bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

RESEARCH

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], Mélina Gallopin[3], Marc Gabriel[3,5], Jérémy Le Coz[3], Thérèse Commes[1,4] and Daniel Gautheret[3,5*]

*Correspondence: daniel.gautheret@u-psud.fr
[3]Institute for Integrative Biology of the Cell, Université Paris-Sud, CNRS, CEA, Gif sur Yvette, France
[5]Institut de Cancérologie Gustave Roussy Cancer Campus (GRCC), AMMICA, INSERM US23/CNRS UMS3655, Villejuif, France
Full list of author information is available at the end of the article

Abstract

Each individual cell produces its own set of transcripts, which is the combined result of genetic variation, transcription regulation and post-transcriptional processing. 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 produces a variety of transcripts that depart from the reference transcriptome and should be properly identified. To address this challenge, we introduce a k-mer-based software protocol for capturing local RNA 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 raw data 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 is subsequently assigned to biological events such as differential lincRNAs, antisense RNAs, splice and polyadenylation variants, introns, expressed repeats, and SNV-harboring or exogenous RNA. We applied DE-kupl to public RNA-seq datasets, including an Epythelial-Mensenchymal Transition model and different human tissues. DE-kupl identified abundant novel events and showed excellent reproducibility when applied to independent deep sequencing experiments. DE-kupl is a new paradigm for analyzing differential RNA-seq data with no preconception on target events, which can also provide fresh insights into existing RNA-seq repositories.

Background

Successive generations of RNA sequencing technologies have established since the 1990's that organisms produce a highly diverse and adaptable set of RNA molecules. Modern transcript catalogs such as Gencode[1] now reach hundreds of thousands of transcripts, reflecting widespread pervasive transcription and alternative RNA processing. However, in spite of years of high throughput sequencing efforts and bioinformatics analysis, we contend that large areas of transcriptomic information remain 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 [2], circularization [3] and cleavage/degradation by various nucleases [4, 5]. Thirdly, an essential, yet often overlooked, source of transcript diversity, is genomic variation. Polymorphism and struc-

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

tural variations within transcribed regions produce RNAs with single nucleotide variations (SNVs), tandem duplications or deletions, transposon integrations, unstable microsatellites or fusion events. These events are major sources of transcript variation that can strongly impact RNA processing, transport and coding potential.

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 [6, 7, 8, 9] rely on the alignment (or pseudo-alignment) of RNA-seq reads to a reference genome or transcriptome, while de novo assembly tools [10] 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 full range of variation observed in organisms 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 any type of event occurring in an RNA-seq library independently of alignment or transcript assembly. Our approach relies on k-mer indexing of sequence files, a technique that recently gained momentum in NGS data analysis [8, 9, 11, 12, 13]. To identify biologically meaningful transcript variation, our method selects k-mers with differential expression (DE) between two experimental conditions, hence its name: DE-kupl. Using public human RNA-seq datasets, we show that a large amount of RNA variation can be captured that is not represented in existing transcript catalogs. As proofs of concept, we applied DE-kupl to RNA-seq data from an Epythelial-Mensenchymal Transition (EMT) model and from different human tissues. DE-kupl identified abundant novel events and showed excellent reproducibility when applied to independent deep sequencing experiments.

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 [14]. Figure 1A-B compares k-mers from Gencode transcripts, the human genome reference and RNA-seq libraries from 18 different individuals [15] corresponding to three primary tissues (6 libraries/tissue). To minimize the risk of including k-mers containing 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.

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 [16], 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 existing RNA-seq analysis procedures in that it does neither "map-first" (a la Tuxedo suite [17]) or "assemble-first" (a la Trinity [18]) 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 exogenous 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

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

contains in the order of 10^8 distinct k-mers and an index for 50 individual samples can reach billions of k-mers. We selected the Jellyfish tool for counting k-mers [14] as it presents very fast computing times and allows to store the full index on disk for further query.

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 captured efficiently by conventional, reference-based protocols [8, 9]. Discarding these k-mers enable us to ignore the overwhelming signal caused by known transcripts and focus on expressed regions harboring differences from the reference transcriptome.

Two modes are available to perform differential analysis of k-mers (Figure S10 & Methods): The *t*-test filter mode is fast and has lower sensitivity, *i.e.* it retrieves only the most significantly differentially expressed k-mers. The DESeq2-based mode [19] is slower, more sensitive and is recommended for small samples (fewer than 6 *vs.* 6 samples). Whenever possible, key steps of the procedure (k-mer table merging, *t*-test, k-mer assembly) were written in C, enabling the whole procedure to run on a relatively standard computer in a reasonable amount of time.

Discovery and assembly of differential RNA contigs with DE-kupl

To assess the capacity of DE-kupl to discover novel differential events, we applied the procedure to 12 RNA-seq samples from an EMT cell-line model [20], in which NSCLC cells were induced 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). The full DE-Kupl procedure was completed in about 4 hours in the *t*-test mode (single-threaded), and 6.5 hours in the DESeq2 mode (multi-threaded), using 8 computing cores, 54 GB RAM and 7 to 42 GB of hard disk space (Table 1). Recurrence filters efficiently reduced k-mer counts from 707M to 92.5M and the Gencode filter further reduced counts to 40.3M. Differential analysis using the *t*-test mode eventually retained 3.8M k-mers that were assembled into 133.690 contigs (Table 2). The resulting contigs ranged in size from 31 bp (corresponding to an "orphan" k-mer) to 3.6 kbp, with a major peak of short 31-40 bp contigs and a minor peak around 61 bp (Fig 5A).

Almost all (99.2%) of the 133k DE contigs mapped to the human genome. Mapping revealed that most 61 bp contigs result from assembly of 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 61 bp (Fig 5B). Contigs that do not map to the human genome are generally shorter than mapped contigs (Fig 5A), indicating a lower signal-to-noise ratio in unmapped contigs. 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 (not shown).

Analysis of contig locations reveals distinct contig classes. Most contigs are located in annotated introns and exons (Fig 6), however intronic contigs are predominantly

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

exact matches while exonic contigs are predominantly mismatched. 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.

