counts for each predictor gene in each category (rows) and in each sample (columns).

Predictor gene selection. We selected a subset of housekeeping genes from the list previously published by Eisenberg and Levanon (16) as well as some widely expressed histone genes that produce non-polyadenylated transcripts barely detected in polyA+ RNA-seq (see Table 1). We also selected specific genes from chromosome Y that have a ubiquitous expression, from Maan *et al.*'s publication (17). For these different sets of genes, we designed specific *k*-mers using Kmerator at the gene level and also computed the *k*-mer reversed complementary counterparts for the orientation category. Housekeeping genes' ubiquitous expression profile in various tissues, chromosome Y genes' specific expression pattern in male tissues and histone genes' low expression in polyA+ RNA-seq samples have been validated by exploring the GTEx database (https://www.gtexportal.org) (see Supplementary Figure S2).

For the detection of 5′/3′-end biases, we used the specific *k*-mers from ubiquitous genes (orientation set) and individually attributed them to their corresponding region, 5′ untranslated region (UTR), 3′ UTR or CDS, depending on their position in the principal transcript, according to the APPRIS database. For that purpose, we used Ensembl annotations with the biomaRt R package that gives the information of the UTR and CDS regions for each transcript. We searched the *k*-mers in transcript CDS and UTR sequences to label them by region. For mycoplasma tag selection, we first selected the most frequent mycoplasma found in cell contamination according to Drexler and Uphoff (18). We then downloaded ribosomal RNA (rRNA) sequences of the six selected mycoplasma species from the SILVA database v132 (19), which provides updated and curated rRNA sequences from Bacteria, Archaea and Eukaryota. Some species have several associated strains and therefore, several rRNA sequences. We have included them all for the *k*-mer design. For HeLa detection, we selected HPV-18 transcripts reported to be expressed in HeLa cells (20). Using UGENE software (21), we manually modified these transcripts to match the mutations reported as HeLa specific in the Cantalupo *et al.* study (20). We then defined sequences taking 30 nt on both sides of each mutation, before passing them to Kmerator to keep only *k*-mers not present in the human genome and transcriptome. For species identification, we selected those principally found in the SRA database. We then downloaded mitochondrially encoded cytochrome *c* oxidase I (MT-CO1) human gene sequence and its orthologs in each of the selected species, using the corresponding animal reference genome and transcriptome sequences (Ensembl v91 for each). Finally, sequences of virus genomes have been downloaded from RefSeq using the common virus list provided by Uphoff *et al.* (22). All these potential contamination sequences were used to produce specific *k*-mers using Kmerator at the chimera level, to select tags that can be found neither in the human reference genome nor in the transcriptome. For the advanced application of KmerExploR, we designed *k*-mers corresponding to new or rare transcriptional events detected in the Leucegene dataset (https://leucegene.ca/). For chimera detection, we used two well-known fusion RNA examples associated with chromosomal translocation and their reciprocal counterparts [RUNX1–RUNXT1 t(*x*,21) RUNXT1–RUNX1, PML–RARA t(15,17) and RARA–PML]. Specific *k*-mers are designed with Kmerator on 60 bp sequences spanning the junction. For mutation detection, we manually designed 31 bp *k*-mers centered on the mutation for reference and alternative sequences of three genes currently used in acute myeloid leukemia (AML) diagnosis: TET2, KRAS and CEBPA. We finally designed *k*-mers with Kmerator at the transcript level for a new lncRNA previously published in (23) as NONE 'chr2-p21'.

RNA-seq dataset. In this paper, we illustrated KmerExploR output on several datasets, depending on the biological question, all described in Supplementary Table S1. Characteristics related to RNA-seq protocol, which we call basic features, are tested on 103 paired-end samples from ENCODE (Dataset-FEATURES). For the contaminations part, we used the 33 single-read samples from the PRJNA153913 study (24) previously described as highly contaminated by mycoplasma (Dataset-MYCO) (25). We also selected three public RNA-seq samples by species to check the relevance of our species-specific *k*-mers (Dataset-SPECIES). HeLa contamination was tested in three cervical cancer CCLE (Cancer Cell Line Encyclopedia) cell lines: one HeLa and two negative controls (Dataset-HELA-CLE). Finally, for virus detection we used 19 samples from the CCLE dataset reported by Uphoff *et al.* (22) as contaminated by viruses and three control non-contaminated cell lines also included in the Uphoff *et al.* study (Dataset-VIRUS-CCLE).

RESULTS

Kmerator performances

To assess the Kmerator methodology, we first extracted *k*-mer signatures from all the human Ensembl transcriptome (i.e. 199 181 transcripts) and genes (i.e. 54 874 coding and non-coding genes). We were able to identify specific *k*-mers ($k = 31$ nt) for 83% of human transcripts and 97% of human genes as shown in Figure 2A and B.

This way, the transcriptome information has been almost entirely summarized by 69 760 957 *k*-mers at the transcript level and 88 003 855 *k*-mers at the gene level, corresponding to 23.8% and 30% of the total number of *k*-mers in the reference transcriptome, respectively. The attribution of specific *k*-mers at the gene and transcript levels is fundamentally different: whereas the gene level (–appris option) accepts specific *k*-mers shared with other isoforms, the transcript

