Telonis et al. BMC Bioinformatics (2016) 17:123
DOI 10.1186/s12859-016-0921-0

BMC Bioinformatics

CORRESPONDENCE Open Access

Consequential considerations when mapping tRNA fragments

Aristeidis G. Telonis, Phillipe Loher, Yohei Kirino and Isidore Rigoutsos*

Abstract

We examine several of the choices that went into the design of tDRmapper, a recently reported tool for identifying transfer RNA (tRNA) fragments in deep sequencing data, evaluate them in the context of currently available knowledge, and discuss their potential impact on the output that the tool generates.

Keywords: tRNAs, tRNA-lookalikes, tRNA fragments, tRFs, i-tRFs, 5′-tRFs, 3′-tRFs, 5′-halves, 3′-halves, RepeatMasker

A recent article by Selitsky and Sethupathy [1] describes tDRmapper, a tool devised for analyzing RNA transcripts that originate from mature or precursor transfer RNAs (tRNAs). Such transcripts, commonly referred to as *tRNA fragments* or *tRFs*, have been commanding increasing attention in recent years [2–4]. In view of the repeating nature and particular sequence characteristics of both nuclear and mitochondrial tRNAs, several of tDRmapper's stated design choices [1] will result in the labeling as tRFs of transcripts that have a significantly-high probability to originate from non-tRNA portions of the genome. This is a particularly relevant consideration given that tDRmapper was proposed as an exploratory tool. In what follows, we define as "tRNA space" the collection of 610 nuclear and 22 mitochondrial genome loci that harbor tRNA genes [5].

Mapping to tRNA space alone, and not to the full genome, cannot unambiguously define tRNA-derived sequences

In [1], the authors present several points to support their decision for mapping to only the tRNA space and *not* to the entire genome (see Results and discussion; Step 2 of [1]). However, as we have discussed previously [6], mapping to the tRNA space alone will result in the unavoidable inclusion of sequenced reads that can be unrelated to tRNAs, thereby leading to the generation and reporting of false positives: this argument is supported by several key facts that pertain to the specifics of tRNA sequences. **First**, the human nuclear genome comprises hundreds of incomplete tRNA sequences with lengths significantly shorter than the average 72 nucleotides (nts) of the typical mature tRNA, which in turn makes them non-tRNAs. For example, for the hg19 assembly (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/index.shtml) the "tRNA" class of RepeatMasker (http://www.repeatmasker.org; Library 2014/01/31) includes 716 genomic segments, outside of tRNA space, each with length ≤ 50 nts. Because of the existing sequence similarity, tDRmapper will consider these deep-sequencing reads as tRFs even though their ambiguous genomic origin makes it distinctly probable that they do not originate from tRNAs [6, 7]. **Second**, many mature tRNAs are spliced from intron-containing precursor transcripts: deep-sequencing reads that span such tRNA exon-exon junctions cannot be claimed to be tRNA-specific without exploring the entirety of the human genome. An example of this complication is highlighted by the junction-spanning sequence ACTTCTAATTCAAA (all sequences are shown in 5′ → 3′ orientation): as implemented, tDRmapper will assign these reads to the five *genes* of ArgTCT despite the fact that this sequence has 99 exact copies in the human genome outside of tRNA space. **Third**, the presence of the non-templated "**CCA**" at the 3′ terminus of mature tRNAs creates novel CCA-ending strings that are candidate 3′-tRFs: however, these strings can also exist at multiple genomic locations that are unrelated to tRNAs. By not examining their instances outside of tRNA space, tDRmapper runs the risk of labeling as

* Correspondence: isidore.rigoutsos@jefferson.edu
Computational Medicine Center, Sidney Kimmel Medical College, Thomas Jefferson University, 1020 Locust Street, Philadelphia, PA 19107, USA

