Exhaustive capture of biological variation in RNA-seq data through k-mer decomposition

Jérôme Audoux[1], Nicolas Philippe[1], Rayan Chikhi[2], Mikaël Salson[2], Marc Gabriel[3], Thérèse Commes[1,4], and Daniel Gautheret[3,*]

[1]INSERM U1183 IRMB, Université de Montpellier, Montpellier, France
[2]Univ. Lille, CNRS, Inria, UMR 9189 - CRIStAL - F-59000 Lille, France
[3]Institute for Integrative Biology of the Cell, Université Paris-Sud, CNRS, CEA. Gif sur Yvette, France
[4]Institut de Biologie Computationnelle, Université Montpellier, Montpellier, France
[*]daniel.gautheret@u-psud.fr

ABSTRACT

Each individual cell produces its own set of transcripts, which is a combinatorial result of genetic, transcriptomic and post-transcriptomic variations. Due to this combinatorial nature, obtaining the exhaustive set of full-length transcripts for a given species is a never ending endeavor. Yet, each RNA deep sequencing experiment turns out a variety of transcripts that depart from reference transcriptomes and should be properly identified. To address this challenge, we introduce a k-mer-based software protocol for capturing local transcriptional variation from a set of standard RNA-seq libraries, independently of a reference genome or transcriptome. Our software, called DE-kupl, analyzes k-mer contents and detects k-mers with differential abundance directly from the sequencing files, prior to assembly or mapping. This enables to retrieve the virtually complete set of unannotated variation lying in an RNA-seq dataset. This variation can be subsequently assigned to lincRNAs, antisense RNAs, splice and polyadenylation variants, retained introns, expressed repeats, chimeric or circular RNA, foreign RNA and SNV-harbouring RNA. We applied DE-kupl to a published differential RNA-seq experiment carried on a human cell line, and were able to discover highly significant unannotated transcript variations. We propose that DE-Kupl could be a valuable tool for extracting in full the untapped transcript information contained in large scale transcriptome projects.

Introduction

cDNA sequencing[1] and massively parallel RNA sequencing (RNA-seq) revealed that organisms produce a variety of RNAs that is far larger than previously expected. Due to the combination of two major phenomena, pervasive transcription and variable RNA processing, modern transcript catalogues such as Gencode may harbor ten times more transcripts than there are protein-coding genes[2]. However, in spite of this apparently broad picture of the transcriptome enabled by high throughput sequencing, we argue that a large segment of transcriptomic information is essentially disregarded.

To illustrate this point, let us consider the biological events that drive transcript diversity. Firstly, transcripts result from transcription initiation events either at promoters of protein-coding and non-coding genes, or at multiple antisense or inter/intra genic loci. Secondly, transcripts are processed by a large variety of mechanisms, including splicing and polyadenylation, editing[3], circularization[4] and cleavage/degradation by various nucleases[5,6]. Thirdly, an essential, yet often overlooked, source of transcript diversity, is genomic variation. Polymorphism and structural variations within transcribed regions produce RNAs with single nucleotide variations, tandem duplications or deletions, transposon integration, unstable microsatellites or fusion events. These events are major sources of transcript variation that can strongly impact coding potential in translated regions and protein-binding abilities in untranslated regions.

Current bioinformatics protocols for RNA-seq analysis do not properly account for this vast diversity of transcripts. Prevalent computing strategies can be roughly classified into two categories: reference-based tools[7–10] rely on the alignment (or pseudo-alignment) of RNA-seq reads to a reference genome or transcriptome, while de novo assembly tools[11] reconstruct full-length transcripts based on the analysis of RNA-seq reads. These protocols fail to account for true transcriptional diversity in several respects: (i) they ignore small-scale variations such as SNP or indels (ii) they rely on full-length transcripts that cannot represent the combinatorials of variations observed in actual populations and that are impacted by bias in alignment procedures, and (iii) they misrepresent transcripts containing repeats due to ambiguity in alignment or assembly.

We propose a new approach to RNA-seq analysis that facilitates the discovery of all types of events occurring in an RNA-seq library, independently of alignment or reconstruction. Our approach relies on k-mer indexing of sequence files, a technique that recently gained momentum in NGS data analysis[9,10,12–14]. In order to focus on biologically meaningful events, we focus on

differentially expressed k-mers, hence the name of our method, DE-kupl. Using published human RNA-seq datasets, we show that a large amount of RNA variation can be captured that is not represented in existing transcript catalogues. As a proof of concept, we applied DE-kupl to a published Epythelial-Mensenchymal Transition (EMT) model dataset and characterized a large number of novel events.

Results

Reference datasets are an incomplete representation of actual transcriptomes

First, we analyzed k-mer diversity in different human references and high-throughput experimental sequences. To this aim, we extracted all 31-nt k-mers from sequence files using the Jellyfish program[15]. Figure 1A-B compares k-mers from Gencode transcripts, the human genome reference and RNA-seq libraries from 18 different individuals[16] corresponding to three primary tissues (6 RNA-Seq libraries/tissue). To minimize the risk of including k-mers that contain sequencing errors, we retained for each tissue only the set of k-mers that appear in 6 or more individuals.

Measures of k-mer abundance show that k-mers are overwhelmingly associated to Gencode transcripts (Fig 1B1). However, when considering k-mer diversity, a large fraction of k-mers are tissue-specific and not found in the Gencode reference (Fig 1A). These tissue-specific k-mers may result from sequencing errors, genetic variation in individuals or novel, or non-reference transcripts. The majority of RNA-seq k-mers that do not occur in Gencode are found in the human genome reference (Fig 1B, 1B2), suggesting polymorphisms and errors represent a minor fraction of tissue-specific k-mers and a lot of k-mers results from expressed genome regions that are not represented in Gencode. Further scrutiny of tissue-specific k-mers shows that a significant fraction can be mapped to the transcriptome with one substitution. However, for each tissue there is an average of 1 million k-mers that cannot be mapped to either reference (1B3).

Non-reference k-mers classify samples as accurately as reference transcripts. We performed a Principal Component Analysis (PCA) of the above human tissue samples using conventional transcript counts and k-mer counts. PCA based on 20,000 randomly selected unmapped k-mers was able to differentiate tissues as well as PCA based on estimated gene expression or transcript expression (Fig 2). This illustrates how a "shadow", non reference transcriptome that is not incorporated in standard analyses comprises biologically relevant expression data.

When comparing RNA-seq and whole genome sequence (WGS) data from the same individual[17], library-specific k-mers represent a much larger fraction of RNA-seq than of WGS k-mers (Fig 3). This shows that non-reference sequence diversity is larger in RNA-seq than in WGS. Altogether these results point towards the existence of a significant amount of untapped biological information in RNA-seq data.

Non-reference k-mers may result from three classes of biological events. First, they may stem from genetic polymorphism in the studied sample. Second, they may result from RNA processing, notably, but not limited to, splicing and polyadenylation. A predominant source of k-mers in this category is intron retention, whose products are not usually incorporated into reference databases and are mostly by-products of regular gene expression. A third, major source of k-mer "innovation" is intergenic expression (eg. lincRNA, antisense RNA, expressed repeats or endogenous viral sequences). Altogether, the combination of these genetic, transcriptional and post-transcriptional events may have a profound impact on transcript function.