Assigning contigs to biological events

We assigned DE contigs generated from the EMT dataset to 11 classes of potential biological events, using the rule set described in Table 3. Since intragenic DE contigs may result from a mere over/under-expression of their host gene and do not necessarily reflect a differential usage (DU) of transcript isoforms, we implemented a simple strategy to distinguish the two situations based on the differential status of the host gene (Methods). We made this distinction for splicing, polyadenylation, SNVs and intron retention (Table 3).

From the total set of 133k DE contigs (supplemental data), we extracted about 6900 contigs matching our rule set for either event class (Table 3). We noted that a single event often generates multiple contigs. We thus further grouped contigs into "loci", defined as independent annotated genes or intergenic regions harboring one or more contigs (Table 3). We describe below the main classes of events identified.

Differential splicing. Analysis of split-mapped contigs found evidence of potentially novel differential splice variants in 1040 contigs (Table 3, Fig 7A,B,C). Note that this class excludes SNV-containing contigs, as to avoid known splice variants associated to genetic polymorphism. Furthermore, 171 of these contigs were classified as DU, suggesting 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 bp difference from the annotated reference, which escaped prior annotation (see eg. Fig S1).

Differential 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 140 such poly-A terminated contigs, 105 (75%) contained an AATAAA or variant polyadenylation signal (Table S1), indicating they result from actual polyadenylated transcripts (Table 3, line "PolyA"). Note these are not necessarily novel polyadenylation sites since polyadenylated transcripts always create k-mers that differ from the reference transcriptome and are hence retained by DE-kupl. Indeed, only 6 of the 105 poly-A contigs mapped intergenic regions. Furthermore, only 17 poly-A contigs mapped to genes with no differential expression ("polyA DU" in 3), and these had relatively poor fold change values (Table S11) raising doubt on their DU status. Altogether this analysis demonstrates that DE-kupl is able to capture *bonafide* polyadenylated transcripts present in the sequencing reads, however we did not observe any clear case of differentially polyadenylated genes in the experiment studied.

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

LincRNA. We identified a subset of 809 DE contigs (282 loci) corresponding to potential long intergenic non-coding RNAs (Table 3, line "lincRNA"). Criteria for lincRNAs were contigs of size > 200nt mapped to an intergenic locus. Visual inspection revealed clear lincRNA-like patterns, whith contigs clustered into well defined transcription units with abundant read coverage and evidence of splicing (Fig 7C, Fig S2). DE-kupl is thus an effective tool for the identification of novel differentially expressed 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 thus be used for identifying antisense RNAs (asRNAs) and disembiguating loci with intricated expression on both strands. We identified 356 contigs from 156 loci mapping to the reverse strand of an annotated gene (Table 3, line "asRNA"). 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 (either base substitution or indel) and mapping to an exon whose host gene was not measured as differentially expressed (Table 3, line "SNV DU"). This was a less than perfect procedure, as we did not explicitly test for a switch in allelic balance among the two conditions. Yet, among the 717 contigs identified, some displayed strong apparent changes in allelic balance between the E and M conditions (*eg.* Fig S4). The ability of DE-kupl to capture differential SNV between datasets may be particularly interesting when looking for recurrent mutations in subpopulations.

Intron retention and other intronic events. As highly expressed transcripts often carry intronic byproducts, we expected DE-kupl to turn out a lot of "parasitic" intronic contigs. Indeed 1909 contigs mapped to intronic loci (Table 3, line "intron"). We thus focused on intronic k-mers from genes that were not DE (line "intron DU"). This filter identified 559 intronic contigs from 185 different genes. Inspection of read mapping at these loci revealed clear instances of novel skipped of extended 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 [21] (Fig S6). Therefore DE-kupl can be used for screening a wide variety of exon/intron processing events in addition to alternative splicing.

Expressed repeats. Assessing the expression of human repeats by conventional RNA-seq analysis protocols is difficult, as ambiguous alignments render repeat regions "unmappable" [22]. Since DE-kupl first measures expression independently of mapping, we were able to collect and analyze differential contigs with multiple genome hits. 4968 contigs of size 50nt or larger have multiple hits (not shown), and 1141 are repeated more than 5 times (Table 3, line "repeat"). RepeatMasker [23] found 664 out of these 1141 sequences to match known repeats, mostly LINEs, LTRs and SINEs (Fig S8). Further inspection showed that most of the remaining

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

multiple-hit contigs correspond to unannotated repeats or low complexity regions. One of the most striking differential repeats is an unannotated 22x66 bp tandem repeat, located about 2 Mbp from the chromosome 8 telomere. This repeat is found about 50-fold overexpressed in the Mesenchymal condition (Fig 7B, S7). These results indicate DE-kupl can serves as a screen for differential expression or activation of endogenous viral sequences and other repeat-containing transcripts.

Unmapped contigs. Finally, we analyzed DE contigs that did not map the human genome. 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 levels of foreign RNA vary across samples. In this test set, where all samples come from an *invitro* cell line, we did not expect to observe such a phenomenon. Indeed, out of 114 unmapped contigs of size > 50 bp (Table 3, line "Unmapped"), 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 were low complexity sequences or aligned to non-human primate sequences, indicating possible contamination and/or misassemblies.

DE-kupl event detection is reproducible across independent datasets

We cross-validated DE-kupl findings using two independent human RNA-seq datasets extracted from the Genotype-Tissue Expression (GTEx) [24] and the Human Protein Atlas (HPA) [15]. DE contigs were first obtained by running DE-kupl on 8 colon vs 8 skin libraries from GTEx. Events were classified as above into intron retentions, lincRNAs, polyadenylation sites, repeats, splice sites and unmapped. The 100 top events from each class (50 for class 'unmapped') were extracted and their k-mer labels saved as a sequence file. We then counted the occurrence of each k-mer in colon and skin libraries from the HPA project and applied DEseq2 [19] to evaluate the significance of the expression change between colon and skin (see Methods). 79% of the 550 DE k-mers identified in GTEX were also significantly DE in the HPA data (Figure 8). Each event class showed clear reproducibility, with particularly strong effects in lincRNAs and splice variants. This demonstrates that novel events identified by DEkupl are reproducible across independent datasets in spite of independent RNA extraction, library preparation and sequencing protocols.

