A combined approach for successful reannotation of animal mitochondrial tRNAs based on pattern-matching and tRNA-predictor programs

Lupi R, Gissi C

Motivation

Transfer RNAs encoded by the mitochondrial genome (mtDNA) of Metazoa present strong deviations from the classical cloverleaf secondary structure, including the loss or size variation of either D- or T-domain. In addition, some taxa show “bizarre” tRNA structures: nematodes possess unconventional mt-tRNAs lacking either the T or D stem [1]; spiders (Araneae, Chelicerata) and gall midges (Cecidomiiydae, Insecta) have many “truncated” tRNAs, i.e. tRNAs lacking a well-paired aminoacyl stem, which can also lost the T-arm [2,3]; annelids belonging to family Questidae have a full set of truncated tRNAs [4] . These peculiarities hamper the annotation of mt-tRNAs in mtDNA sequences, since conventional tRNA detection programs perform poorly (as tRNAscan-SE) or lead to the detection of a significant number of false positives (as Arwen) [5,6]. Finally, mt-tRNA annotations of are affected by numerous errors in gene name, boundaries and strand definition occurring during the sequence submission to primary databases [7] . In the effort to construct a curated database of complete mtDNAs of Metazoa, we have developed a specific pipeline including both pattern-matching and tRNA-predictor programs, aimed at automatically check/rectify the annotation of both standard and “bizarre” mt-tRNAs.

Methods

The developed mt-tRNA reannotation pipeline analyses the single tRNA sequences through two different programs: PatSearch, a pattern-matching program [10]; and Arwen, a mt-tRNA secondary structure predictor [6] . Two modules, made of several home-made Python scripts, specifically parse the results of PatSearch and Arwen using several empirically-settled criteria. As for PatSearch, two main tRNAs patterns were specifically set for each mt-tRNA category. These patterns are able to detect the overall tRNA secondary structure based on the identification of only the aminoacyl (AA) and anticodon (AC) arms: the first pattern assumes perfectly annotated tRNAs with correct limits and a single 3'-discriminant base in the AA stem, while the second pattern searches for tRNAs having incorrect boundaries. In addition, patterns looking for a perfect AC arm in the correct tRNA position were also defined in order to look for “truncated” tRNAs, only in taxa where such unusual tRNA structures are expected to be present (Araneae, Cecidomiiydae and

Dipartimento di Scienze Biomolecolari e Biotecnologie, Università degli Studi di Milano, Italy

Questidae). All mt-tRNA patterns assume the presence of canonical anticodon sequences. In this pipeline, Arwen was preferred to tRNAscan-SE because it has a detection rate close to 100% for mt-tRNAs, however, given the high false positive rate, Arwen results were taken into account only for mt-tRNAs not identified by the PatSearch patterns. Arwen itself has the advantage to find tRNAs with unusual anticodons and uncommon secondary structures, moreover the program was run applying specific options and extending the original tRNA boundaries from 5 to 45 bp at both gene sides, using an incremental step of 5 or 15 bp. The extension of the original tRNA boundaries forced the program to identify mt-tRNAs having erroneous gene limits.

Results

A total dataset of 42,617 mt-tRNA sequences collected in the MitoZoa database v2.0 [8] was analyzed by our pipeline: 95.9% mt-tRNAs were identified/corrected by the PatSearch module; 3.8% were identified/corrected only by the Arwen module; 0.3% of the total tRNAs were not identified at the end of the whole pipeline and correspond mainly to erroneously annotated tRNAs. Thus, our pipeline represents a reliable tool for improving the annotation quality of metazoan mt-tRNAs both in complete and partial mtDNA sequences, since it was able to resolve (i.e. correct or validate) the annotation of >99% of the analyzed sequences, taking into account either taxon-specific and secondary-structure peculiarities of tRNA genes. In order to compare the resolving power and accuracy of the two-core modules of our pipeline, the PatSearch-confirmed tRNAs (42,617 minus the 184 “truncated” tRNAs) were re-analyzed by the Arwen module. The results were straightforward: only 0.06% mt-tRNAs were predicted with different gene name/strand by the PatSearch compared to the Arwen module (these cases will be fully discussed in the poster), while 1.5% mt-tRNAs identified by PatSearch were not found by the Arwen module. This is mainly due to the low detection rate of Arwen for tRNA-Ser(AGY), tRNA-Cys and nematode mt-tRNAs, which together correspond to 44% of the total tRNAs not identified by Arwen. Among the 184 “truncated” tRNAs, all sequences were found by the PatSearch-settled patterns, while only 64 tRNAs (34.8%) were identified by the Arwen module. These results demonstrate that adequate patterns describing the tRNA secondary structure outperforms a good tRNA predictor such as Arwen in the present mt-tRNA reannotation pipeline, and could be useful for the identification of tRNA-like structure.

Contact e-mail

renato.lupi@unimi.it

Supplementary information

References

[1] Watanabe, Y., Tsurui, H., Ueda, T., Furushima, R., Takamiya, S., Kita, K., Nishikawa, K. and Watanabe, K. (1994) J Biol Chem, 269, 22902-22906. [2]

Beckenbach, A. and Joy, J. (2010) Genome Biology and Evolution, in press, 278-287. [3] Masta, S. and Boore, J. (2008) Mol Biol Evol, 25, 949-959. [4] Bleidorn, C., Hill, N., Erséus, C. and Tiedemann, R. (2009) Mol Phylogenet Evol., 52, 57-69. [5] Lowe, T.M. and Eddy, S.R. (1997) Nucleic Acids Res., 25, 955-964. [6] Laslett, D. and Canback, B. (2008) Bioinformatics., 24, 172-175. [7] Boore, J. (2006) OMICS, 10, 119-126. [8] Lupi, R., D'Onorio De Meo, P., Picardi, E., D'Antonio, M., Paoletti, D., Castrignanò, T., Pesole, G. and Gissi, C. (2010) Mitochondrion, 10, 192-199.