A new k-mer based protocol for deriving transcriptome variation from RNA-seq data

We designed the DE-kupl computational protocol with the aim to capture all k-mer variation in an input set of RNA-seq libraries. This protocol is composed of four main components (Figure 4):

1. Indexing: index and count all k-mers (k=31) in the input libraries
2. Filtering: delete k-mers representing potential sequencing errors or perfectly matching known transcripts
3. Differential Expression (DE): select k-mers with significantly different abundance across conditions
4. Assembly and annotation: build contigs of assembled k-mers and annotate contigs based on sequence alignment.

DE-kupl departs radically from all existing RNA-seq analysis procedures in that it does neither "map-first" (a la Tuxedo suite) or "assemble-first" (a la Trinity) but instead directly analyzes contents of the raw fastq files, displacing assembly and mapping to the final stage of the procedure. In this way, DE-kupl guarantees that no variation in the input sequence (even at the level of a single nucleotide) is lost at the initial stage of the analysis. Even unmappable k-mers such as sequences from repeats, low complexity regions or exogeneous organisms, are retained up to the final stage and can be analyzed.

The DE-kupl protocol is detailed in Methods. We highlight here some of its key features. First, DE-kupl must deal with the large size of the k-mer index. A single human RNA-seq library has in the order of 10^8 k-mers and an index for 50 individual samples can reach 10^{10} k-mers (not shown). We selected the Jellyfish tool for counting k-mers[15] as it presents very fast computing times and allows to store the full index on disk for further query. Other key steps of the procedure (k-mer table

merging, DE tests, k-mer assembly) were written in C, enabling the whole procedure to run on a relatively standard computer in a reasonable amount of time.

The central process in DE-kupl is k-mer filtering. Filtering out unique or rare k-mers is relatively straightforward and considerably reduces k-mer diversity and the amount of sequence errors. Another stringent filter is the removal of k-mers matching reference Gencode transcripts. The rationale for this is that the bulk of k-mers in RNA-seq data comes from expressed exons, and we are not interested in this canonical exon expression, as it can be easily captured by conventional, reference-based protocols[9,10]. Discarding these k-mers enable us to ignore the very strong signal caused by known transcripts and focus on expressed regions harboring differences from the reference transcriptome.

To demonstrate the capacity of DE-kupl to discover novel biological events, we applied the procedure to 12 RNA-samples from an EMT cell-line model[18]. EMT was induced in NSCLC cells by ZEB1 expression over a 7-day time course. We compared 6 RNA-seq libraries from the "Epythelial" stage of the time course (uninduced and Day 1) with 6 libraries from the "Mesenchymal" stage (Day 6 and 7, Table 1).

DE-kupl discovers and assembles unannotated differentially-expressed events

The full DE-Kupl procedure was completed in less than 10 hours using 4 computing cores, 16 GB RAM and 60 GB of hard disk space (Table 2). Recurrence filters efficiently reduced k-mer counts from 707M to 92.5M and the Gencode filter further reduced counts to 40.3M. Differential analysis eventually retained 3.6M k-mers that were assembled into 128k contigs (Table 3). The resulting contigs range in size from 31nt (corresponding to an "orphan" k-mer) to 3.6kb, with a major peak of short 31-40nt contigs and a minor peak around 61nt (Fig 5A). 61nt-contigs are formed by 31 overlapping k-mers harboring a single nucleotide variation (SNV) at every position of the k-mer. This phenomenon also causes a higher mismatch ratio for contigs around 61nt (Fig 5B). Therefore 61nt contigs are predominantly associated to SNVs.

Contigs that do not map onto the human genome are generally shorter than mapped contigs (Fig 5A), indicating a lower signal-to-noise ratio in the former class. Expectedly, shorter mapped contigs tend to map at multiple loci more often than longer ones (Fig 5C), however 80% of all contigs are uniquely mapped (data not shown).

Contig locations reveal distinct classes of biological events. Most contigs are located in annotated introns and exons (Fig 6), however intronic contigs are predominantly exact matches while exonic contigs predominantly carry one mismatch. This effect is due to Gencode filtering : contigs with exact matches to introns are usually not filtered, as they do not pertain to a Gencode transcript, while contigs that match exons are filtered out unless they differ from the reference. This difference might be in the form of SNVs, or through exons extending in flanking intergenic or intronic regions. Under the same rationale, contigs mapping to intergenic and antisense regions are depleted in SNVs (Fig 6), consistent with their location in unannotated lncRNAs and antisense-RNAs, while contigs overlapping exon-exon junctions behave like exonic contigs (high rate of SNV). However, a significant fraction of exon junction contigs are exact matches, indicating they may correspond to novel junctions.

Annotating new EMT events from mapped and unmapped contigs

Almost all (99.2%) the 128k DE contigs mapped to the human genome, at 1633 different loci (Table 4). Our annotation procedure decomposed DE contigs into eight classes of biological events using the rule set described in Table 4. We describe below the different classes of events. We grouped contigs into "loci", here defined as independent annotated genes or intergenic regions. We remind that all the reported contigs are both differentially expressed and composed of k-mers that are absent from Gencode transcripts.

Alternative splicing. Analysis of split-mapped contigs found evidence of potentially novel, differential splice variants at 338 loci derived from a set of 371 contigs. (Table 4, Fig 7A,B,C). We purposedly excluded SNV-containing splice sites, and splice events in genes described as differentially expressed by the conventional DE procedure. Therefore differential splicing at these sites may not be a consequence from DE of the whole gene. Remarkably, these novel events include a number of subtle variations at 5' and 3' splice sites with 3-15 nt difference from the annotated reference, which escaped prior annotation (see eg. Fig S1).

Alternative polyadenylation. We extracted all contigs aligned with 5 or more clipped (e.g. non-reference) bases at their 3' end, and containing 5 or more trailing As. Out of 166 such poly-A terminated contigs, 125 (75%) contain an AATAAA or variant polyadenyalation signals, indicating they result from actual polyadenylated transcripts. Although 95 such "polyA contigs" come from genes predicted as differentially expressed by the standard procedure, 30 come from previous non-genic regions or genes not predicted as differentially expressed, thus suggesting the occurrence of differential polyadenylation at these sites (Table 4, S1, S2).

LincRNA. We identified a subset of 927 DE contigs corresponding to potential long intergenic non-coding RNAs (Table 4). Criteria for lincRNAs were contigs of size $>$ 200nt and mapped to an intergenic locus. These potential lincRNAs were found in 220 distinct intergenic regions. Visual inspection reveals clear lincRNA-like patterns at these loci, whith contigs clustered

into well defined transcription units with abundant read coverage and abundant splicing (Fig 7C, Fig S2). DE-kupl is thus an effective tool for the identification of novel DE lincRNAs.

Antisense RNAs. When DE-kupl is applied to stranded RNA-seq libraries (as with the EMT libraries used in this study), the resulting contigs are strand-specific and can be used for disambiguating sense/antisense expression. We identified 400 contigs from 173 loci mapping to the reverse strand of an annotated gene (Table 4). These antisense RNAs include very strong cases of differential expression (Fig 7D), sometimes combined to apparent repression of the sense gene (Fig S3).

Allele-specific expression. As DE-kupl quantifies every SNV-containing k-mer, we set out to exploit this capacity to identify potential allele-specific expression events. We extracted all contigs including an SNV and mapping to an exon whose host gene was not measured as differentially expressed. This was a less than perfect procedure, as we did not explicitly tested for a switch in allelic balance among the two conditions. Yet, among the 732 contigs identified (Table 4), some display strong apparent changes in allelic balance between the E and M conditions (eg. Fig S4).

