A read mapper for investigation of U-insertion/deletion RNA editing

Pavel Flegontov

Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava, Czech Republic;
Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic;
Insitute for Information Transmission Problems RAS, Moscow, Russia
pflegontov@gmail.com

Evgeny Gerasimov

Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia;
Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia;
Insitute for Information Transmission Problems RAS, Moscow, Russia
jalgard@gmail.com

Vojtěch David

Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic
mortis.gw@seznam.cz

We have developed software tools for accurate and exhaustive mapping of RNA-seq reads having extensive U-insertions/deletions, allowing a detailed investigation of U-insertion/deletion (U-indel) RNA editing typical for the mitochondria of kinetoplastid protists. Importantly, we have conducted what is to our knowledge the first investigation of U-indel-edited mitochondrial transcripts based on deep transcriptome sequencing. Since the discovery of this type of RNA editing in 1986, editing mechanisms have been unraveled via targeted sequencing on a clone-by-clone basis. Recently, deep sequencing of guide (g) RNA libraries in *Trypanosoma brucei* [1] has uncovered an unexpected degree of complexity and disorder inherent in gRNA-mediated editing. By deep sequencing of mRNAs in kinetoplastids *Perkinsela* sp. and *Leptomonas pyrrhocoris*, we show that up to 50% of reads for a given edited region contain errors of the editing system, also termed 'misediting' [2,3], although we have not detected alternative translatable mRNAs of considerable abundance [4].

Our tool package for investigation of U-indel RNA editing includes two programs: a modified version of *Bowtie2* and a dedicated read mapper, *T-aligner*.

Bowtie2 is an open-source fast and accurate short read mapper written in the C++ programming language [5]. It uses a fast multiseeding procedure to find candidate alignment locations, and then proceeds with the Smith-Waterman algorithm using SIMD (single instruction, multiple data) to create the best gapped alignment [6]. Like any other read mapper we are aware of, Bowtie2 uses a scoring system with equal gap open and extension penalties for the four nucleotides, A, G, T, and C. We modified Bowtie2 to facilitate accurate alignment of U insertion/deletion-edited RNA reads, while preserving mapping speed and accuracy. Edited reads of the mitochondrial genomes of kinetoplastids have U-indels only, therefore they can be aligned correctly when gap penalties for T (corresponding to U in RNA) are different from those for A, G, and C.

We modified the Bowtie2 v.2.0.2 source code and implemented a more complex nucleotide-specific gap scoring system that allows separate penalty values for A, G, T, and C. Source code modifications were made both in the aligner module, which fills the dynamic programming table, and in the backtrack module of the program, which reconstructs the alignment using the filled dynamic programming table. Branch and array access instructions were minimized for each step, ensuring minimal time cost for more complex scoring. Using this scoring matrix, U-indel edited reads can be successfully mapped and accurately aligned with a low T-indel penalty and high penalties for other nucleotides. Additional modifications of the alignment procedure were necessary in order to let reads have a gap/mismatch after the last nucleotide of the read. This option allows the seeding of more extensively edited reads on a pre-edited RNA sequence and prevents a significant fraction of edited reads from being discarded. Alignments made with Bowtie2 were cut into overlapping windows, and examined to find sequences appropriate for seeding further read mappings with T-aligner.

T-aligner is a new program written in the C++ programming language with the source code posted online (https://github.com/jalgard/T-Aligner). The algorithm is specially designed to map extensively edited RNA-seq reads on pre-edited transcript references, taking into account the biological mechanism of RNA editing, i.e. its 3’ to 5’ progression along the transcript. Exact matches between short substrings (seeds) are first found using a hash table, but only in a short pre-defined region near the 3’-end of the edited domain. A local optimal alignment is then produced with the Smith-Waterman algorithm, allowing ‘T,-‘ and ‘-,T’ gaps with zero penalty, thus taking into account the biological mechanism of U-indel RNA editing. The general T-aligner workflow is as follows: a fixed seed is chosen in a never-edited or universally edited 3′-terminus of the transcript (or editing domain in appropriate cases). Reads are then mapped if they satisfy the following criteria: (i) they contain the seed; (ii) at least part of the read lies 5′ to the seed; (iii) the alignment may contain any number of U-indels of any length; (iv) the alignment contains no other indels and no or few mismatches.

After the alignments are produced, T-aligner classifies all editing events (U insertion or U deletion) and clusters the reads into three groups: (i) those matching the reference sequence, (ii) those matching the putative main ‘editing pathway’ (i.e., the user-defined final edited product) and (iii) all other reads containing alternative editing events. Reads matching the main pathway are defined as follows: (i) those with no additional edited sites compared to the main pathway; (ii) reads with insertions/deletions that are shorter or equal in size to those in the main pathway; and (iii) reads in which all sites are edited in the same direction as in the main pathway (e.g., insertion in the main pathway versus insertion in a sequence read). Reads in violation of any of these conditions are placed in the ‘alternative editing’ group. Reads that are exact substrings of other reads are then merged into ‘editing intermediates’. The support value, i.e. average read count, associated with an editing intermediate can be used to determine the most abundant sequences, which is useful when examining alternative editing.

One to three iterations of read mapping with T-aligner (with the original seed shifting in the 3′ to 5′ direction) were enough to cover the whole transcript or its edited region, and then reconstruct the main editing pathway. Repeating T-aligner-assisted read mapping with prior knowledge of the main edited product allowed us to reveal and quantify alternative editing products.

1. D.Koslowsky et al. (2013) The insect-phase gRNA transcriptome in *Trypanosoma brucei, Nucleic Acids Res.*, **42:**1873–1886.

2. D.A.Maslov et al. (1994) Editing and misediting of transcripts of the kinetoplast maxicircle G5 (ND3) cryptogene in an old laboratory strain of *Leishmania tarentolae, Mol. Biochem. Parasitol.*, **68:**155–159.

3. N.R.Sturm et al. (1992) Generation of unexpected editing patterns in *Leishmania tarentolae* mitochondrial mRNAs: misediting produced by misguiding, *Cell*, **70:**469–476.

4. T.Ochsenreiter, S.L.Hajduk (2006) Alternative editing of cytochrome c oxidase III mRNA

in trypanosome mitochondria generates protein diversity, *EMBO Rep.*, **7:**1128–1133.

5. B.Langmead, S.L.Salzberg (2012) Fast gapped-read alignment with Bowtie 2, *Nat. Methods,* **9:**357–359.

6. M. Farrar (2007) Striped Smith-Waterman speeds database searches six times over other SIMD implementations, *Bioinformatics,* **23:**156–161.