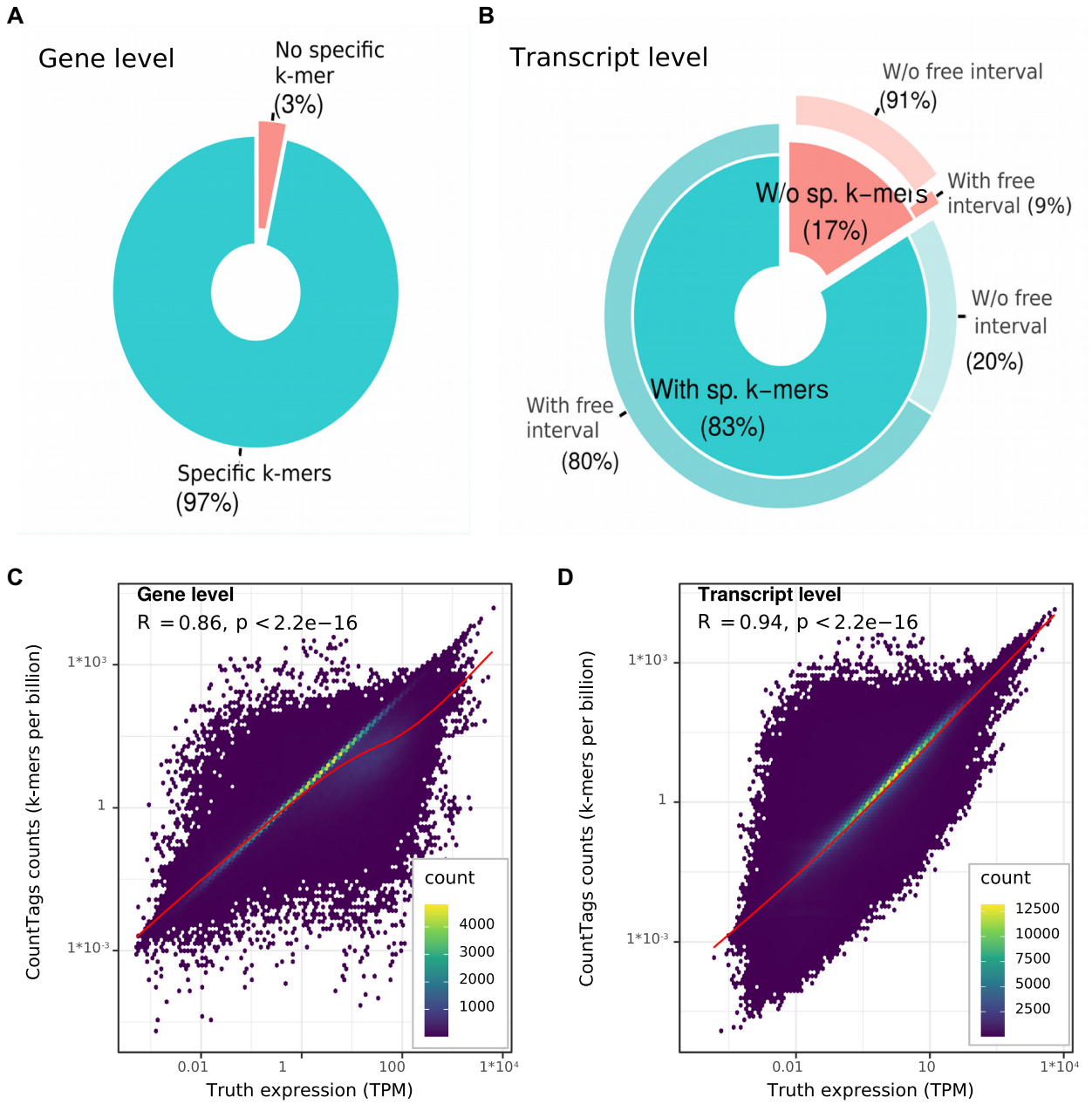

Figure 2. Kmerator performances on the whole transcriptome. We extracted *k*-mer signatures from all the human Ensembl transcriptome v91 at both gene (54 874 coding and non-coding genes, left) and transcript (i.e. 199 181 transcripts, right) levels. **(A)** The first pie chart represents the proportion of genes having specific *k*-mers (turquoise) versus those without specific *k*-mers (red). **(B)** In the same way, we represented the proportion of transcripts having specific *k*-mers (turquoise) or not (red). For these two classes, we looked at the percentage having free intervals, i.e. regions in the transcript not shared with other isoforms (secondary pie). Most of the transcripts lacking specific *k*-mers do not have free intervals (91%). We tested Kmerator sensitivity to quantify simulated data, at both gene **(C)** and transcript **(D)** levels. We represented the *k*-mer counts normalized per billion of *k*-mers in the sample (*Y*-axis) as a function of the true expression in TPM (*X*-axis), on the whole simulated dataset. *R* is the Spearman's correlation coefficient between *k*-mer counts and TPM. Each point on the graph is a transcript and the color scale depends on the transcript density on the graph.

level is more stringent and eliminates each *k*-mer shared by other ones. This explains the higher percentage of transcripts without specific *k*-mer compared to the gene level. To explain the absence of specific *k*-mers for some transcripts, we used BiomaRt genomic intervals to calculate the part of each transcript not covered by other isoforms, considering the strand, and named it 'free interval' (see Figure 1B). As expected, 91% of transcripts without specific *k*-mer have no 'free interval', which means that they are completely covered by other transcripts, thus confirming the validation of the Kmerator process. The set of specific *k*-mers designed with Kmerator strongly depends on the input sequence and on the level of selection. At the gene level, we observed that the length of the input sequence was corre-

lated with the number of designed specific *k*-mers ($R = 0.91$, $P < 2.2e{-}16$; see Supplementary Figure S3A) but not at the transcript level ($R = 0.22$, $P < 2.2e{-}16$; see Supplementary Figure S3B). On the contrary, the transcript level depends on the overlap between the input transcript and the different isoforms. A high number of isoforms is correlated to a low number of specific *k*-mers ($R = 0.79$, $P < 2.2e{-}16$; see Supplementary Figure S3C) and, in addition, the length of free intervals is strongly correlated to the number of specific *k*-mers ($R = 0.94$, $P < 2.2e{-}16$; see Supplementary Figure S3D). Finally, *k*-mer design differs between biotypes and selection levels: the biotypes without specific *k*-mers mainly correspond to small RNAs (miRNAs, rRNA) at the gene level (see Supplementary Figure S3E) and to coding and pseudo-genes at the transcript level (see Supplementary Figure S3F).

The Kmerator Suite has been designed as a new way to explore RNA-seq data and rapidly quantify some chosen sequences called predictors. Kmerator, the first key element of this suite, can extract unique *k*-mers from any sequence. In combination with countTags, it is used to generate large *k*-mer count tables. To situate our tool in relation to a widely used, referenced and benchmarked quantification tool, we tested the Kmerator + countTags pipeline accuracy to estimate gene and transcript expression using simulated data (see the 'Materials and Methods' section). Indeed, using a simulated dataset, for which we have the exact counts, even if it fails to capture the complexity of real data, is the best way to proceed to illustrate our purpose (26). We have run Kmerator and countTags to search for all human gene and transcript expression levels in a set of 10 simulated data. We assessed Spearman's correlation between normalized *k*-mer counts and the ground truth. We used countTags *k*-mer mean count per transcript reported to the total of *k*-mers contained in the input fastq. As shown in Figure 2, the Spearman's correlation factor comparing Kmerator + countTags results to the truth is 0.86 for the gene level (Figure 2C) and 0.94 for the transcript level (see Figure 2D), indicating a highly positive relationship with normalized counts ($P < 2e{-}16$).

Quantification results are comparable when using the Kallisto pseudo-alignment method, despite slightly higher correlation factors (gene and transcript $R = 0.97$; see Supplementary Figure S4A and B). This result is consistent with the recent paper describing Matataki (9), another quantification tool based on *k*-mers. Our pipeline being not specifically dedicated to gene quantification but for rapid exploration of large datasets is accurate enough to evaluate gene and transcript expression levels in RNA-seq data. Interestingly, the precision of Kallisto quantification decreases strongly with transcripts/genes not covered by Kmerator (see Supplementary Figure S4C and D), showing that each protocol using the *k*-mer principle struggles to correctly quantify sequences that do not possess distinctive *k*-mers.

Finally, we tested speed performance of countTags processing time on random subparts of sample simulated data (10 million, 101 nt paired-end reads), while increasing the number of quantified *k*-mers (1/1000/1 million). It appears that processing time remains low compared to alignment-based protocols (~1 min for 10 million reads) and depends on the number of *k*-mers quantified (see Supplementary Figure S4E). These results support an optimized usage of the Kmerator Suite protocol for its primary usage: the research of a limited number of signatures in large RNA-seq datasets.