Intron retention. As highly expressed transcripts often carry intronic byproducts, we expected DE-kupl to turn out a lot of "parasitic" intronic contigs. To mitigate this artifact, we focused on intronic k-mers from genes that were not DE. This filter identified 547 intronic contigs from 200 different genes (Table 4). This included cases of strong localized intronic expression suggesting novel exons (Fig S5) as well as cases where a specific short intronic region was differentially expressed, reminiscent of the pattern observed at intronic processed miRNAs and snoRNAs[19] (Fig S6).

Expressed repeats. Assessing the expression of human repeats by conventional RNA-seq analysis protocols is difficult, as mapping ambiguities render repeat regions "unmappable"[20]. Since DE-kupl first measures expression independently of mapping, we were able to collect and analyze differential contigs with multiple genome hits. 7243 contigs of size 50nt or larger have multiple hits, and 1111 are repeated more than 5 times. RepeatMasker[21] found 747 out of these 1111 sequences to match known repeats, mostly LINEs, LTRs and SINEs (Fig S8). Further inspection showed that most of the remaining multiple-hit contigs correspond to unannotated repeats or low complexity regions. One of the most striking differential repeats is an unannotated 22x66nt tandem repeat, located about 2Mb from the chromosome 8 telomere. This repeat is found about 50-fold overexpressed in the Mesenchymal condition (Fig 7B, S7).

Unmapped contigs. Unmapped contigs may result from transcripts produced by highly rearranged genes or by exogenous viral genomes and could thus be highly relevant biologically. In principle, DE-kupl is able to detect such events when analyzing samples with varying levels of foreign RNA. However, here we compared samples from the same cell line, thus we did not expect to observe such phenomena. Indeed, out of 112 unmapped contigs of size $>$ 50nt (Table 4), the vast majority (76%) correspond to vector sequences overexpressed in the "M" condition (not shown), indicating these contigs come from the expression vector used for EMT induction. The remaining unmapped contigs are low complexity sequences or align to non-human primate sequences, indicating they result from yet incorrectly assembled regions of the human genome.

Discussion

K-mer decomposition followed by filtering and differential expression analysis is a novel way of analysing RNA-seq data that is capable of detecting a wider spectrum of transcript variation than previous protocols. Contrarily to popular RNA-seq analysis software, DE-kupl does not attempt full transcript recognition or reconstruction but focuses instead on local transcript variations. In some way it is closer in spirit to methods analyzing local RNA-seq coverage such as RNAprof[22] and DERfinder[23], with the notable exception that DE-kupl does not involve mapping and thus avoids mapping-related pitfalls and simplifications. In fact, we do not consider full-length transcript reconstruction to be a realistic, or even desirable goal when all levels of variation are considered, as the combinatorial nature of genomic, transcriptomic and post-transcriptomic variations would create indefinitely expanding transcript catalogues.

DE-kupl explores all k-mers in the input RNA-seq files (vs. only k-mers from annotated transcripts in existing protocols) which potentially entails heavy computational time and memory requirement. Using the Jellyfish k-mer indexing software and C-programming code for key table manipulation, we achieved time/memory requirements on par with popular mapping-based protocols for similarly sized datasets. Another key aspect of our protocol that rendered the "full k-mer" approach tractable was applying successive filters for rare k-mers, Gencode transcripts and Differential Expression, which altogether resulted in a 200-fold reduction in k-mer counts. These filters are not only useful for technical reasons (they reduce runtimes and enable to get rid of most sequence errors), but they allow to focus on k-mers which (i) vary significantly between the conditions under study, and (ii) would not be captured by conventional reference-based protocols.

Using an RNA-seq dataset from a human cell line, we showed that DE-kupl is able to detect events caused by alternative transcription and alternative RNA processing, as well as SNVs in DE genes and potential allele-specific expression events. This ability to identify transcriptome differences linked to genetic variation is an important benefit of the DE-kupl approach.

Furthermore, DE-kupl takes advantage of read direction in oriented RNA-seq libraries, which facilitates the disambiguation of sense and antisense transcription units at complex loci, as illustrated in some of the events shown.

In this proof of concept study, we applied DE-kupl to RNA-seq libraries from a single cell line, with no expected polymorphism among samples. The next step will be application to libraries from multiple individual organisms. Although K-mer diversity is much higher in a set of individuals than in a cell line, our initial tests with with RNA-seq data from 60 individual tumors were completed successfully on a medium size server (data not shown). The capacity of DE-kupl to simultaneously detect genetic variation and RNA processing events opens exciting perspective for the analysis of patient samples. Finally, it did not escape our notice that k-mers of interest uncovered by DE-kupl can be used for the efficient querying of large scale RNA-seq repositories such as the Cancer Genome Atlas in order to retrieve similar events.

Methods

Characterization of k-mer diversity in human RNA-seq libraries

RNA-seq data for bone marrow, skin and colon where retrieved from the human protein atlas project (10.1126/science.1260419, E-MTAB-2836). A total of 18 datasets from different individuals, corresponding to 6 replicates per tissue, were downloaded from EBI/ENA (bone marrow: ERR315469, ERR315425, ERR315486, ERR315396, ERR315404, ERR315406, colon: ERR315348, ERR315403, ERR315357, ERR315484, ERR315400, ERR315462, skin: ERR315401, ERR315464, ERR315460, ERR315372, ERR315376, ERR315339). The reference GRCh38 genome and Ensembl 86 transcripts were downloaded from Ensembl.

First we counted k-mers in each RNA-Seq and references sequence set using Jellyfish (2.2.0) count, with options k = 32 and -C (canonical k-mers). The k-mer list for each tissue (Fig 1A and B) was produced by merging counts for all 6 samples and conserving only those found in all replicates.

For mapping statistics (Fig 1B3), we extracted k-mers specific of each tissue and mapped them to the Ensembl 86 transcript reference using Bowtie 1 (version 1.1.2). Unmapped k-mers were mapped a second time with Bowtie 1 to the GRCh38 genome reference. Reads with 3 or more mismatches are not mapped by Bowtie 1 and, therefore, are considered as unmapped.

The intersection of k-mers between RNA-Seq and WGS data (Fig 1C), is based on the transcriptome and genome of lymphoblastoid cell lines[17]. K-mers were counted in these libraries with the same procedure as described before. In order to reduce bias from sequencing errors, k-mers with only one occurrence were filtered out.

DE-kupl Implementation

The DE-kupl pipeline (Fig 4) is implemented using the Snakemake[24] workflow manager. A configuration file is filled up by the user with location of FASTQ files, the condition of each sample, as well as global parameters such as k-mer length, cpu number, maximum memory and other parameters for each step of the pipeline which will be described hereinafter.

K-mer counting

Raw sequences (FASTQ files) are first processed with the "jellyfish count" command of the Jellyfish software, which produces one library (a dump of the Jellyfish hash-table) for each sample. For stranded RNA-seq libraries, reads in reverse direction relative to transcript are reverse-complemented, ensuring proper orientation of k-mers. At this point, for each library, only k-mers having at least 2 occurrences are recorded (user-defined parameter). Once a jellyfish index has been constructed, we use the "jellyfish dump" command to output the raw-counts in a two column text file, that contains at each line a k-mer and its number of occurrences. These raw counts are sorted alphabetically by k-mer sequence with the Unix "sort" command.