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. DE-kupl explores all k-mers in the input RNA-seq files (*vs.* only k-mers from annotated transcripts in recent software [8, 9]) 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 software for similarly sized datasets. Another key aspect of our protocol that rendered a "full k-mer" analysis tractable was applying successive

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 also 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.

Contrarily to popular RNA-seq analysis software, DE-kupl does not attempt full-length transcript assignment or assembly but focuses instead on local transcript variations. Indeed, we do not consider full-length transcript analysis to be realistic when screening for unspecified RNA variation, since the combinatorial nature of genomic, transcriptomic and post-transcriptomic events would require an indefinitely expanding transcript catalog. In some way DE-kupl is closer in spirit to methods analyzing local RNA-seq coverage such as RNAprof [25] and DERfinder [26], with the notable exception that DE-kupl does not involve mapping and thus avoids mapping-related pitfalls while considerably widening the range of detectable events. Another important benefit of the k-mer strategy is that k-mers representing events of interest can be used to efficiently assess the occurence of similar events in the huge public compendium of RNA-seq data.

We showed that DE-kupl is able to detect a wide range of differential transcription and RNA processing events. Although specialized software may perform better at assessing specific event classes such as differential splicing, to our knowledge no software provides such an extensive screen. As differential RNA-seq analysis is often conducted with an exploratory spirit, we argue that it is preferable to cast a wide net with no preconception on target events, using DE-kupl along with a conventional gene-by-gene differential expression analysis. Note that DE-kupl might also be an interesting option for exploring other types of NGS data such as small-RNA-seq, ChIP-seq or CLIP-seq, with simple adjustments of k-mer size and event annotation rules.

In this proof of concept study, we focused on RNA-seq libraries from a cell line, where no genetic polymorphism was expected among samples. The next step will be application to libraries from multiple individual organisms. Although k-mer diversity will be higher in such datasets, preliminary tests with RNA-seq data from 60 human tumors were completed successfully on a single computer server (data not shown). Analysis of patient samples open exciting perspectives. For instance, the ability of DE-kupl to simultaneously detect genetic variation and RNA expression/processing events may serve as a basis for studying genotype-phenotype relations. Analysis of patient RNA-seq data may also reveal event classes not explored in this work, such as fusion transcripts and circular RNAs.

Methods

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

RNA-seq data for bone marrow, skin and colon were 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,

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 reference 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 (version 1.1.2). Unmapped k-mers were mapped a second time with Bowtie to the GRCh38 genome reference. Reads with 3 or more mismatches are not mapped by Bowtie 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 [16]. K-mers were counted in these libraries with the same procedure as above. In order to reduce noise from sequencing errors, k-mers with only one occurrence were filtered out.

DE-kupl Implementation

The DE-kupl pipeline (Fig S12) is implemented using the Snakemake [27] 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, as described hereinafter.

K-mer counting

Raw sequences (FASTQ files) are first processed with the `jellyfish count` command of the Jellyfish software, which produces one index (a disk representation of the Jellyfish hash-table) for each sequence library. For stranded RNA-seq libraries, reads in reverse direction relative to the 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 is built, we use the `jellyfish dump` command to output the raw-counts in a two column text file, which contains at each line a k-mer and its number of occurrences. Raw counts are then sorted alphabetically by k-mer sequence with the Unix `sort` command.

K-mer filtering

All samples counts are joined together using the `dekupl-joinCounts` binary to produce a single matrix will all k-mers and their abundance in all samples. Given an integer $a \geq 0$, we define the *recurrence* of a k-mer x as the number of samples where x appears more than a times, i.e. $recurrence(x, a) = \sum_{i=1}^{n} \mathbb{1}_{\{x_i > a\}}$, where n is the total number of samples and x_i is the number of times the k-mer x appears in sample i. The k-mer filtering step involves two user-defined parameters: an integer `min_recurrence_abundance` and an integer `min_recurrence` such that a k-mer x is filtered out if $recurrence(x, \texttt{min_recurrence_abundance}) < \texttt{min_recurrence}$, i.e. if the k-mer x appears more than `min_recurrence_abundance` times in fewer than

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

`min_recurrence` of the samples. Usually `min_recurrence` is set to the number of replicates in each conditions, and `min_recurrence_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-mers.

Differential k-mer expression

Prior to differential analysis, we compute normalization factors (NFs) using the "median ratio method" [28] on the table of k-mers after recurrence filter: for each sample, the NF is the median of the ratios between sample counts and counts of a pseudo-reference obtained by taking the geometric mean of each k-mer across all samples. To avoid dealing with the complete table of k-mers, we extracted a random subset of 30% of the k-mers and computed NFs on this subset. Computing NFs on the complete table of k-mers, on the table of k-mers after recurrence and Gencode filters, or on the table of transcripts abundances produced by Kallisto [8] led to similar values (Fig S9).

To perform differential analysis, two options are implemented (Fig. S10). The first option is to apply a t-test for each k-mer on the log transformed counts, normalized with the previously computed NF. Transformation of raw counts in conjunction with linear model analysis have been successfully used for differential analysis of counts [29]. We perform the t-test independently on each k-mer and avoid complex variance modeling strategies to reduce execution time of the analysis. The t-test option has been implemented in C in the `dekupl-TtestFilter` binary. Note that this t-test option is not appropriate for small samples [30]. To increase the power of the analysis, in particular for small samples (typically less than 6 vs 6 libraries), we strongly advise to use the second option based on generalized linear model, implemented in the R package `DESeq2` [19]. On top of modeling raw counts (normalization or prior log-transformation of the counts is not required), this approach performs information sharing across k-mers to improve variance estimation and differential analysis results. However, given the large number of k-mers, we do not apply this approach on the complete matrix of k-mer counts. We divide the matrix of k-mer counts into random chunks of approximately equal size (around one million k-mers) and apply the DESeq2 approach on each chunk independently. For each chunk, previously computed NFs are used as an input of the method, and are not computed independently on each chunk. Raw p-values, not adjusted for multiple testing, are collected as an output for each chunk, and merged into one single vector containing the raw p-values for all k-mers to test. Subsequently, raw p-values obtained from either the t-test or the DESeq2 test are adjusted for multiple comparisons using the Benjamini-Hochberg procedure [31] and k-mers with adjusted p-values above a user-set cutoff are filtered out.