KmerExploR for inspecting large RNA-seq datasets

We developed KmerExploR to improve the characterization of large RNA-seq datasets using the quantification of selected predictor genes. Predictor genes have been selected based on the literature to answer specific questions (see Table 1). As described in the 'Materials and Methods' section, we first extracted with Kmerator sets of specific *k*-mers from gene sequences and use KmerExploR to count the *k*-mer occurrences in RNA-seq datasets and visualize the results. Here, we present the results obtained with specific datasets (Table 1 and Supplementary Table S1) selected to highlight the rapid control of biological and technical parameters using KmerExploR. The results of the basic features, including sample sex, polyA or ribo-depletion, orientation and 5′/3′ bias, are presented in Figure 3.

As previously described, sample sex is determined by searching for *k*-mers corresponding to genes located on the Y chromosome. The *k*-mer signature clearly separates samples depending on the sex. To help the user classify his samples, we defined, in KmerExploR, a threshold of five *k*-mers per billion, above which we expect with confidence that it is a male. Moreover, Y chromosome gene expression variance between the samples can be explained by the variability of cell types and public RNA-seq experiment parameters, including sequencing depth and methods of RNA extraction and selection. For instance, the four male samples with the lowest expression (ENCFF232KGN, ENCFF434EMO, ENCFF831HCD and ENCFF992HBZ) come from a unique study (ENCSR999ZCI). However, the sex classification is more complicated in case of cancerous data. When we are looking at cancerous RNA-seq cell lines, some samples with male metadata show low Y chromosome-specific gene expression (data not shown). This extreme downregulation of chromosome Y gene expression has already been described in previous studies and strongly associated with cancer risk in men (27).

Gene abundance can be measured in RNA-seq data through sequencing of mRNA or ribo-depleted total RNA samples. The mRNA protocol relies on polyA selection, when the total RNA method is based on rRNA depletion (Ribozero protocol). However, non-polyadenylated transcripts should only be found in data produced using this procedure, when they should barely be detectable in mRNA samples. As the majority of histone transcripts are known to be non-polyadenylated, we used this characteristic first to detect sample contamination by non-polyadenylated RNA, and second to infer from the result the RNA preparation procedure. We first investigated the expression level of all histone genes and retained the most highly expressed according to the literature. Second, we analyzed their expression pattern using the GTEX resource. As RNA-seq from GTEX are exclusively produced from polyA selected RNA samples, we used this database to select histone genes showing the lowest expression levels (see Supplementary Figure S2B). We used this set of histone genes to test a se-

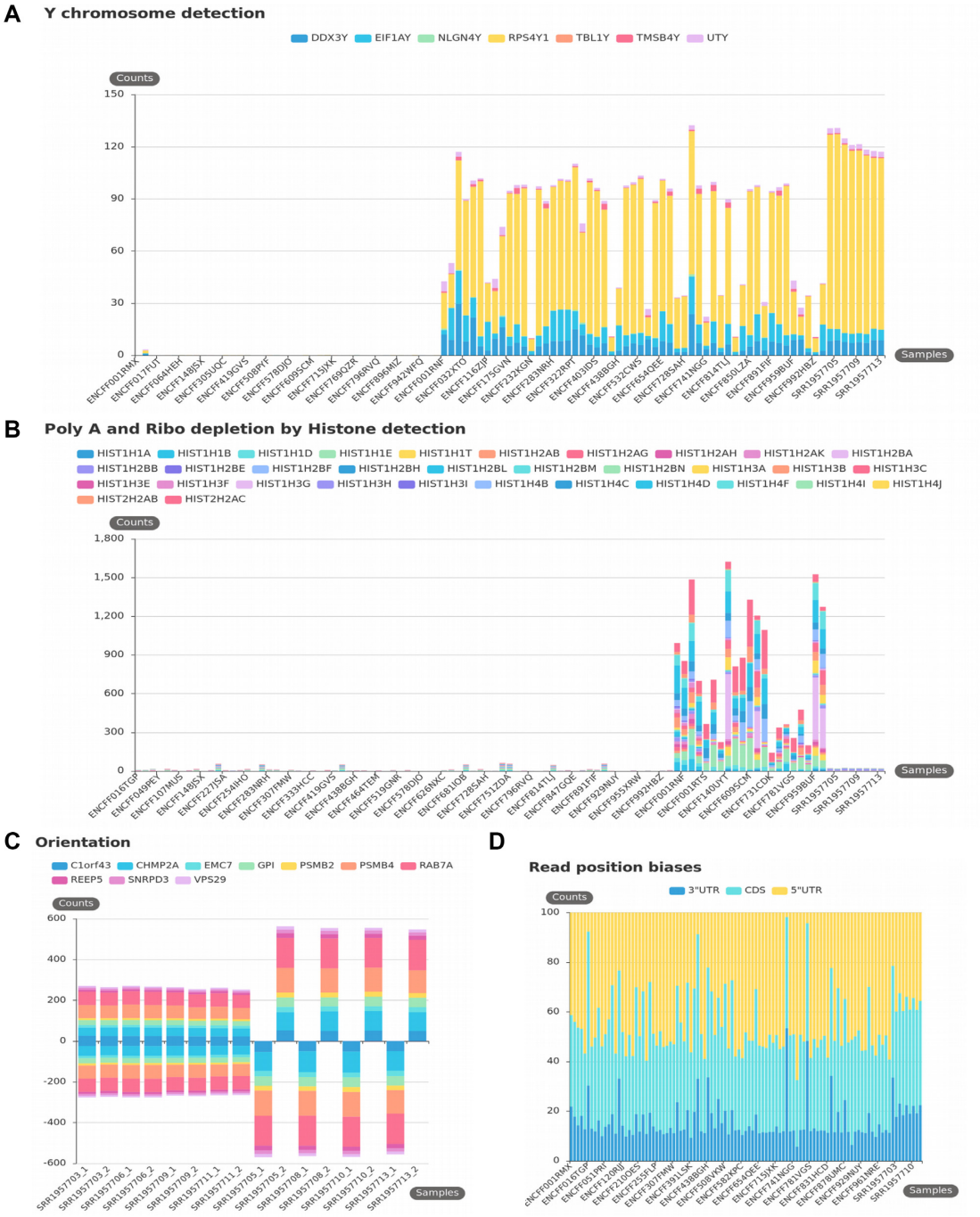

Figure 3. KmerExploR default usage: basic features. All presented bar plots are direct output of KmerExploR and they are generated from the Dataset-FEATURES described in Supplementary Table S1 (103 paired-end ENCODE samples) except for the orientation (C), which is a subset of eight RNA-seq from the Dataset-FEATURES. For each bar plot, the legend lists the set of predictor genes for which *k*-mer mean counts are computed (see also Table 1). Samples are on the *X*-axis. Panels (A), (B) and (C) have the mean *k*-mer counts by gene normalized per billion of *k*-mers on the *Y*-axis. **(A)** Sex determination. Samples are sorted by sex in the order female, then male. **(B)** PolyA+ selection versus ribo-depletion by histone detection. Samples are sorted by protocol in this order: polyA, ribo-depletion, unknown. **(C)** Stranded versus unstranded sequencing protocol. For this category, both fastq files by sample are shown. The first four samples are unstranded and the last four samples are stranded. **(D)** Read position biases along 5′ UTR, 3′ UTR and CDS regions. After computing *k*-mer mean counts by gene, they are summed up by 5′ UTR, 3′ UTR or CDS regions and converted in % (*Y*-axis).