K-mer filtering

All samples counts are then joined together with "dekupl-joinCounts" binary to produce a single matrix will all k-mer and their abundances in all samples. During this step, the user can set two parameters: "min_reccurence" which define the minimum number of samples to have counts for a given k-mer and "min_reccurence_abundance" which define the minimum count value for one sample to be considered for the "min_reccurence" filter. Given n samples, k_i the count for the k-mer k in the ith sample, and a the value of "min_reccurence_abundance", the recurrence is computed with $reccurence(k,a) = \sum_{i=1}^{n} k_i > a_i$. Usually "min_reccurence" is set to the number of replicates in each conditions, and "min_reccurence_abundance" is set to 5. In order to remove known transcripts sequences from our set of experimental k-mers, we also use our Jellyfish-based procedure to create the set of k-mers appearing in the reference transcriptome and we subtract this set from the experimental k-mer. The final matrix we obtain from this subtraction will be later referred as the "raw counts".

Differential gene expression

In parallel to the k-mer counting and filtering procedure, we analyze the RNA-Seq data with a more conventional pipeline. First reads are processed with kallisto to estimate transcripts abundances. The transcripts estimated counts are then collapsed to the

gene-level and processed with DESeq2 statistical framework in order to produce a set of differentially expressed genes (DEGs) and normalization factors (NF) for each sample.

Differential k-mer expression

In order to identify k-mers having differential expression between two conditions, we apply a T-test on the log transformed counts previously normalized with the NF produced from the differential gene expression procedure. Since conventional DE statistical procedure (DESeq2, EdgeR) cannot be used for millions of k-mers, we use a t-test on log transformed counts to approach a normal distribution, similar to the procedure used in other studies[25]. The p-values obtained from the T-test are then corrected with the Benjamini-Hochberg procedure and k-mers not rejecting the null hypothesis ($FDR \geq 0.05$) are filtered-out. This procedure has been implemented in C in the "dekupl-TtestFilter" binary, for performance purposes. It should be noted that, since the DE statistics are computed on a very large number of k-mers, multiple testing correction strongly affects the resulting P-values. Pending improved DE statistics, we recommand using DE-kupl for designs with at least 5 vs 5 libraries. Our tests with fewer libraries often yield no DE contig.

K-mer assembly

DE k-mers are assembled de novo in order to group k-mers that potentially overlap the same event (ie. all k-mer overlapping a new differential splice junction or SNV). To this aim, we developed our own procedure called "mergeTags", which works as follows: first we try to merge k-mers having non-ambiguous $k-1$ prefix-suffix overlap. For example, given the set of k-mers : ATG, TGA, TGC, CAT, the following contigs are produced : $contigs = CATG, TGA, TGC$. We repeat this assembly step using assembled k-mers until no overlap is found. We then repeat the assembly process with $k-2$ prefix-suffix overlaps, using as input the assemblies produced at the previous step, and so forth. Finally, a set of DE contigs is produced and each contig is labelled by the assembled k-mer having the lowest p-value. This assembly procedure is implemented in C in the "dekupl-mergeTags" binary. By default the assembly process stops after assembling sequences with 15nt overlaps.

Annotation

Finally, DE contigs are annotated to facilitate biological inference. Annotated features, summarized in Table S3, are reported in a contig summary table. First, contigs are aligned with BLAST[26] against Illumina adapters. Contigs matching adapters are discarded. Retained contigs are further mapped to the reference genome using the GSNAP short read aligner[27], which showed the best speed/sensitivity ratio for aligning both short and long contigs. GSNAP is used with option `-N 1` to enable identification of new splice junctions. Contigs not mapped by GSNAP are collected and re-aligned using BLAST. Alignment characteristics are extracted from GSNAP and BLAST alignments. Alignment coordinates are then compared with Ensembl (v86) annotations (in GFF3 format) using BEDTools[28] and a set of locus-related features is extracted. Finally we generate two supplementary files, one containing a "per locus" summary of contigs, and a BED file of contig locations that can be used as a display track in genome browsers. In the "per locus" table, a locus is defined as either an annotated gene, the genomic region located on the opposite strand of an annotated gene, or the genomic region separating two annotated genes. The table records the number of contigs overlapping each locus as well as the contig with lowest FDR for this genomic interval.

Availability

The DE-kupl pipeline software is available at `https://github.com/Transipedia/dekupl`.

References

1. Carninci, P. *et al.* The Transcriptional Landscape of the Mammalian Genome. *Science* **309**, 1559–1563 (2005). DOI 10.1126/science.1112014.
2. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22**, 1760–1774 (2012). DOI 10.1101/gr.135350.111.
3. Nishikura, K. Functions and Regulation of RNA Editing by ADAR Deaminases. *Annual review of biochemistry* **79**, 321–349 (2010). DOI 10.1146/annurev-biochem-060208-105251.
4. Chen, L.-L. The biogenesis and emerging roles of circular RNAs. *Nature Reviews Molecular Cell Biology* **17**, 205–211 (2016). DOI 10.1038/nrm.2015.32.
5. Kirchner, S. & Ignatova, Z. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. *Nature Reviews Genetics* **16**, 98–112 (2015). DOI 10.1038/nrg3861.
6. Dieci, G., Preti, M. & Montanini, B. Eukaryotic snoRNAs: A paradigm for gene expression flexibility. *Genomics* **94**, 83–88 (2009). DOI 10.1016/j.ygeno.2009.05.002.
7. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511–515 (2010). DOI 10.1038/nbt.1621.

8. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011). DOI 10.1186/1471-2105-12-323.

9. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**, 525–527 (2016). DOI 10.1038/nbt.3519.

10. Patro, R., Duggal, G. & Kingsford, C. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. *bioRxiv* 021592 (2015). DOI 10.1101/021592.

11. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644–652 (2011). DOI 10.1038/nbt.1883.

12. Nordström, K. J. V. *et al.* Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nature Biotechnology* **31**, 325–330 (2013). DOI 10.1038/nbt.2515.

13. Shajii, A. R., Yorukoglu, D., Yu, Y. W. & Berger, B. Fast genotyping of known SNPs through approximate k-mer matching. *bioRxiv* 063446 (2016). DOI 10.1101/063446.

14. Ondov, B. D. *et al.* Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17**, 132 (2016). DOI 10.1186/s13059-016-0997-x.

15. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011). DOI 10.1093/bioinformatics/btr011.

16. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015). DOI 10.1126/science.1260419.

17. Griffith, M. *et al.* Genome Modeling System: A Knowledge Management Platform for Genomics. *PLOS Computational Biology* **11**, e1004274 (2015). DOI 10.1371/journal.pcbi.1004274.

18. Yang, Y. *et al.* Determination of a Comprehensive Alternative Splicing Regulatory Network and Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition. *Molecular and Cellular Biology* **36**, 1704–1719 (2016). DOI 10.1128/MCB.00019-16.

19. Miyoshi, K., Miyoshi, T. & Siomi, H. Many ways to generate microRNA-like small RNAs: Non-canonical pathways for microRNA production. *Molecular genetics and genomics: MGG* **284**, 95–103 (2010). DOI 10.1007/s00438-010-0556-1.

20. Derrien, T. *et al.* Fast Computation and Applications of Genome Mappability. *PLOS ONE* **7**, e30377 (2012). DOI 10.1371/journal.pone.0030377.

21. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0 (2013-2015).

22. Tran, V. D. T., Souiai, O., Romero-Barrios, N., Crespi, M. & Gautheret, D. Detection of generic differential RNA processing events from RNA-seq data. *RNA biology* **13**, 59–67 (2016). DOI 10.1080/15476286.2015.1118604.

23. Frazee, A. C., Sabunciyan, S., Hansen, K. D., Irizarry, R. A. & Leek, J. T. Differential expression analysis of RNA-seq data at single-base resolution. *Biostatistics (Oxford, England)* **15**, 413–426 (2014). DOI 10.1093/biostatistics/kxt053.

24. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012). DOI 10.1093/bioinformatics/bts480.

25. Love, M. I., Hogenesch, J. B. & Irizarry, R. A. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature Biotechnology* **34**, 1287–1291 (2016). DOI 10.1038/nbt.3682.

26. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). DOI 10.1186/1471-2105-10-421.

27. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010). DOI 10.1093/bioinformatics/btq057.

28. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841–842 (2010). DOI 10.1093/bioinformatics/btq033.

29. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* bts635 (2012). DOI 10.1093/bioinformatics/bts635.

Acknowledgements

We thank Damien Drubay and Mélina Gallopin for useful statistical discussions. This project was funded in part by "Plan Cancer – Systems Biology" grant #bio2014-04 to DG.

Additional information

Table 1. EMT experience design.

	Condition 1	Condition 2	Total
Experiments	Day 0 + Day 1 (triplicates)	Day 6 + Day 7 (triplicates)	
Files number *(paired-end gzip-compressed fastq)*	2 * 3	2 * 3	24 files (R1+R2)
Sizes	61.2 G	56.7 G	117.9 G

Table 2. DE-kupl parameters and ressource used for the EMT experiment

Parameter/Ressource	value
nb_threads	4
min_recurrence	6
min_recurrence_abundance	5
pvalue_threshold	0.05
Max memory usage	23 GB
Running time	9h 47m
Max disk used	59 GB

Table 3. DE-kupl pipeline results for EMT experience. Description of output files sequentially generated by DE-kupl. Number of kmers/contigs, correspond to the number of lines in each file.

Files	Description	Nb of kmers or contigs	Sizes
raw_counts (no filter)	Matrix of k-mers counts from all libraries	707,067,278	(not generated)
raw_counts.tsv.gz	Matrix of k-mer counts from all libraries with recurrence filters	92,525,450	1.9 GB
noGENCODE-counts.tsv.gz	Counts filtered with Gencode k-mers	40,398,848	728 MB
diff-counts.tsv.gz	Counts with differential expression test, filtered on FDR	3,642,688	177 MB
merged-diff-counts.tsv.gz	DE k-mers assembled into contigs	128,275	8.4 MB

Table 4. Classes of annotated DE contigs. Each class of contigs is described by the set of conditions applied to select annotated contigs. Column "Other condt" refers to the following criteria: 1. Contig ends with AAAAA, 2. Mean counts in both conditions > 20, 3. Mean counts > 20 in at least 1 condition & mapped region $< 1kb$, 4. Mean counts in cond1+cond2 > 70 & mapped region $< 1kb$. Column "Contigs" indicates the number of contigs of each class found in the EMT experiment. Column "Loci" is the number of loci implicated by these contigs (see Methods).

	Conditions													
	gene is diff	nb junctions	maps gene	maps AS gene	clipped 3p	is mapped	SNV	Exonic	Intronic	nb hit	contig length	other condt	Contigs	Loci
Alternative Splicing	F	>0	T			T							371	338
Alternative polyA					>=5	T						1	166	144
lincRNA			F	F		T					>200	2	927	330
asRNA			F	T		T					>200	2	400	173
Allele specific expression	F		T	F		T	T	T		1		3	732	525
Intron retention	F		T	F		T			T	1		4	547	200
Expressed Repeats						T				>4	>50		1111	587
Unmapped						F					>50		112	

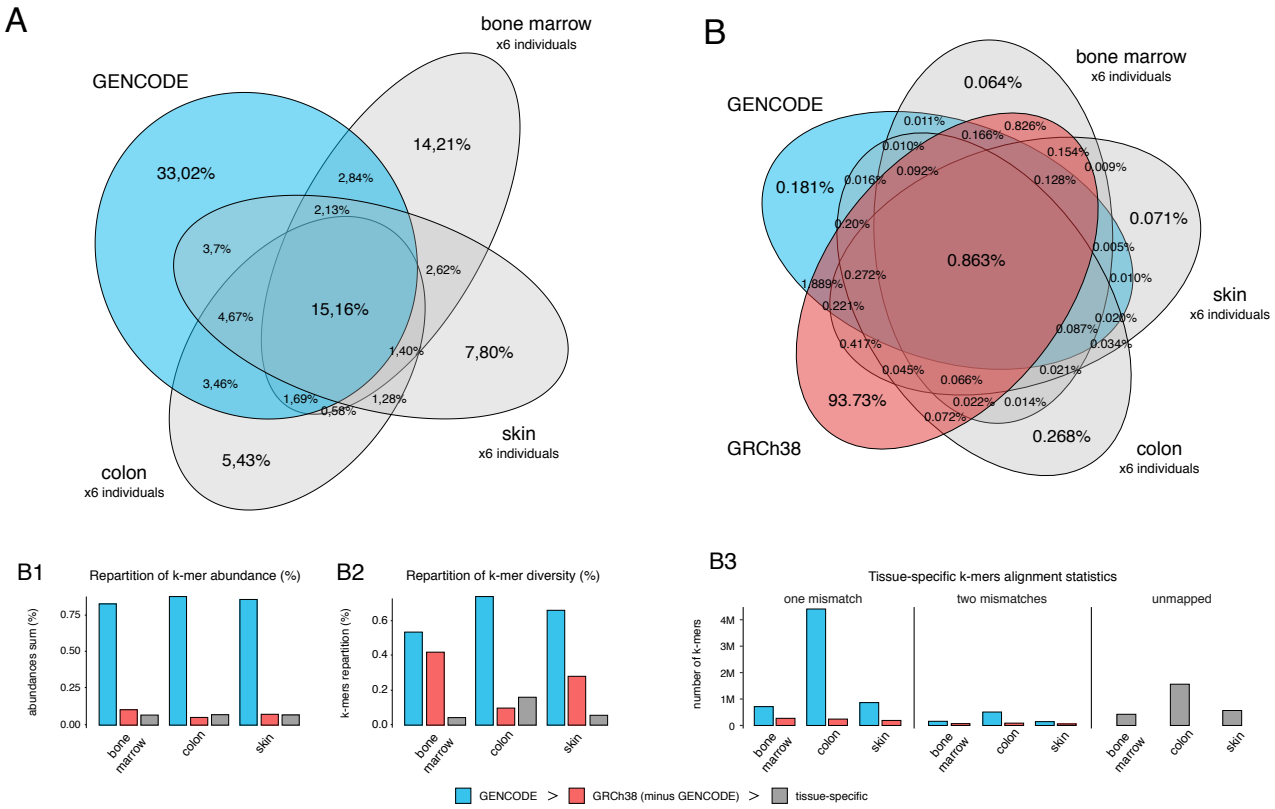

Figure 1. The diversity of 31nt k-mers in RNA-Seq exceeds that of reference sequences. A. Intersections of k-mers between Gencode transcripts and RNA-Seq data from three tissues: bone marrow, skin and colon. The set of k-mers for each tissue was defined as the common k-mers shared by all six individuals. **B**. Intersections of k-mers between Gencode transcripts, the reference human genome (GRCh38) and RNA-Seq data (same as in A). **B1**. Repartition of k-mers abundances for each tissue represented in A and B. K-mers shared with Gencode are labelled as "GENCODE", then k-mers shared with the human genome are labelled as "GRCh38", other k-mers are labelled as "tissue-specific". The same procedure was applied in B2 and B3. **B2**. Repartition of k-mer diversity for each tissue. **B3**. Mapping statistics of k-mers labeled as "tissue-specific" in B2. These k-mers were first mapped to Gencode transcripts, and unmapped k-mers were then mapped to the GRCh38 reference, using Bowtie1. Bowtie1 allows up to 2 mismatches in a 31nt k-mer.