3′-tRFs sequenced reads that are not linked to tRNAs. For example, the sequence CTCACTGGAACCT-**CCA** occurs in a single mature tRNA anticodon (GlnCTG, negative strand of chr1: 146476760-146476831, inclusive, or TRQ-CTG10-1), yet, the sequence also exists identically at 421 locations of the human genome outside of tRNA space. This 16-mer is not an isolated example. As a matter of fact, for 480 of the 632 nuclear and mitochondrial tRNAs in tRNA space, their CCA-ending 16-mers have a grand total of 2643 identical copies in the human genome outside of tRNA space: tDRmapper lets these 16-mers support a 3′-tRF even though their ambiguous origin suggests that they may not be related to tRNAs. **Fourth**, we recently showed that there are 497 loci in the human genome whose length matches that of the typical mature tRNA and whose sequences resemble, at different levels of sequence similarity, known tRNAs [5, 8]. Any deep-sequenced reads that resemble these lookalikes are of ambiguous origin, yet will be assigned by tDRmapper to currently known *bona fide* tRNA genes, which will in turn result in miscalculating a given tRNA's potential to produce tRFs. The resulting cross-talk can potentially be very substantial: for example, 350 of the 632 nuclear and mitochondrial tRNAs share 1493 *exact* 16-mers with these 497 tRNA-lookalike sequences. For some of the tRNAs, the above four considerations will be exacerbated by tDRmapper's allowance of mismatches and deletions. We note that strategies for overcoming localization ambiguity and controlling for potential false positive hits have already been developed and published by others [7] and by us [6].

The inclusion of short sequenced reads will bias the generated profiles

In addition to the large number of long repeat elements, the human genome is riddled with numerous identical repetitions of sequences of length ≥ 16 nts [9]; shorter sequences appear identically at even more locations. As shown in Table 1, 92.4 % of all the 14-mers that exist in the *mature* nuclear and mitochondrial tRNAs (which include the CCA addition) also exist elsewhere on the genome. Therefore, their inclusion will result in the labeling as tRFs of 14-mers with high likelihood to originate outside of tRNA space. At the same time, as these non-tRNA 14-mers will be allowed to contribute "votes" to the accumulating profiles, they have the potential to also skew the endpoints of true tRFs that are present in the analyzed samples. In the case of 15-mers, there is still a very large fraction (79.0 %) of instances outside of tRNA space. Altogether dismissing 14-mers, and perhaps 15-mers, should reduce the rate of false positives without appreciable loss of signal. However, as we highlighted in the previous section and discussed at length in [6], for higher values of *k*, e.g., 16-mers (Table 1) or longer ones, disambiguation of the genomic origin needs to be carried out separately for each sequenced read before it can be labeled a tRF.

Associating a 'dominant' tRF with each tRNA sequence runs counter to the available evidence

The authors of tDRmapper purport to being able to associate a "dominant" tRF with each tRNA family, *isodecoder,* or *isoacceptor* (Results and Discussion; Step 3). This is based on their implicit assumption that the *most abundant* tRF (i.e., "dominant" per their definition) is also the *most important* one. However, as we already reported in [6], there are many examples of human tRFs that are *not* the most abundant tRFs in the analyzed samples yet can differentiate among groups of samples by gender, population origin, race, tissue, disease subtype, etc. Moreover, tDRmapper's assigning of counts from *multiple distinct* tRFs to a *single* sequence further complicates matters, because such an aggregation of reads by tDRmapper does not acknowledge the following important fact. As we showed by analyzing hundreds of human tissue samples (see [6]), distinct tRFs that arise from the same anticodon template have uncorrelated abundances even though they can differ only slightly in sequence (the abundances of such distinct fragments are collapsed by tDRmapper into a composite 'count' that will suppress any available signal and skew the endpoints of the molecules that are actually present). This lack of correlation is not specific to tRNAs: indeed, it also holds true for another category of non-coding RNAs the microRNAs (miRNAs), as others [10, 11] and we reported previously [12, 13]. Specifically, in both health and disease, we have shown that different miRNA isoforms from the *same* miRNA locus can: a) have uncorrelated abundances even though they have very similar sequences [13]; b) differ greatly in their ability to distinguish between gender, population origin, race, tissue, disease subtype, etc. [12, 13]; and, c) have very distinct regulatory roles and downstream targets [13]. The numerous dependencies that tRF sequences have been shown to have on the cell/tissue type of origin, the disease type/sub-type, and the attributes of the person from whom

Table 1 Number of 14-/15-/16-mers from the 632 nuclearly- and mitochondrially- encoded tRNA sequences (including the non-templated CCA) with instances inside and outside of tRNA space

k-mer length	# of distinct *k*-mers present in nuclear and mitochondrial tRNAs	# of *k*-mers with instances exclusively inside tRNA space	# of *k*-mers with instances inside and outside of tRNA space
14	16,380	1251 (7.6 %)	15,129 (92.4 %)
15	16,733	3519 (21.0 %)	13,214 (79.0 %)
16	17,006	6972 (41.0 %)	10,034 (59.0 %)

the RNA is sourced, are evidence that does not support the hypothesis of "dominant" tRFs.

We hope that these insights prove useful in the quest to capture the tRF repertoire of samples of interest from RNA sequencing data.