Table 1. List of predictor genes, by category, included in KmerExploR and associated RNA-seq dataset names used in this paper

	Datasets	Predictor genes	Total *k*-mer number	References and details
PolyA/RiboD	Dataset-FEATURES	HIST2H2AC, HIST2H2AB, HIST1H4J, HIST1H4I, HIST1H4F, HIST1H4D, HIST1H4C, HIST1H4B, HIST1H3I, HIST1H3H, HIST1H3G, HIST1H3F, HIST1H3E, HIST1H3C, HIST1H3B, HIST1H3A, HIST1H2BN, HIST1H2BM, HIST1H2BL, HIST1H2BH, HIST1H2BF, HIST1H2BE, HIST1H2BB, HIST1H2BA, HIST1H2AK, HIST1H2AH, HIST1H2AG, HIST1H2AB, HIST1H1T, HIST1H1E, HIST1H1D, HIST1H1B, HIST1H1A	24 512	Supplementary Figure S2
Orientation	Dataset-FEATURES	VPS29, SNRPD3, REEP5, RAB7A, PSMB4, PSMB2, GPI, EMC7, CHMP2A, C1orf43, VPS29_rev, SNRPD3_rev, REEP5_rev, RAB7A_rev, PSMB4_rev, PSMB2_rev, GPI_rev, EMC7_rev, CHMP2A_rev, C1orf43_rev	36 638	Supplementary Figure S2 (16)
Sex	Dataset-FEATURES	UTY, TMSB4Y, TBL1Y, RPS4Y1, NLGN4Y, EIF1AY, DDX3Y	21 996	Supplementary Figure S2 (17)
5′/3′ bias	Dataset-FEATURES	VPS29, SNRPD3, REEP5, RAB7A, PSMB4, PSMB2, GPI, EMC7, CHMP2A, C1orf43	12 705	Supplementary Figure S2 (16)
Mycoplasma	Dataset-MYCO	Mycoplasma_orale, Mycoplasma_hyorhinis, Acholeplasma_laidlawii, Mycoplasma_hominis, Mycoplasma_arginini, Mycoplasma_fermentans	363 025	(18)
Virus	Dataset-VIRUS-CCLE	Human_gammaherpesvirus_4, Human_herpesvirus_4, Human_herpesvirus_8, Murine_leukemia_virus, Hepatitis_C_virus_genotype, Human_immunodeficiency_virus_1, Human_T_lymphotropic_virus_1, Squirrel_monkey_retrovirus, Human_T_lymphotropic_virus_2, Human_papillomavirus_type_92, Hepatitis_B_virus_strain, Human_immunodeficiency_virus_2, MuLV_related_virus_22Rv1/CWR, Bovine_viral_diarrhea_virus	516 882	(22)
HeLa	Dataset-HELA-CCLE	L1_mut7486, L1_mut7258, L1_mut6842, L1_mut6625, L1_mut6460, L1_mut6401, L1_mut5875, E7_mut806, E7_mut751, E6_mut549, E6_mut485, E6_mut287, E6_mut104, E1_mut2269, E1_mut1994, E1_mut1843, E1_mut1807, E1_mut1353, E1_mut1012	589	(20)
Species	Dataset-SPECIES	Homo_sapiens_MT_CO1, Danio_rerio_mt_co1, Zea_mays_COX1, Saccharomyces_cerevisiae_COX1, Rattus_norvegicus_Mt_co1, Mus_musculus_mt_Co1, Gallus_gallus_MT_CO1, Drosophila_melanogaster_mt_CoI, Caenorhabditis_elegans_ctc_3_MTCE, Arabidopsis_thaliana_COX1	12 119	MT-CO1 (and orthologs)
Chimeras	Dataset-LEUCEGENE	PML–RARA, RARA–PML, RUNX1T1–RUNX1, RUNX1–RUNX1T1	724	
lncRNA	Dataset-LEUCEGENE	NONE	78	(23)
Mutations	Dataset-LEUCEGENE	TET2, KRAS, CEBPA	10	

The samples included in each dataset and some metadata are detailed in Supplementary Table S1.

lection of ENCODE samples that metadata indicates either polyA or ribo-depletion protocol (Supplementary Table S1). The results clearly demonstrate differences between libraries prepared by ribo-depletion versus polyA selection for most of the chosen histone genes. We observe histone gene expression variability between the samples demonstrating again the disparity of public data. To help users categorize their RNA-seq data, we defined in the KmerExploR tool a threshold of 200 *k*-mer counts per billion for this category, above which we expect to have only the ribo-depleted samples and not the polyA ones.

Strand-specific and unstranded library preparation are two commonly used preparation protocols that differ by their ability to retain or not RNA strand information. To detect this characteristic from RNA-seq data, we designed *k*-mers, specific for a set of ubiquitous genes (Table 1) and their reverse complement counterparts. *K*-mers on the forward strand are counted as positive and their reverse complement as negative, permitting to determine the orientation of the library. If forward and reverse tags are found in equivalent proportions in the same fastq file, data are considered as 'unstranded'. This leads graphically to a balanced distribution between positive and negative counts. As shown in Figure 3, using this property we are able to clearly separate unstranded and stranded libraries. 5′ to 3′-end bias is a difference of reads' repartition along the tran-

scripts, classically linked to library preparation: incomplete retrotranscription or specific protocols. A comparison between polyA selection and ribo-depletion protocols has previously shown coverage differences across transcripts with a poor 5′-end coverage with the polyA selection method (28). Knowing whether an RNA-seq sample possesses a read repartition bias is critical for isoform detection, or simply to give an indication on the library construction protocol used in large-scale analysis of public data. Using previously described housekeeping genes (Table 1), we have selected different sets of specific *k*-mers depending on their position in the regions defined as 5′ UTR, 3′ UTR and CDS. Figure 3C shows the repartition in percent of these *k*-mers across the Dataset-FEATURES samples. Representing the mean *k*-mer counts as a percentage allows us to evaluate the distribution homogeneity across 5′ UTR, 3′ UTR and CDS regions between the 103 ENCODE samples. This global representation grouping together several genes allows us to identify samples for which one region has a very little coverage. Here, four samples have <10% 5′ UTR coverage (ENCFF734ZAD, ENCFF770NYA, ENCFF419GVS and ENCFF016TGP). We can also notice a better homogeneity of coverage for ribo-depleted samples.