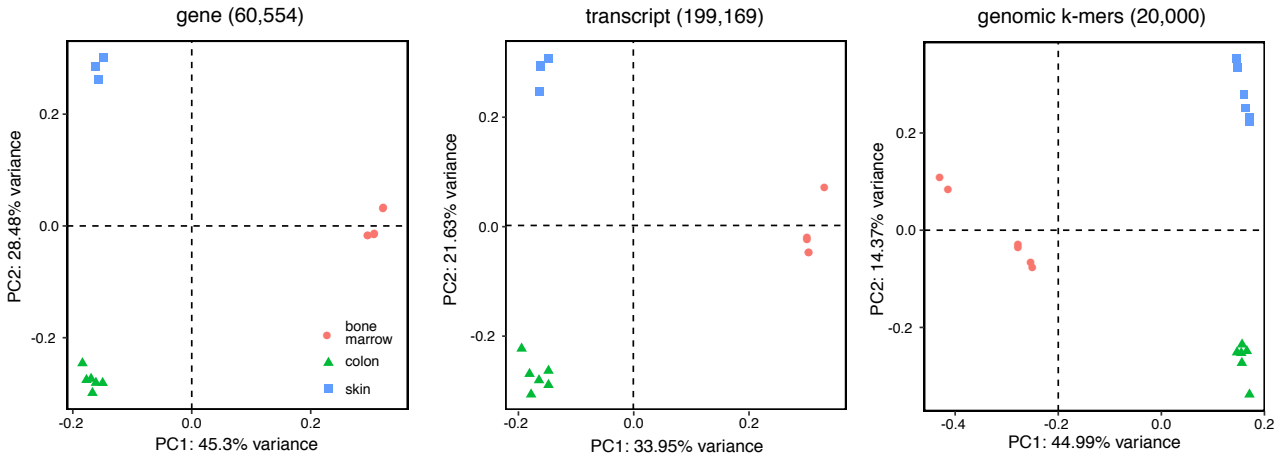

Figure 2. Principal Component Analysis on non-reference k-mers discriminates tissues. Samples are labeled according to their tissues (bone marrow, colon, skin). PCA were produced with normalized, log transformed, counts. For genes, and transcripts, counts were generated with Kallisto based on Gencode V25. Genomic k-mers correspond to 20k random k-mers not mapped to Gencode transcripts but successfully mapped to GRCh38.

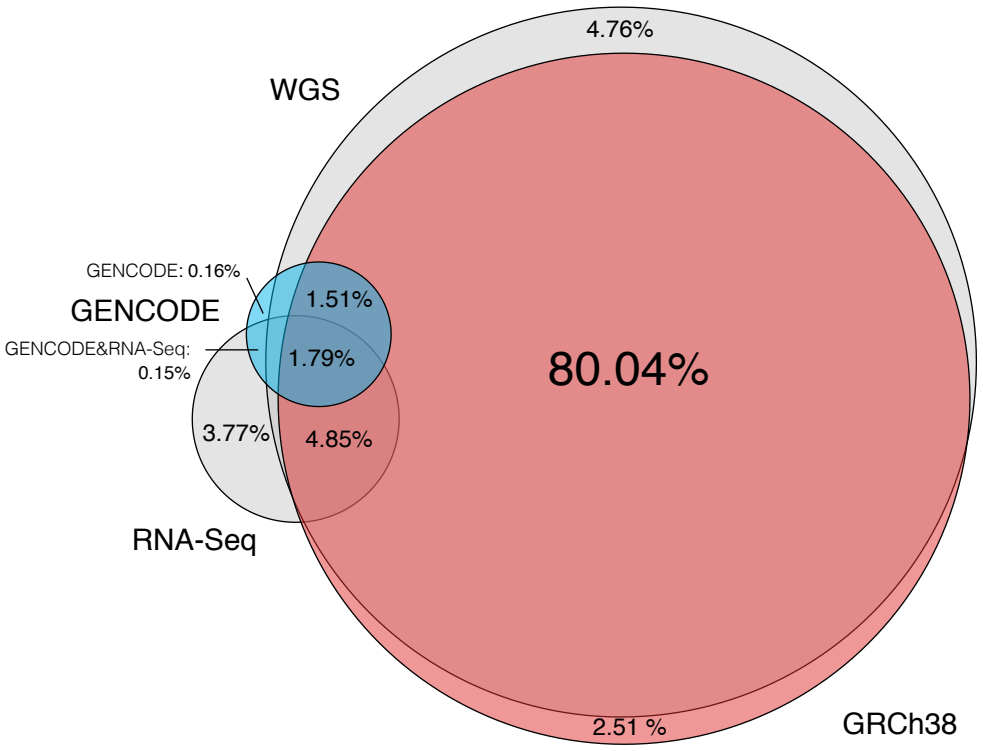

Figure 3. The diversity of non-reference k-mers is greater for RNA-Seq than for whole genome sequencing (WGS). Intersection of k-mers between Gencode transcripts, the human genome (GRCh38), RNA-Seq and WGS data. RNA-Seq and WGS data originate from the same lymphoblastoid cell line (HCC1395).