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-mers overlapping a splice junction or SNV). To this aim, we developed our own procedure called `mergeTags`, which works as follows: we first identify all exact $k-1$ prefix-suffix overlaps between k-mers. We consider only k-mers

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

that overlap with exactly one other k-mer, and merge all pairs of k-mers involved in such overlaps. For example, given the set of k-mers : ATG, TGA, TGC, CAT, the following contigs are produced : $contigs = CATG, TGA, TGC$. We repeatedly merge contigs that overlap exactly over $k-1$ bp with exactly one other contig. We then repeat this assembly process with $k-2$ exact prefix-suffix overlaps, using as input the contigs produced at the previous step, and so forth for increasing values of i such that $k-i > 15$ bp (default value). Finally, a set of DE contigs is produced and each contig is labelled by its constitutive k-mer of lowest p-value. This assembly procedure is implemented in C in the `dekupl-mergeTags` binary.

Contig Annotation

Finally, DE contigs are annotated in order to facilitate biological event identification. First, contigs are aligned with BLAST [32] against Illumina adapters. Contigs matching adapters are discarded. Retained contigs are further mapped to the reference Hg38 human genome using the GSNAP short read aligner [33], which showed the best speed/sensitivity ratio for aligning both short and long contigs in internal tests (not shown). 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 outputs. Alignment coordinates are compared with Ensembl (v86) annotations (in GFF3 format) using BEDTools [34] and a set of locus-related features is extracted. The final set of annotated features (Table S3) is reported in a contig summary table. The annotation procedure generates two additional files: a "per locus" summary of contigs (one line per genic or intergenic locus), 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.

In parallel to k-mer counting and filtering, we analyze the RNA-Seq data libraries a conventional differential expression protocol. Reads are processed with Kallisto [8] to estimate transcript abundances. Transcript-level counts are then collapsed to the gene-level and processed with DESeq2 [19] to produce a set of differentially expressed genes. This information is stored in the contig summary table and used later on for defining events with differential usage ("DU" in Table 3).

DE-kupl run on EMT data

DE-kupl was run using RNA-seq libraries from [20] retrieved on the GEO web site under accession GSE75492. For stage "E" we used libraries GSM1956974, GSM1956975, GSM1956976, GSM1956977, GSM1956978, GSM1956979, and for stage "M" GSM1956992, GSM1956993, GSM1956994, GSM1956995, GSM1956996, GSM1956997. DE-kupl parameters were `kmer_length 31`, `min_recurrence 6`, `min_recurrence_abundance 5`, `pvalue_threshold 0.05`, `lib_type stranded`, `diff_method Ttest`. Output files are provided as supplementary material. The DE-kupl contig summary table was analyzed interactively using R commands to

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

extract lists of contigs based on the filtering rules described in Table 3. Visualization of selected contigs was performed with IGV [35], using the bed file produced by DE-kupl and read mapping files produced by STAR [36].

Cross-validation

DEkupl was applied to 8 skin and 8 colon libraries from GTEx [24]: skin library IDs: SRR1308800, SRR1309051, SRR1309767, SRR1310075, SRR1311040, SRR1351501, SRR1400467, SRR1479595; colon library IDs: SRR1316343, SRR1396146, SRR1397292, SRR1477732, SRR1488307, SRR807751, SRR812697, SRR819486. DE-kupl parameters were `kmer_length 31`, `min_recurrence 6`, `min_recurrence_abundance 5`, `pvalue_threshold 0.05`, `lib_type unstranded`, `diff_method Ttest`. DE-kupl contigs were interactively classified using R commands, applying the same rules as in Table 3. Classes asRNA and SNV-DU were not included since asRNA identification is not possible using the unstranded GTEx and HPA libraries, and we had no reason to expect common SNVs with differential usage in this dataset. DE contigs were sorted by fold-change and k-mer labels of the top 100 DE contigs in each class were extracted (50 for class 'unmapped' due to lower event number). DE-kupl output files and selected k-mers are provided in supplementary material. For cross-validation, we used the same 6 skin and 6 colon RNA-Seq data as in Figure 1 (10.1126/science.1260419, E-MTAB-2836). K-mers were counted in each library using Jellyfish with options `k = 31` and `-C` (canonical k-mers) as GTEx data were unstranded. All k-mers selected from the GTEx analysis were queried against the jellyfish databases using `jellyfish query` command. Finally the extracted k-mers counts were processed with DESeq2 [19] and the resulting adjusted p-values were plotted for each event class (Figure 8).

Availability

The DE-kupl software, documentation and supplemental material presented herein are available at `https://transipedia.github.io/dekupl/`.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

JA, NP, RC, MS, MeG, TC and DG designed the study. JA, MaG, MeG and JLC developed the code. JA, NP, RC, MS, MeG, TC and DG analyzed the results. DG and JA drafted the manuscript and all authors proofread and corrected the manuscript.

Acknowledgements

We thank Damien Drubay for useful statistical discussions and Emilie Drouineau for setting up the DE-kupl Docker container. This project was supported in part by grants "Plan Cancer – Systems Biology" #bio2014-04 and ANR-10-INBS-0009 (France Génomique) from Agence Nationale pour la Recherche to DG. GTEx data was downloaded from the dbGaP web site under phs000178/GRU.

Author details

[1] INSERM U1183 IRMB, Université de Montpellier, Hopital St Eloi 80 avenue Augustin Fliche, 34295 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. [5]Institut de Cancérologie Gustave Roussy Cancer Campus (GRCC), AMMICA, INSERM US23/CNRS UMS3655, Villejuif, France.