Detection of potential contamination

Different microorganisms like mycoplasma and virus can contaminate samples and cell cultures, modifying the metabolism of the cell and therefore biasing the results of ensuing analysis. Moreover, cancer research has shown that viruses are responsible for ∼20% of human cancers (29). To detect contaminants in RNA-seq data, tools relying on alignment like DecontaMiner (30) or viGEN (31) have been widely used, but the alignment step is time and memory consuming. Exact alignment of *k*-mer-based approaches like Kraken (32) and Taxonomer (33) is an alternative for taxonomic classification. However, these tools are complex and involve data cleaning from adaptors (trimming), the use of internal and external databases and/or probabilistic models for contaminant classification. Using a specific and reduced set of *k*-mers, we have seen an advantage to quickly detect principal contaminants of human cells in RNA-seq datasets, free from alignment methods.

Because mycoplasma is a common source of cell culture sample contamination and could affect host gene expression (25), we choose to control its presence in RNA-seq data. Mycoplasma contamination is evaluated through the detection of specific *k*-mers corresponding to 16S rRNA sequences according to the literature. In fact, Olarerin-George and Hogenesch showed that 90% of the specific mycoplasma-mapped reads from human RNA-seq samples mapped to mycoplasma rRNA. We selected six species that have the highest record rate of detection in cell culture samples (i.e. *Acholeplasma laidlawii*, *Mycoplasma fermentans*, *Mycoplasma hominis*, *Mycoplasma hyorhinis*, *Mycoplasma orale* and *Mycoplasma arginini*) (18) to design our *k*-mers. We used part of an RNA-seq data series previously described as highly contaminated (25) (PRJNA153913 study) to test the relevance of our approach. As shown in Figure 4, we can easily detect the six selected mycoplasma species in some samples, with a prevalence for the *M. hyorhinis* species. Comparing our results with the Olarerin-George and Hogenesch study that used Bowtie 1 alignment and BLAST+ to filter non-specific reads, we were able to confirm mycoplasma rRNA presence for the same samples (see Supplementary Figure S5A). Moreover, we observe a high proportionality between our *k*-mer counts and their read counts on the 33 single-read samples (Dataset-MYCO described in Supplementary Table S1), for each of the six common *Mycoplasma* species.

Viruses are a significant cause of human cancers. Several studies interrogate for the presence of major viruses known to infect human and other mammalian cells (22,34,35). Recently, Uphoff *et al*. screened >300 CCLE RNA-seq data using the Taxonomer interactive tool and compared the results to virus-specific polymerase chain reaction (PCR) analysis, revealing 20 infected cell lines with different viruses (22). To rapidly explore the potential presence of viruses in RNA-seq datasets with our *k*-mer-based approach, we used the same virus reference genomes as described in the Uphoff *et al*. study. Using Kmerator at the chimera level (absent from human annotations), we designed specific *k*-mers for each virus and searched them in a subset of contaminated CCLE data according to Uphoff *et al*. (19 CCLE paired-end samples) and in negative controls (3 CCLE paired-end samples), to validate our protocol ability to detect viruses. Among the contaminated samples, we were able to detect the main viruses in the same samples as in the Uphoff *et al*. study, except for the SRR8615677 sample where we do not detect any virus, as the bovine polyomavirus is not included in our list of common viruses. Our results are shown in Figure 4B and Taxonomer results from the Uphoff *et al*. study are presented in Supplementary Figure S5B. Epstein–Barr virus (EBV) is a very common virus detected in most of the samples; we have therefore analyzed it in more detail in Supplementary Figures S5C (our approach) and S5D (Taxonomer quantification). Indeed, our EBV quantification is correlated with the one from Taxonomer (Pearson's and Spearman's correlation coefficients are 0.99 and 0.89, respectively).

HeLa is the first immortal human cell line, coming from Henrietta Lacks' cancerous tissue samples. Her cancer was triggered by an infection with human papillomavirus type 18 (HPV-18). Nowadays, this cell line is largely used in medical research. Looking for several viruses in public RNA-seq cancer-related databases revealed the presence of HPV-18 sequences in many cancers (36) that closely resemble the HPV-18 viral sequence that is integrated into HeLa cells, suggesting a contamination. Three segments of HPV-18 are integrated into the HeLa genome on chromosome 8 and include the long control region, the E6, E7 and E1 genes, and partial coding regions for the E2 and L1 genes (20). These genes are expressed in HeLa cells, and mutations have been found specifically in HeLa cells. Thus, selecting these mutated HeLa HPV-18 gene-specific *k*-mers and counting them into three CCLE RNA-seq datasets (one positive sample and two negative controls), we validated the accuracy of our selection as we are able to find our *k*-mer selection specifically in HeLa cells. We also checked the results in other HeLa samples from the PRJNA639358 study (see Supplementary Figure S5E).