Abbreviations

tRNA: transfer RNA; tRF: tRNA fragment; 5′-tRF: 5′ tRNA fragment; i-tRF: internal tRNA fragment; 3′-tRF: 3′ tRNA fragment; miRNA: microRNA; hg19: human genome assembly 19.

Competing interests

All four authors are actively involved in tRNA research and have published previously in this area. The authors declare no competing financial interests.

Authors' contributions

AGT, IR and PL carried out the described analyses. IR, AGT, YK and PL wrote the commentary. All authors read and approved the final manuscript.

Response to "Consequential Considerations When Mapping tRNA Fragments"

Praveen Sethupathy[1,2]

[1]Department of Genetics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[2]Lineberger Comprehensive Cancer Center, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

We must first address a citation oversight. Specifically, we wish to explicitly mention here two important studies that were not cited in our original tDRmapper publication [1]. The first is Honda et al., 2015, in which the authors implicate tRNA-halves in breast and prostate cancer cell proliferation [14]. The second is Telonis et al., 2015, in which the authors develop a bioinformatic mapping scheme to establish the diversity of human tRNA-derived RNAs (tDRs) [15]. This paper was published while our tDRmapper manuscript was already under review. These two studies represent contributions to the tDR literature that merit attention.

Our primary goals with tDRmapper [1] were five-fold:

(1) To present the major challenges associated with analysis of tDRs
(2) To provide a publicly-available bioinformatic tool for annotation and quantification of tDRs from human small RNA-sequencing data
(3) To demonstrate the abundance of tDRs in several human tissues
(4) To offer new graphical visualization strategies for the analysis of tDR profiles and tDR chemical modifications
(5) To catalyze discussion among scientists in the RNA and genomics research communities about improved approaches for tDR nomenclature and quantification

In the spirit of goal #5, we welcome the commentary from Telonis et al., and we hope that it will lead to a broader conversation about the limitations of current approaches and ideas for new strategies. We provide below our responses to their two major critiques:

(1) Mapping to tRNA space instead of whole genome

Most tools in bioinformatics represent a trade-off between false negative and false positive rates. We decided in tDRmapper to reduce false negatives, while sacrificing a potential increase in false positives.

We made the decision to map to tRNA space in part because tRNA annotation in the human genome is incomplete. It remains unclear whether currently unannotated genomic regions that resemble tRNA sequences are *bona fide* tRNAs, truncated tRNAs, or unrelated to tRNAs entirely. Mapping to the whole genome would result in ambiguity about the origin of many reads. If we discard all ambiguous reads, then we run the risk of artificially lowering the true signal from annotated tRNAs. However, Telonis et al. correctly state that mapping only to annotated tRNAs in the genomic tRNA database (GtRNAdb) as opposed to the entire genome can lead to false positives, because some reads might map equally well elsewhere in the genome at incomplete-tRNA or non-tRNA loci, and these loci may be the true origin of the reads.

Consider supplemental table 6 of our recent publication in *Scientific Reports* [16], in which we list the putative tDR sequences found in human liver, as well as the coordinates of the locations to which these sequences map. The most abundant sequence, which we designate as a tRNAhalf (tRH), maps to the 5'-end of the annotated full-length Glycine tRNAs, as well as to two other locations. These other locations, which are embedded within repeat regions, harbor only incomplete Glycine tRNA sequences (it is not clear whether these regions are robustly transcribed in the liver). Therefore, the origin of the "tRH" is technically ambiguous – it could derive from the full-length Glycine tRNA or the

incomplete Glycine tRNA sequences with unknown transcription status. If these reads were excluded (as suggested by Telonis et al.), a very large proportion of the tDR signal in liver tissue would be discarded, thereby running the risk of greatly inflating the false negative rate. If the reads are included (as proposed by Selitsky et al.[2]), a very large proportion of multi-mapped reads would be counted as tDRs, thereby running the risk of greatly inflating the false positive rate. In tDRmapper, we have made the decision to reduce false negative rates.

Furthermore, and perhaps more importantly, mapping to the entire genome will miss tDRs that are modified, such as by non-templated nucleotide addition (*e.g.*, 3'-terminal CCA), and therefore do not match the genomic sequence. Mapping only to annotated tRNA space eliminates these issues and further reduces the false negative rate.

We emphasize that mapping to tRNA space alone is only a short-term solution. As human genomic annotation of tRNAs improves, and as transcription status across the human genome in different tissues is clarified (via techniques such as PRO-seq), we agree that it will be important to consider alternate strategies for mapping that balance both false positive and false negative rates.