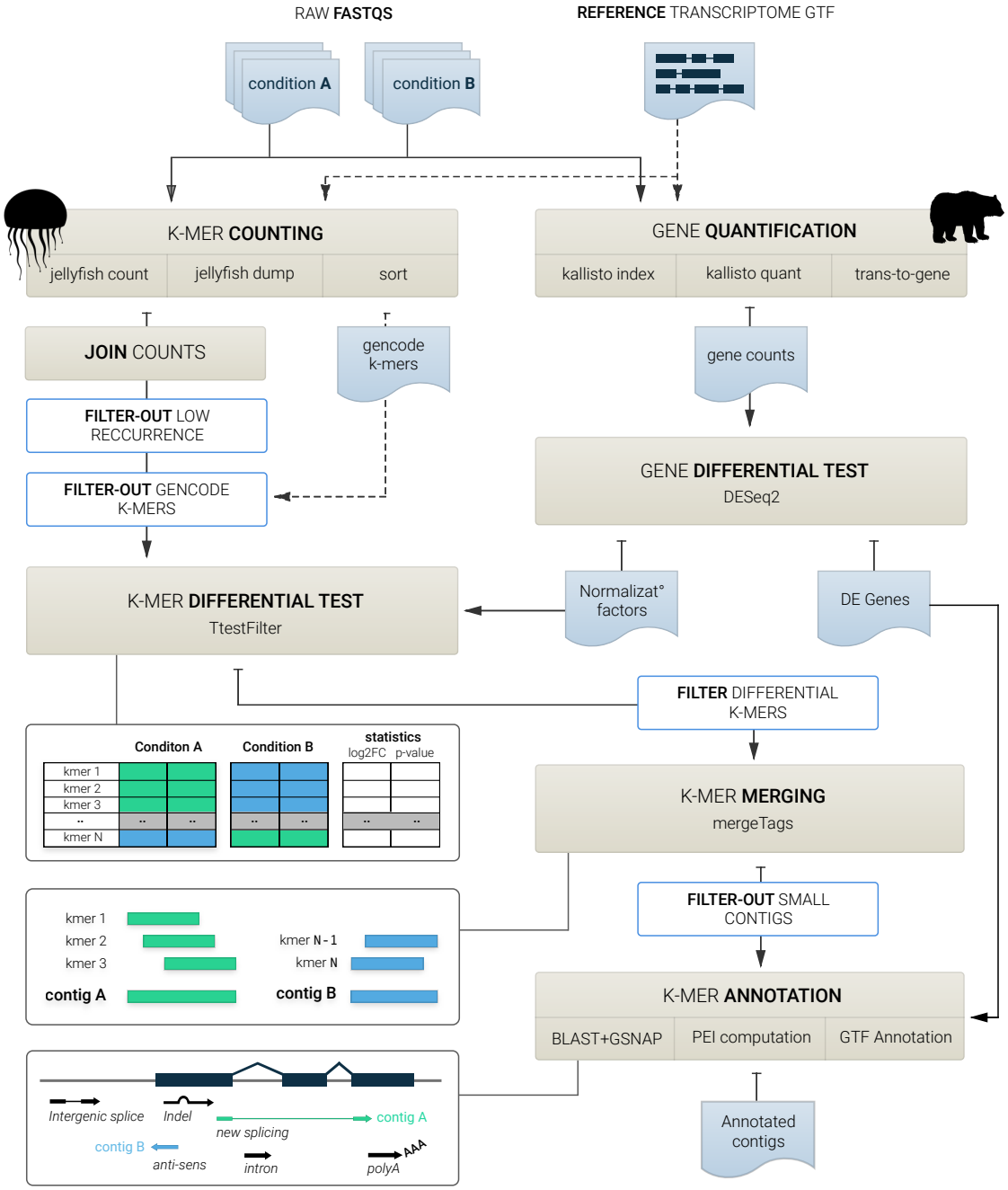

Figure 4. The DE-kupl pipeline for the discovery and analysis of differentially expressed k-mers. Raw FASTQ files are processed with two independent quantification procedures. On the left side, we use Jellyfish to count k-mers in all libraries. K-mers counts are then joined into a count matrix and filtered for low-recurrence and matching to the reference transcriptome. On the right side, FASTQs are processed with Kallisto to estimate gene counts based on the reference transcriptome. Gene counts are then processed with DESeq2 to compute normalization factors (NF) and differentially expressed genes. K-mer counts from the left side procedure are then normalized with NF computed from gene counts and the DE procedure is applied. Finally overlapping DE k-mers are merged into contigs and annotated based on their alignment to reference and overlap with annotations.

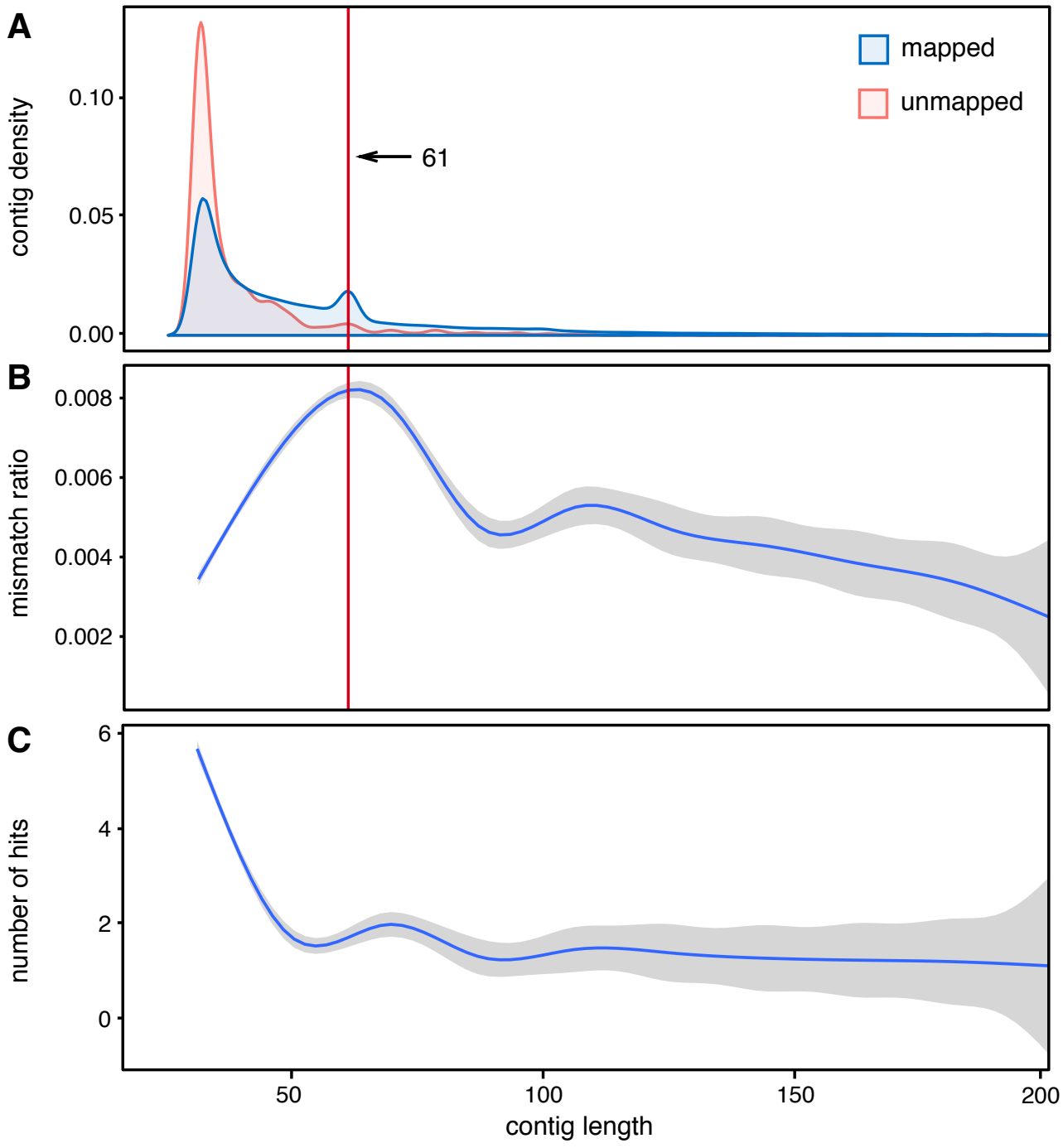

Figure 5. Specificity of differentially expressed contig. **A**. Density estimation plot of contigs length between mapped and unmapped contigs. The red line indicates contigs assembled from *k* k-mers and likely corresponding to SNVs. **B**. Mismatch ratio (number of mismatch / contig size) as a function of contig length. **C**. Number of hits in the reference genome as function of contig length. The B and C curves were obtained using a smoothing function.