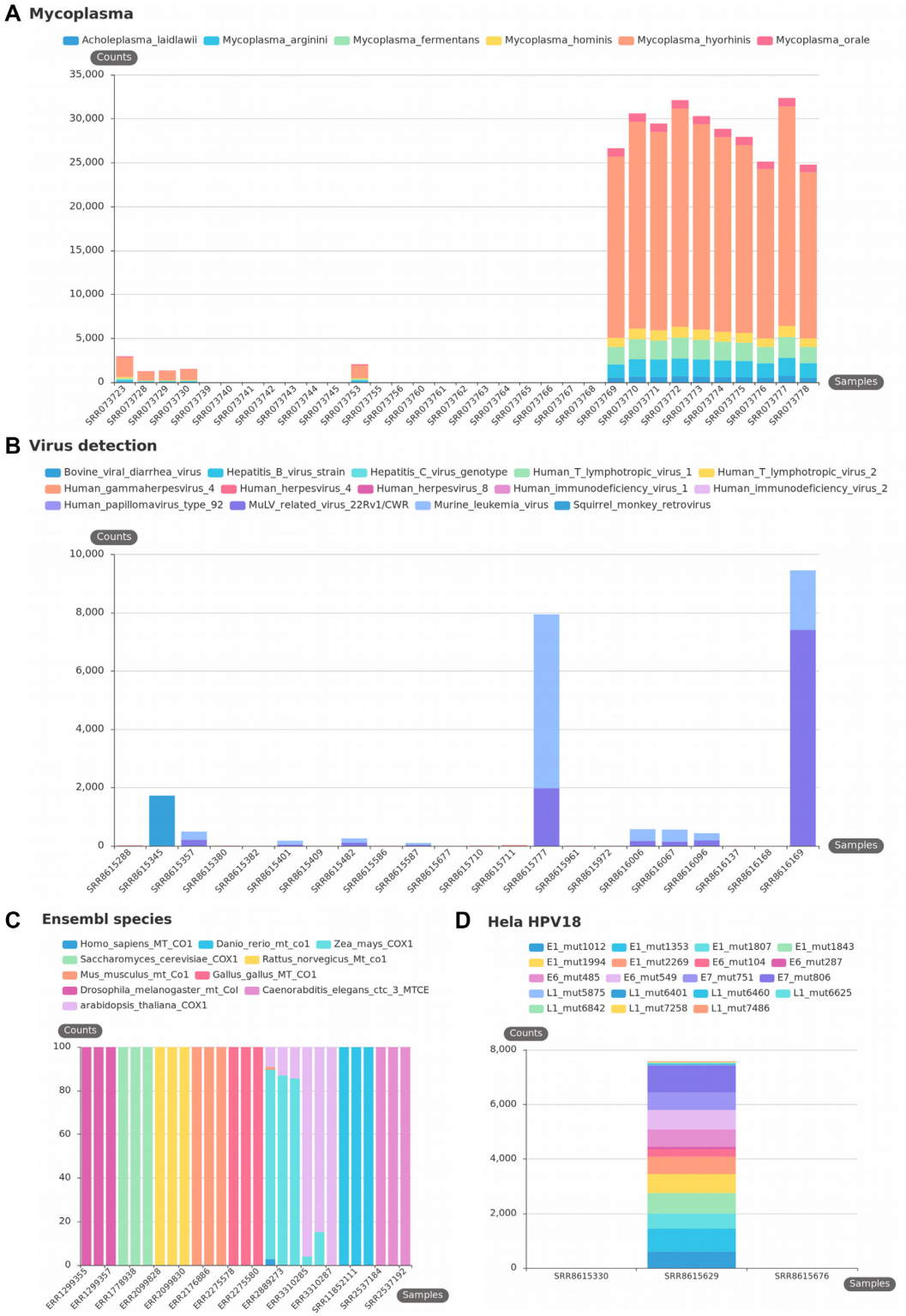

Figure 4. KmerExploR default usage: contaminations. All presented bar plots are direct output of KmerExploR and all bar plot datasets are described in Supplementary Table S1. For each bar plot, the legend lists the set of predictors for which *k*-mer mean counts are computed (details in Table 1). Samples are on the *X*-axis. Panels (A), (B) and (D) have the mean *k*-mer counts by gene normalized per billion of *k*-mers on the *Y*-axis. **(A)** Mycoplasma contamination on the Dataset-MYCO (33 single-read samples). **(B)** Virus detection on the Dataset-VIRUS-CCLE (22 paired-end samples). **(C)** Species determination on the Dataset-SPECIES (27 paired-end samples). For this category, after computing *k*-mer mean counts by species, they are converted in % (*Y*-axis) to avoid big expression differences between species. **(D)** HeLa determination on the Dataset-HELA-CCLE (three paired-end samples). The sample in the middle is a HeLa cell line and the two others are negative controls (SF767 and SiHa cells).

As for HeLa cells, cross-species contamination remains a documented 'danger' for the interpretation of results in molecular biology (37). The probability of mixed cell lines in sample preparation, usage of PCR that can accidentally amplify the wrong piece of DNA, and an unknown probability of error in metadata assignation motivated us to create a quality check to determine the species of an RNA-seq sample. In (38), the usage of mitochondrial DNA for phylogenetic and taxonomic inference was discussed and two extreme viewpoints emerged: using exclusively the mitochondrial DNA or fully excluding it. It appears that mitochondrial DNA does not fully answer or impairs the perspectives of advanced phylogenetics. However, the 'mitochondrial barcode' approach does show an interesting gene marker, MT-CO1 (39), that could be sufficient for a quick check of the species of RNA-seq data. Indeed, this gene is highly expressed and reference sequences from many distinct species of animals are available. Thus, we selected specific *k*-mers with Kmerator, at the gene level, for MT-CO1. We repeated the procedure for MT-CO1 orthologs in different species, principally found in the SRA database, using the appropriate species reference genome and transcriptome. These *k*-mers have been then quantified in three public data by species to check the efficiency of their usage. As shown in Figure 4C, the research of MT-CO1 *k*-mers alone can discriminate most of the common Ensembl species and can be usable for a quick quality check. However, without proper experiments we cannot support its usage with phylogenetically close species.

To conclude, we developed KmerExploR to rapidly control RNA-seq raw data quality and filter samples on unusual profiles or presence of contaminations. KmerExploR is a tool that provides a modular set of analyses like fastQC (https://qubeshub.org/resources/fastqc). It can be used in a complementary way to fastQC analysis to complete missing metadata in public datasets or to give a quick profile of the RNA-seq contents. The modular analysis is based on a *k*-mer selection from predictor genes, included in KmerExploR. The tool can be used to control any human RNA-seq dataset, and it can also be easily modified adding any other modular function.

KmerExploR, an advanced usage for the detection of genomic or transcriptomic events

The above 'checking application' of KmerExploR demonstrated all its potential in the rapid exploration of large public RNA-seq datasets before performing any biological query. However, the KmerExploR tool can also be used in a more advanced way such as biomarker search or discovery in human health. This application is a powerful one as it can compensate for the lack of completeness in genomic or transcriptomic references and we currently know that much important information may be missed by ignoring the unreferenced RNA diversity (12). As a proof of concept, we used a set of *k*-mers designed with Kmerator to identify events outside reference annotations including fusion or chimeric RNA, oncogene mutations and new lncRNA expression. We then applied *k*-mer quantification in a tumoral and a non-tumoral dataset to evaluate the specificity and performance of the approach. The results obtained with a part of the Leucegene cohort are presented in Figure 5.

The selection includes different AML subtypes and normal $CD34^+$ cells as control (Dataset-LEUCEGENE described in Supplementary Table S1). The results obtained with two well-known fusion RNAs associated with chromosomal translocation, RUNX1–RUNXT1 t(x,21) and PML–RARA t(15,17), and their reciprocal counterparts RUNXT1–RUNX1 and RARA–PML are presented in Figure 5A. In this case, the *k*-mers, once designed by Kmerator, are restricted to those spanning the fusion junction with at least 10 nucleotides in gene 1 or gene 2 of the fusion. All the normal $CD34^+$ cells are negative and we only observe an expression in corresponding positive AML subtypes. Figure 5B illustrates the results obtained for mutations in TET2, KRAS and CEBPA genes currently used in AML diagnosis. Once again, we only observe the presence of these mutations in positive samples, demonstrating the high specificity of the approach by *k*-mers. The expression of a new lncRNA was also quickly searched in the Leucegene dataset (see Figure 5C); we observe a homogeneous and low expression in CD34 normal cells compared to a heterogeneous one in AML subtypes. This lncRNA candidate was already described in (23), using for the first time the '*k*-mer concept' for checking new biomarker candidates, and we have demonstrated a restricted expression of the NONE 'chr2-p21' lncRNA in the hematopoietic lineage using the Leucegene and ENCODE datasets. Hence, for lncRNA candidates, following their discovery in a tissue/disease type, their specificity could be easily evaluated through quantification in a wide range of RNA-seq data including normal and pathological conditions as recently described by Riquier *et al*. (40).

In conclusion, the high specific expression of transcriptional events may lead them to be used as biomarkers for biological and health applications, including cell therapy, diagnosis, prognosis or patient follow-up as it is already done with fusion RNAs and mutations.