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., Hubbard, T.J.: GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research **22**(9), 1760–1774 (2012). doi:10.1101/gr.135350.111
2. 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
3. Chen, L.-L.: The biogenesis and emerging roles of circular RNAs. Nature Reviews Molecular Cell Biology **17**(4), 205–211 (2016). doi:10.1038/nrm.2015.32
4. Kirchner, S., Ignatova, Z.: Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. Nature Reviews Genetics **16**(2), 98–112 (2015). doi:10.1038/nrg3861
5. Dieci, G., Preti, M., Montanini, B.: Eukaryotic snoRNAs: A paradigm for gene expression flexibility. Genomics **94**(2), 83–88 (2009). doi:10.1016/j.ygeno.2009.05.002
6. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L.: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology **28**(5), 511–515 (2010). doi:10.1038/nbt.1621
7. 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
8. Bray, N.L., Pimentel, H., Melsted, P., Pachter, L.: Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology **34**(5), 525–527 (2016). doi:10.1038/nbt.3519
9. 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
10. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A.: Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology **29**(7), 644–652 (2011). doi:10.1038/nbt.1883
11. Nordström, K.J.V., Albani, M.C., James, G.V., Gutjahr, C., Hartwig, B., Turck, F., Paszkowski, U., Coupland, G., Schneeberger, K.: Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. Nature Biotechnology **31**(4), 325–330 (2013). doi:10.1038/nbt.2515
12. 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
13. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M.: Mash: Fast genome and metagenome distance estimation using MinHash. Genome Biology **17**, 132 (2016). doi:10.1186/s13059-016-0997-x
14. Marçais, G., Kingsford, C.: A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics **27**(6), 764–770 (2011). doi:10.1093/bioinformatics/btr011
15. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C.A.-K., Odeberg, J., Djureinovic, D., Takanen, J.O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J.M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., Pontén, F.: Tissue-based map of the human proteome. Science **347**(6220), 1260419 (2015). doi:10.1126/science.1260419
16. Griffith, M., Griffith, O.L., Smith, S.M., Ramu, A., Callaway, M.B., Brummett, A.M., Kiwala, M.J., Coffman, A.C., Regier, A.A., Oberkfell, B.J., Sanderson, G.E., Mooney, T.P., Nutter, N.G., Belter, E.A., Du, F., Long, R.L., Abbott, T.E., Ferguson, I.T., Morton, D.L., Burnett, M.M., Weible, J.V., Peck, J.B., Dukes, A., McMichael, J.F., Lolofie, J.T., Derickson, B.R., Hundal, J., Skidmore, Z.L., Ainscough, B.J., Dees, N.D., Schierding, W.S., Kandoth, C., Kim, K.H., Lu, C., Harris, C.C., Maher, N., Maher, C.A., Magrini, V.J., Abbott, B.S., Chen, K., Clark, E., Das, I., Fan, X., Hawkins, A.E., Hepler, T.G., Wylie, T.N., Leonard, S.M., Schroeder, W.E., Shi, X., Carmichael, L.K., Weil, M.R., Wohlstadter, R.W., Stiehr, G., McLellan, M.D., Pohl, C.S., Miller, C.A., Koboldt, D.C., Walker, J.R., Eldred, J.M., Larson, D.E., Dooling, D.J., Ding, L., Mardis, E.R., Wilson, R.K.: Genome Modeling System: A Knowledge Management Platform for Genomics. PLOS Computational Biology **11**(7), 1004274 (2015). doi:10.1371/journal.pcbi.1004274
17. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L.: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols **7**(3), 562–578 (2012). doi:10.1038/nprot.2012.016
18. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A.: De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols **8**(8), 1494–1512 (2013). doi:10.1038/nprot.2013.084
19. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology **15**(12), 550 (2014). doi:10.1186/s13059-014-0550-8
20. Yang, Y., Park, J.W., Bebee, T.W., Warzecha, C.C., Guo, Y., Shang, X., Xing, Y., Carstens, R.P.: 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**(11), 1704–1719 (2016). doi:10.1128/MCB.00019-16

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

21. 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**(2), 95–103 (2010). doi:10.1007/s00438-010-0556-1
22. Derrien, T., Estellé, J., Sola, S.M., Knowles, D.G., Raineri, E., Guigó, R., Ribeca, P.: Fast Computation and Applications of Genome Mappability. PLOS ONE **7**(1), 30377 (2012). doi:10.1371/journal.pone.0030377
23. Smit, A., Hubley, R., Green, P.: RepeatMasker Open-4.0 (2013-2015)
24. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N.J., Nicolae, D.L., Gamazon, E.R., Im, H.K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E.T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson, J.M., Wilder, E.L., Derr, L.K., Green, E.D., Struewing, J.P., Temple, G., Volpi, S., Boyer, J.T., Thomson, E.J., Guyer, M.S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T.R., Koester, S.E., Little, A.R., Bender, P.K., Lehner, T., Yao, Y., Compton, C.C., Vaught, J.B., Sawyer, S., Lockhart, N.C., Demchok, J., Moore, H.F.: The Genotype-Tissue Expression (GTEx) project. Nature Genetics **45**(6), 580–585 (2013). doi:10.1038/ng.2653
25. 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**(1), 59–67 (2016). doi:10.1080/15476286.2015.1118604
26. 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**(3), 413–426 (2014). doi:10.1093/biostatistics/kxt053
27. Köster, J., Rahmann, S.: Snakemake—a scalable bioinformatics workflow engine. Bioinformatics **28**(19), 2520–2522 (2012). doi:10.1093/bioinformatics/bts480
28. Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome Biology **11**, 106 (2010). doi:10.1186/gb-2010-11-10-r106
29. Law, C.W., Chen, Y., Shi, W., Smyth, G.K.: Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology **15**, 29 (2014). doi:10.1186/gb-2014-15-2-r29
30. Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G., Guedj, M.: Should we abandon the t-test in the analysis of gene expression microarray data: A comparison of variance modeling strategies. PloS One **5**(9), 12336 (2010). doi:10.1371/journal.pone.0012336
31. Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) **57**(1), 289–300 (1995)
32. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L.: BLAST+: Architecture and applications. BMC Bioinformatics **10**, 421 (2009). doi:10.1186/1471-2105-10-421
33. Wu, T.D., Nacu, S.: Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics **26**(7), 873–881 (2010). doi:10.1093/bioinformatics/btq057
34. Quinlan, A.R., Hall, I.M.: BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England) **26**(6), 841–842 (2010). doi:10.1093/bioinformatics/btq033
35. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P.: Integrative genomics viewer. Nature Biotechnology **29**(1), 24–26 (2011). doi:10.1038/nbt.1754
36. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: STAR: Ultrafast universal RNA-seq aligner. Bioinformatics, 635 (2012). doi:10.1093/bioinformatics/bts635

Figures

Tables

Table 1 DE-kupl parameters and resources used for analyzing EMT data (12 libraries) using the t-test and DESeq2 methods.