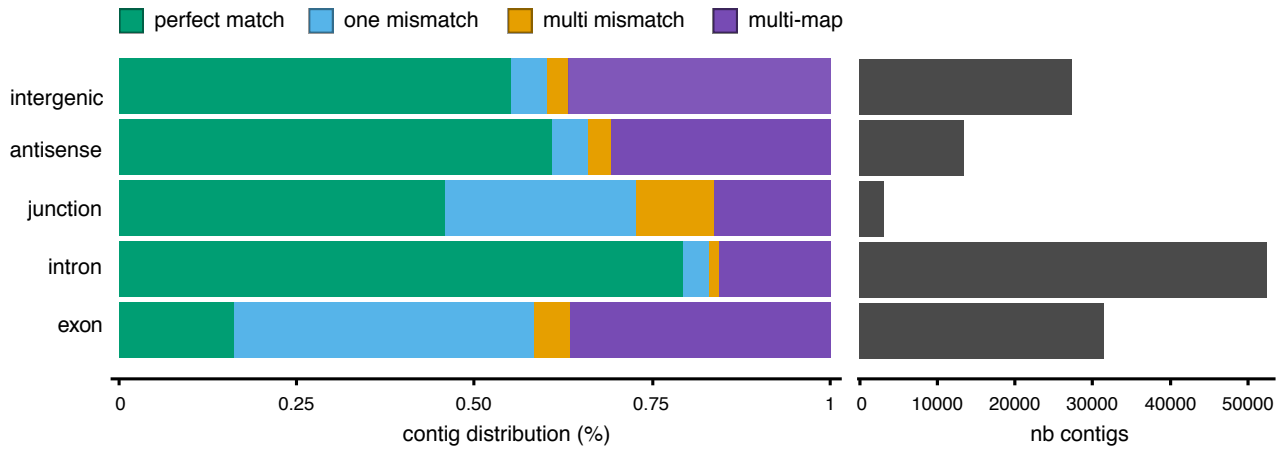

Figure 6. Genomic location of differentially expressed contigs. Contigs are separated by genomic location as follows. Exons: alignment overlaps a known exon. Intron: alignement overlaps an intron of an annotated gene. Junction: alignement is splitted (at least once). Antisense: alignment overlaps an annotated gene located on the opposite strand. Intergenic: mapped elsewhere in the genome. Contigs with a single location are labeled according of their number of mismatches: 0 as ”perfect match”, 1 as ”one mismatch”, >1 as ”multi mismatches”. Contigs having multiple locations on the genome are labeled as ”multi-map”.

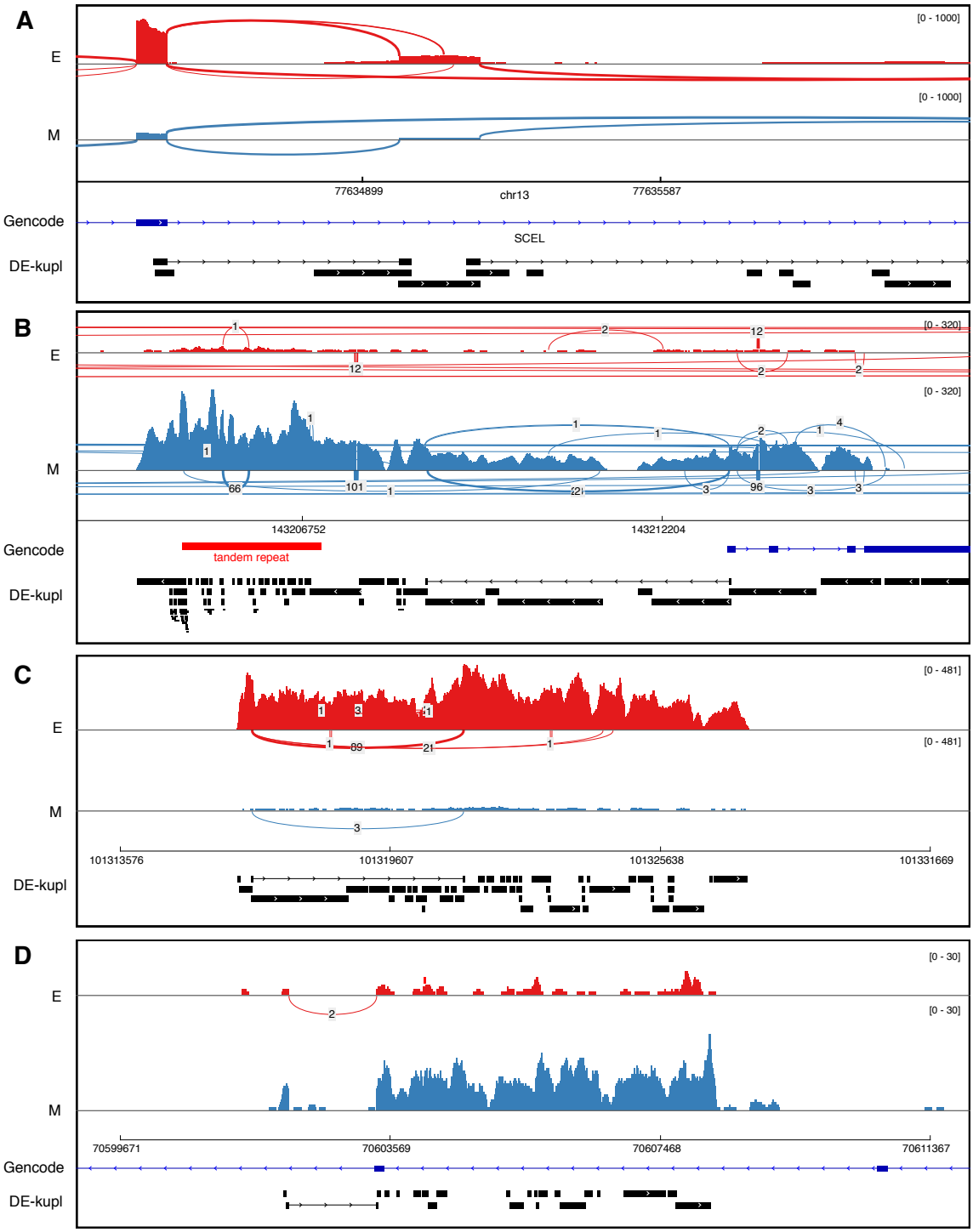

Figure 7. Examples of DE contigs. Sashimi plots generated from IGV using BAM alignments produced with STAR[29]. Sample SRR2966453 from condition D0 is labed as ”E” (epithelial). Sample SRR2966474 from condition D7 is labeled as ”M” (mesenchymal). Annotations from Gencode and DE-kupl DE contigs are shown at the bottom of each frame. **A**. New splicing variant involving an unannotated exon, overexpressed in condition ”E”. **B**. Tandem repeat at chr8:143,204-870-143,206,916 (red region) that is overexpressed in condition ”M” *vs*. ”E”. Note that the overexpressed tandem repeat is part of a larger overexpressed unannotated locus. **C**. A novel lncRNA overexpressed in condition ”E”. **D**. A novel antisense RNA. RNA-seq reads are aligned in the forward orientation while the gene at this locus is in the reverse orientation. Note that the annotated gene is not expressed.