DISCUSSION

Considering the growing number of RNA-seq data, the use of raw data sequences is an important step to check with RNA-seq protocols or bioinformatic pipelines bias. Here, we demonstrated that the Kmerator Suite is an efficient and useful set of tools to verify RNA-seq quality and control intrinsic method and biological characteristics that often failed in technical description. We also showed that the Kmerator Suite can be used to quantify gene/transcript-specific expression as well as to explore sequence variations at the transcriptional level. In this first version, the tool is adapted to human data Ensembl entry, as main public data are available for this species (164 000 RNA-seq with >30 million reads for *Homo sapiens* in the SRA database). A new implementation with adapted predictors is necessary for other species.

The meta-analyses performed in the present study with KmerExploR are a proof of concept of the procedure potential and could be extended to other biological RNA-seq questioning: (i) to extend the application to an enlarged set of microorganisms including new ones like SARS-Cov2

Figure 5. KmerExploR advanced usage: quantification of transcriptomic events outside the annotations. All presented bar plots are direct output of KmerExploR and they are all generated from the Dataset-LEUCEGENE described in Supplementary Table S1 (131 paired-end samples). This dataset includes normal CD34^{+} cells as control (in green on the *X*-axis) and different AML subtypes (in black on the *X*-axis). For each bar plot, the legend lists the set of predictors for which *k*-mer mean counts normalized per billion (*Y*-axis) are computed. **(A)** Chimera detection. Two well-known fusion RNAs associated with chromosomal translocation and their reciprocal counterparts are presented: RUNX1–RUNXT1 t(*x*,21) and PML–RARA t(15,17). **(B)** Mutation detection. TET2, KRAS and CEBPA genes are used in AML diagnosis. The bar plot shows four different mutations for these genes, detected specifically in some AML samples. The reference allele for each of these mutations is detected in almost all samples. **(C)** New lncRNA detection: NONE 'chr2-p21' lncRNA described in (23). This transcript is expressed in the whole dataset but shows different levels of expression depending on AML subtype.

detection and (ii) to search for immunophenotyping profile in cancer datasets as already published by Mangul *et al*. (41,42). Considering advanced applications, we also demonstrated the potential of *k*-mers to explore gene expression in RNA-seq to reinforce biological questions or biomarker usage and discovery. Moreover, many other requests could be easily considered for annotated gene exploration like gene co-expression, or to compensate the lack of completeness in genomic or transcriptomic references to cover unreferenced RNA diversity and search for new spliced events, intron retention or new transcript categories including circular RNAs. In order to increase the potential of the *k*-mer approach, access to very large-scale datasets like SRA level (164 000 human samples) could be considered with efficient indexing structure development (43).

Finally, we showed that the Kmerator Suite can be used to quantify gene/transcript expression as well as to explore sequence variations at the transcriptional level. The simplicity of specific *k*-mer extraction principle and quantification provide flexibility of usage. Indeed, Kmerator Suite quantification does not use probabilistic methods or expectation–maximization algorithms like in Kallisto (7), Sailfish (44) or RNA-Skim (45). Therefore, the sets of specific *k*-mers for quantification can be created, merged and updated at will, without consequence on the quantification itself. The principle of user-owned collection of signatures of interest that can be searched broadly among datasets is the core of KmerExploR application.

DATA AVAILABILITY

RNA-seq libraries were downloaded from the European Nucleotide Archive of the European Bioinformatics Institute (46). The reference GRCh38 genome and Ensembl v91 transcripts were downloaded from Ensembl. Kmerator is distributed under the MIT license. The Kmerator, KmerExploR and countTags software, documentation and supplementary material presented herein are available from https://github.com/Transipedia/kmerator, https://github.com/Transipedia/kmerexplor and https://github.com/Transipedia/countTags, respectively.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research. The HeLa cell line that was established from her tumor cells without her knowledge or consent in 1951 has made significant contributions to scientific progress and advances in human health. The authors are also grateful to Rayan Chikhi for his comments and corrections.

Author contributions: S.R. and T.C. designed the study. S.R., C.B. and T.C. wrote the manuscript. A.-L.B., N.G. and D.G. were contributors in the design of the study and manuscript corrections. S.R. developed the code of Kmerator, selected and downloaded public datasets, and analyzed data. B.G. and C.B. participated in Kmerator code improvements. C.B. analyzed RNA-seq data and generated figures. B.G. developed KmerExploR code, and generated *k*-mer counting and figures. J.A. and A.B. computed and corrected countTags. F.R. validated the RNA-seq data for mutation and chimeric RNAs, and helped in the interpretation of results. H.X. participated in Kmerator testing and checking. All authors read and approved the final manuscript.

FUNDING

Agence Nationale de la recherche [ANR-10-INBS-09]; Canceropole Grand Ouest [2017-EM24]; Region Occitanie [R19073FF].

Conflict of interest statement. None declared.

REFERENCES

1. Collado-Torres,L., Nellore,A., Kammers,K., Ellis,S.E., Taub,M.A., Hansen,K.D., Jaffe,A.E., Langmead,B. and Leek,J.T. (2017) Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.*, **35**, 319–321.
2. Byron,S.A., Van Keuren-Jensen,K.R., Engelthaler,D.M., Carpten,J.D. and Craig,D.W. (2016) Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.*, **17**, 257–271.
3. Xi,X., Li,T., Huang,Y., Sun,J., Zhu,Y., Yang,Y. and Lu,Z.J. (2017) RNA biomarkers: frontier of precision medicine for cancer. *Non-Coding RNA*, **3**, 9.
4. Hippen,A.A. and Greene,C.S. (2020) Expanding and remixing the metadata landscape. *Trends Cancer*, **7**, 276–278.
5. 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.
6. Philippe,N., Salson,M., Commes,T. and Rivals,E. (2013) CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biol.*, **14**, R30.
7. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
8. 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.
9. Okamura,Y. and Kinoshita,K. (2018) Matataki: an ultrafast mRNA quantification method for large-scale reanalysis of RNA-seq data. *BMC Bioinformatics*, **19**, 266.
10. Yu,Y., Liu,J., Liu,X., Zhang,Y., Magner,E., Qian,C. and Liu,J. (2018) SeqOthello: querying RNA-seq experiments at scale. *Genome Biol.*, **19**, 167.
11. 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.
12. Morillon,A. and Gautheret,D. (2019) Bridging the gap between reference and real transcriptomes. *Genome Biol.*, **20**, 112.
13. Marçais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics*, **27**, 764–770.
14. Soneson,C. (2014) compcodeR: an R package for benchmarking differential expression methods for RNA-seq data. *Bioinformatics*, **30**, 2517–2518.
15. Frazee,A.C., Jaffe,A.E., Langmead,B. and Leek,J.T. (2015) Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, **31**, 2778–2784.
16. Eisenberg,E. and Levanon,E.Y. (2013) Human housekeeping genes, revisited. *Trends Genet.*, **29**, 569–574.
17. Maan,A.A., Eales,J., Akbarov,A., Rowland,J., Xu,X., Jobling,M.A., Charchar,F.J. and Tomaszewski,M. (2017) The Y chromosome: a blueprint for men's health? *Eur. J. Hum. Genet.*, **25**, 1181–1188.