Parameter/Resources	value	
nb_threads	8	
min_recurrence	6	
min_recurrence_abundance	5	
pvalue_threshold	0.05	
lib_type	stranded	
	t-test	DESeq2
Max memory usage	54Gb	53Gb
Max disk used (1)	7Gb	42Gb
Running time (1)	4h 02m	6h 33m

(1): excluding reference genome and transcriptome indexing for the annotation step.

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

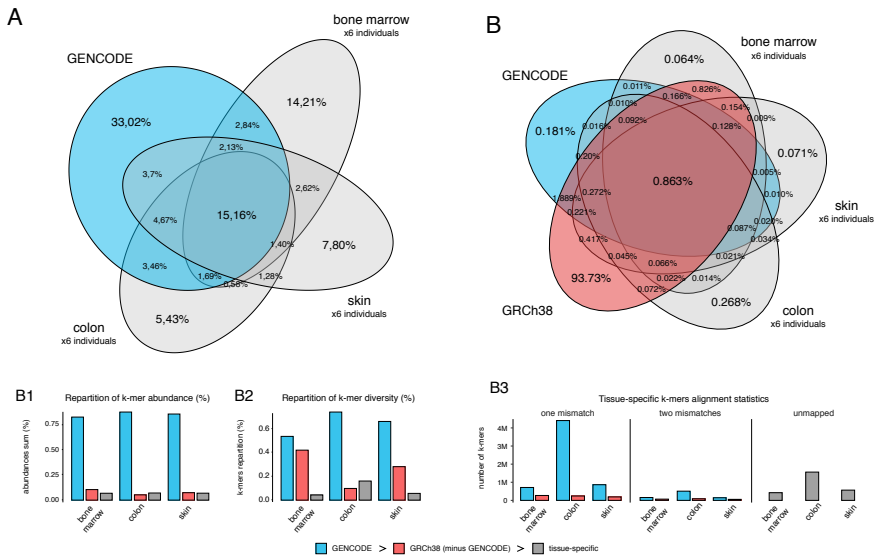

Figure 1 The diversity of 31-mers in RNA-Seq libraries exceeds that of reference sequences. A. Intersection of k-mers present in Gencode transcripts and RNA-Seq data from three tissues: bone marrow, skin and colon. The set of k-mers for each tissue was defined as the common k-mers shared by all six individuals. **B**. Intersection of k-mers present in Gencode transcripts, the reference human genome (GRCh38) and RNA-Seq data (same as in A). **B1**. Repartition of k-mer abundances for each tissue represented in A and B. K-mers shared with Gencode are labelled as "GENCODE", among other k-mers, those shared with the human genome are labelled as "GRCh38", the remaining 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, allowing up to 2 mismatches in a 31-mer.

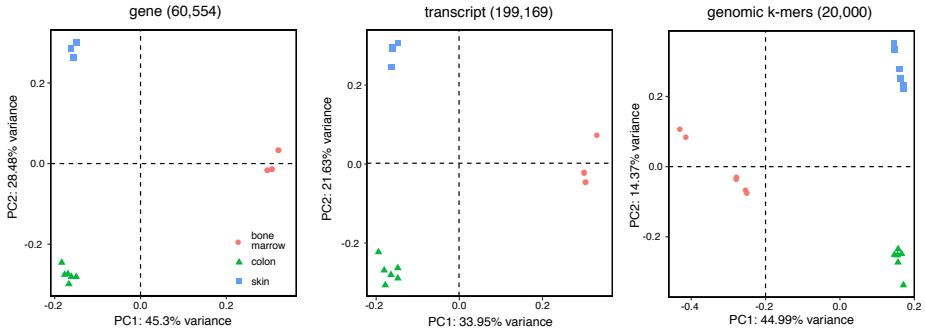

Figure 2 Principal Component Analysis on non-reference k-mers discriminates tissues. Samples are labeled according to their tissues (bone marrow, colon, skin). PCA were produced with normalized, log transformed counts. For genes and transcripts, counts were generated with Kallisto based on Gencode V25.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

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

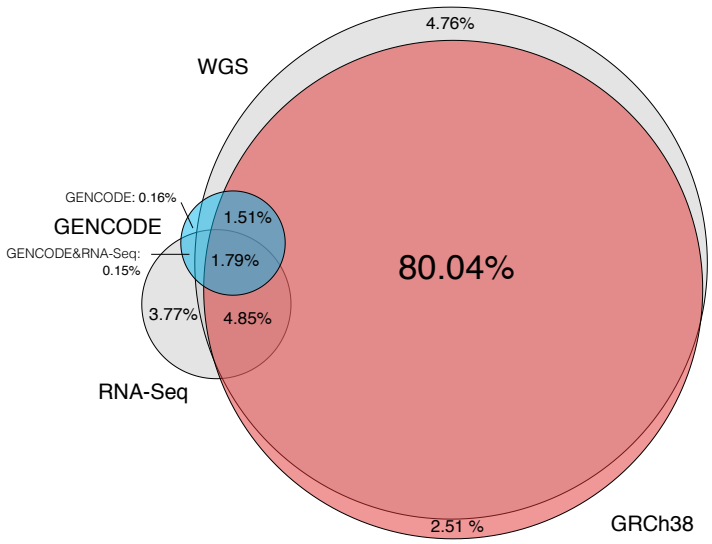

Figure 3 The diversity of non-reference k-mers is greater for RNA-Seq than for whole genome sequencing (WGS). Intersection of k-mers between Gencode transcripts, the human genome (GRCh38), RNA-Seq and WGS data. RNA-Seq and WGS data originate from the same lymphoblastoid cell line (HCC1395).