(2) Associating a single 'dominant' tDR with each sequence

Telonis et al. are concerned about the assignment of a single, dominant tDR to each tRNA sequence. We would like to emphasize that tDRmapper only denotes a tDR as 'dominant' if the reads mapping to it represent > 66 % of the reads mapping to the parent tRNA. In such cases, the 'dominant' tDR is at least 2-fold greater in abundance than any other tDR that may be processed from the parent tRNA. Nonetheless, we appreciate the comment from Telonis et al., because we wish to avoid what Telonis et al. referred to as the "implicit assumption that the most abundant tDR is also the most important one". It behooves us to learn from the microRNA field, which now appreciates that non-canonical and less-abundant microRNA isoforms from the same precursor sequence can be functionally relevant and should be separately annotated and quantified. In the future, we will consider revising tDRmapper to annotate and quantify all tDRs that are above a certain user-specified threshold of read coverage.

We agree wholeheartedly with Telonis et al. that there are important considerations when analyzing tDRs and we welcome a robust discussion, as well as improved annotation of human tRNAs, and greater understanding of tDR function, in order to name and quantify tDRs in a more rational and accurate manner. We appreciate Telonis et al's efforts to engage the discussion and contribute their important points, and we look forward to continuing to refine tDRmapper in order to address current limitations and facilitate the discovery of novel tDR biology.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
I would like to thank Sara R. Selitsky for her assistance in reviewing and editing this response.

Received: 9 December 2015 Accepted: 29 January 2016
Published online: 10 March 2016

References
1. Selitsky SR, Sethupathy P. tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data. BMC Bioinformatics. 2015;16:354.
2. Gebetsberger J, Polacek N. Slicing tRNAs to boost functional ncRNA diversity. RNA Biol. 2013;10(12):1798–806.
3. Shigematsu M, Honda S, Kirino Y. Tranfer RNA as a source of small functional RNA. J Mol Biol Mol Imaging. 2014;1:2.
4. Sobala A, Hutvagner G. Transfer RNA-derived fragments: origins, processing, and functions. Wiley Interdiscip Rev RNA. 2011;2(6):853–62.
5. Telonis AG, Loher P, Kirino Y, Rigoutsos I. Nuclear and mitochondrial tRNA-lookalikes in the human genome. Front Genet. 2014;5:344.
6. Telonis AG, Loher P, Honda S, Jing Y, Palazzo J, Kirino Y, et al. Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies. Oncotarget. 2015;6(28):24797–822.
7. Kumar P, Anaya J, Mudunuri SB, Dutta A. Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. BMC Biol. 2014;12(1):78.
8. Telonis AG, Kirino Y, Rigoutsos I. Mitochondrial tRNA-lookalikes in nuclear chromosomes: could they be functional? RNA Biol. 2015;12(4):375–80.
9. Rigoutsos I, Huynh T, Miranda K, Tsirigos A, McHardy A, Platt D. Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. Proc Natl Acad Sci U S A. 2006;103(17):6605–10.
10. Llorens F, Banez-Coronel M, Pantano L, del Rio JA, Ferrer I, Estivill X, et al. A highly expressed miR-101 isomiR is a functional silencing small RNA. BMC Genomics. 2013;14:104.
11. Vaz C, Ahmad HM, Bharti R, Pandey P, Kumar L, Kulshreshtha R, et al. Analysis of the microRNA transcriptome and expression of different isomiRs in human peripheral blood mononuclear cells. BMC Res Notes. 2013;6:390.
12. Loher P, Londin ER, Rigoutsos I. IsomiR expression profiles in human lymphoblastoid cell lines exhibit population and gender dependencies. Oncotarget. 2014;5(18):8790–802.
13. Telonis AG, Loher P, Jing Y, Londin E, Rigoutsos I. Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity. Nucleic Acids Res. 2015;43(19):9158–75.
14. Honda S, Loher P, Shigematsu M, Palazzo JP, Suzuki R, Imoto I, Rigoutsos I, Kirino Y. Sex hormone-dependent tRNA halves enhance cell proliferation in breast and prostate cancers. Proc Natl Acad Sci USA, 2015;112:E3816–25.
15. Telonis AG, Loher P, Honda S, Jing Y, Palazzo J, Kirino Y, Rigoutsos I. Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies. Oncotarget, 2015;6:24797–822.
16. Selitsky SR, Baran-Gale J, Honda M, Yamane D, Masaki T, Fannin EE, Guerra B, Shirasaki T, Shimakami T, Kaneko S, Lanford RE, Lemon SM, Sethupathy P. Scientific Reports. 2015;5:7675.