18. Drexler,H.G. and Uphoff,C.C. (2002) Mycoplasma contamination of cell cultures: incidence, sources, effects, detection, elimination, prevention. *Cytotechnology*, **39**, 75–90.
19. Yilmaz,P., Parfrey,L.W., Yarza,P., Gerken,J., Pruesse,E., Quast,C., Schweer,T., Peplies,J., Ludwig,W. and Glöckner,F.O. (2014) The SILVA and 'All-species Living Tree Project (LTP)' taxonomic frameworks. *Nucleic Acids Res.*, **42**, D643–D648.
20. Cantalupo,P.G., Katz,J.P. and Pipas,J.M. (2015) HeLa nucleic acid contamination in The Cancer Genome Atlas leads to the misidentification of human papillomavirus 18. *J. Virol.*, **89**, 4051–4057.
21. Okonechnikov,K., Golosova,O. and Fursov,M. and UGENE Team (2012) Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, **28**, 1166–1167.
22. Uphoff,C.C., Pommerenke,C., Denkmann,S.A. and Drexler,H.G. (2019) Screening human cell lines for viral infections applying RNA-seq data analysis. *PLoS One*, **14**, e0210404.
23. Rufflé,F., Audoux,J., Boureux,A., Beaumeunier,S., Gaillard,J.-B., Bou Samra,E., Megarbane,A., Cassinat,B., Chomienne,C., Alves,R. *et al.* (2017) New chimeric RNAs in acute myeloid leukemia. *F1000Res.*, **6**, https://doi.org/10.12688/f1000research.11352.2.
24. Prensner,J.R., Iyer,M.K., Balbin,O.A., Dhanasekaran,S.M., Cao,Q., Brenner,J.C., Laxman,B., Asangani,I.A., Grasso,C.S., Kominsky,H.D. *et al.* (2011) Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.*, **29**, 742–749.
25. Olarerin-George,A.O. and Hogenesch,J.B. (2015) Assessing the prevalence of mycoplasma contamination in cell culture via a survey of NCBI's RNA-seq archive. *Nucleic Acids Res.*, **43**, 2535–2542.
26. Mangul,S., Martin,L.S., Hill,B.L., Lam,A.K.-M., Distler,M.G., Zelikovsky,A., Eskin,E. and Flint,J. (2019) Systematic benchmarking of omics computational tools. *Nat. Commun.*, **10**, 1393.
27. Cáceres,A., Jene,A., Esko,T., Pérez-Jurado,L.A. and González,J.R. (2020) Extreme downregulation of chromosome Y and cancer risk in men. *J. Natl Cancer Inst.*, **112**, 913–920.
28. Cui,P., Lin,Q., Ding,F., Xin,C., Gong,W., Zhang,L., Geng,J., Zhang,B., Yu,X., Yang,J. *et al.* (2010) A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics*, **96**, 259–265.
29. McLaughlin-Drubin,M.E. and Munger,K. (2008) Viruses associated with human cancer. *Biochim. Biophys. Acta*, **1782**, 127–150.
30. Sangiovanni,M., Granata,I., Thind,A.S. and Guarracino,M.R. (2019) From trash to treasure: detecting unexpected contamination in unmapped NGS data. *BMC Bioinformatics*, **20**, 168.
31. Bhuvaneshwar,K., Song,L., Madhavan,S. and Gusev,Y. (2018) viGEN: an open source pipeline for the detection and quantification of viral RNA in human tumors. *Front. Microbiol.*, **9**, 1172.
32. Wood,D.E., Lu,J. and Langmead,B. (2019) Improved metagenomic analysis with Kraken 2. *Genome Biol.*, **20**, 257.
33. Flygare,S., Simmon,K., Miller,C., Qiao,Y., Kennedy,B., Di Sera,T., Graf,E.H., Tardif,K.D., Kapusta,A., Rynearson,S. *et al.* (2016) Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol.*, **17**, 111.
34. Cao,S., Strong,M.J., Wang,X., Moss,W.N., Concha,M., Lin,Z., O'Grady,T., Baddoo,M., Fewell,C., Renne,R. *et al.* (2015) High-throughput RNA sequencing-based virome analysis of 50 lymphoma cell lines from the Cancer Cell Line Encyclopedia Project. *J. Virol.*, **89**, 713–729.
35. Cantalupo,P.G., Katz,J.P. and Pipas,J.M. (2018) Viral sequences in human cancer. *Virology*, **513**, 208–216.
36. Selitsky,S.R., Marron,D., Hollern,D., Mose,L.E., Hoadley,K.A., Jones,C., Parker,J.S., Dittmer,D.P. and Perou,C.M. (2020) Virus expression detection reveals RNA-sequencing contamination in TCGA. *BMC Genomics*, **21**, 79.
37. Ballenghien,M., Faivre,N. and Galtier,N. (2017) Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC Biol.*, **15**, 25.
38. Rubinoff,D. and Holland,B.S. (2005) Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. *Syst. Biol.*, **54**, 952–961.
39. Hebert,P. D.N., Cywinska,A., Ball,S.L. and deWaard,J.R. (2003) Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B: Biol. Sci.*, **270**, 313–321.
40. Riquier,S., Mathieu,M., Bessiere,C., Boureux,A., Ruffle,F., Lemaitre,J.-M., Djouad,F., Gilbert,N. and Commes,T. (2021) Long non-coding RNA exploration for mesenchymal stem cell characterisation. *BMC Genomics*, **22**. 412.
41. Mangul,S., Yang,H.T., Strauli,N., Gruhl,F., Porath,H.T., Hsieh,K., Chen,L., Daley,T., Christenson,S., Wesolowska-Andersen,A. *et al.* (2018) ROP: dumpster diving in RNA-sequencing to find the source of 1 trillion reads across diverse adult human tissues. *Genome Biol.*, **19**, 36.
42. 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.
43. Marchet,C., Boucher,C., Puglisi,S.J., Medvedev,P., Salson,M. and Chikhi,R. (2021) Data structures based on *k*-mers for querying large collections of sequencing datasets. *Genome Research*, **31**, 1–12.
44. Patro,R., Mount,S.M. and Kingsford,C. (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.*, **32**, 462–464.
45. Zhang,Z. and Wang,W. (2014) RNA-Skim: a rapid method for RNA-seq quantification at transcript level. *Bioinformatics*, **30**, i283–i292.
46. Silvester,N., Alako,B., Amid,C., Cerdeño-Tarrága,A., Clarke,L., Cleland,I., Harrison,P.W., Jayathilaka,S., Kay,S., Keane,T. *et al.* (2018) The European Nucleotide Archive in 2017. *Nucleic Acids Res.*, **46**, D36–D40.