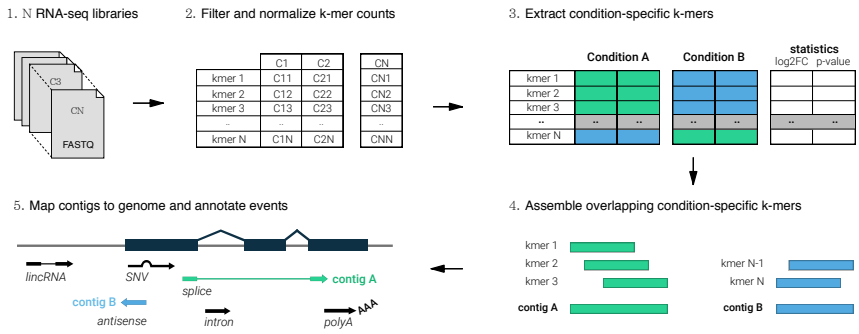

Figure 4 The DE-kupl pipeline for the discovery and analysis of differentially expressed k-mers. First, Jellyfish is applied to count k-mers in all libraries. K-mers counts are then joined into a count matrix and filtered for low-recurrence and matching to the reference transcriptome. Normalization factors are computed from raw K-mer 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 annotated genes

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

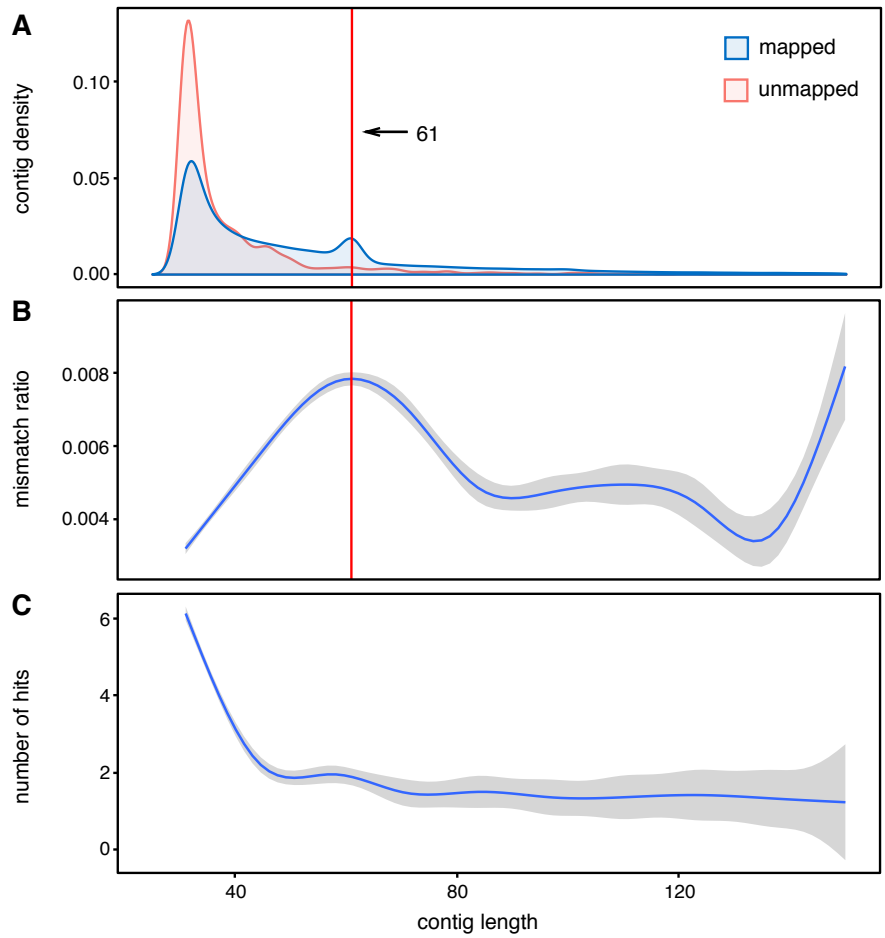

Figure 5 Specificity of differentially expressed contig. A. Density plot of contig lengths for mapped and unmapped contigs. The red line indicates contigs assembled from k k-mers and likely corresponding to SNVs. **B**. Mismatch ratio (number of mismatch / contig size) as a function of contig length.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

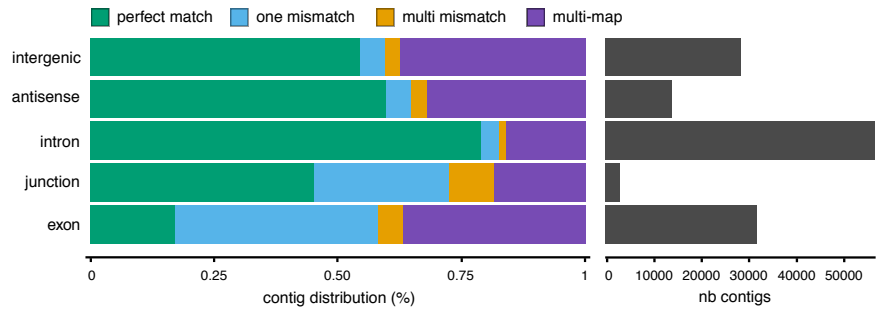

Figure 6 Genomic location of differentially expressed contigs. Contigs are separated by genomic location, according to their overlap with exons, exon-exon junctions, introns, antisense regions of annotated genes or intergenic regions. The right panel shows the total number of contigs in each class; the left panel shows contig distribution according to their alignment status: contigs with a single mapping location are labeled as "perfect match", "one mismatch" or "multi mismatches", contigs with multiple mapping locations are labeled as "multi-map".

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

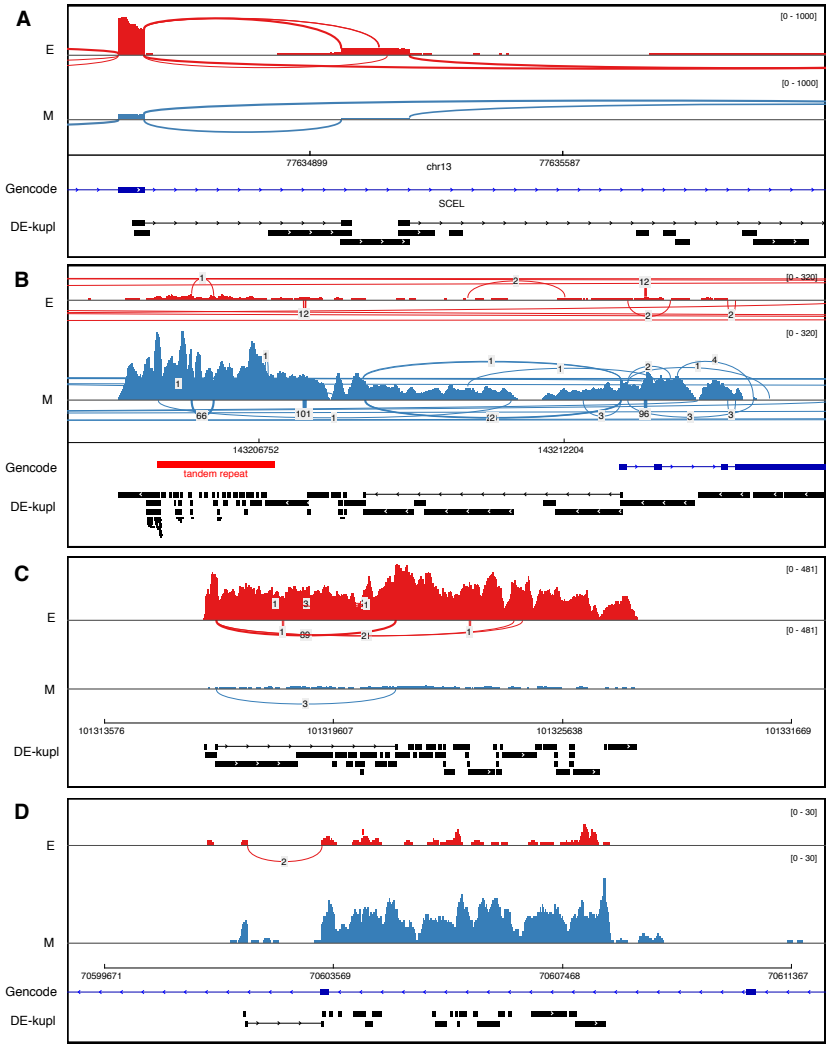

Figure 7 Examples of DE contigs. Sashimi plots generated from IGV using read alignments produced with STAR [36]. Sample SRR2966453 from condition D0 is labeled as "E" (epithelial). Sample SRR2966474 from condition D7 is labeled as "M" (mesenchymal). Annotations from Gencode and DE-kupl DE contigs are 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 lincRNA 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. The annotated gene is not expressed.

bioRxiv preprint doi: https://doi.org/10.1101/122937; this version posted June 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

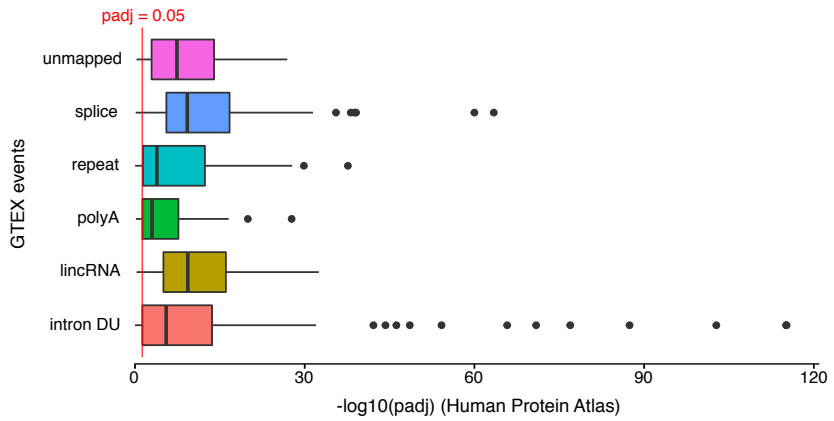

Figure 8 Cross validation of DE events in GTEx and Human Protein Atlas data. 550 DE contigs from 6 different event classes (intron with differential usage, lincRNA, polyA site, repeat, splice site, unmapped) were identified using DE-kupl on GTEx libraries from two human tissues (skin and colon). A representative k-mer from each contig was then tested for differential expression in skin and colon libraries in the Human Protein Atlas. Box plots represent distributions of DESeq2 adjusted P-values for all k-mers in the different classes. The red line shows the adjusted P-value cutoff of 0.05.

Table 2 DE-kupl pipeline results for the EMT experiment. Description of output files sequentially generated by DE-kupl. Number of kmers/contigs, correspond to the number of lines in each file.

Files	Description	Number of k-mers/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 all k-mer counts from all libraries with recurrence filters	92,525,450		1.9GB	
noGENCODE-counts.tsv.gz	Counts filtered with GENCODE k-mers	40,398,848		728MB	
		t-test	DESeq2	*t*-test	DESeq2
diff-counts.tsv.gz	Counts with differential expression test, filtered on adjusted P value	3,813,418	6,102,447	186MB	510MB
merged-diff-counts.tsv.gz	DE k-mers assembled into contigs	133,690	169,613	3.0MB	18MB

Table 3 Assignment rules for DE contigs. Each class of event is defined by a set of rules applied to annotated contigs. Keyword "DU" indicates Differential Usage. Column "Other rules" refers to the following: 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													
Event class	gene is diff	nb junctions	maps gene	maps AS gene	clipped 3p	is mapped	SNV	Exonic	Intronic	nb hit	contig length	other rules	Contigs	Loci
Splicing		>0	T			T	F			1			1040	812
Splicing DU	F	>0	T			T	F			1			171	146
PolyA					≥5	T				1		1	105	96
PolyA DU	F		T		≥5	T				1		1	17	16
lincRNA			F	F		T				1	>200	2	809	282
asRNA			F	T		T				1	>200	2	356	156
SNV DU	F		T	F		T	T	T		1		3	717	493
Intron			T	F	0	T			T	1		4	1909	576
Intron DU	F		T	F	0	T			T	1		4	559	185
Repeats						T				≥5	>50		1141	618
Unmapped						F					>50		114	