Human NumtS: features and bioinformatics approaches for their location and quantification

ID - 125

Attimonelli Marcella[1], Lascaro Daniela[1], Castellana Stefano[1]
[1]Department of Biochemistry and Molecular Biology, University of Bari, Bari

Motivation

Eukaryotic Nuclear genomes present, to a greater or lesser extent, fragments of their mitochondrial genome counterpart derived from the random insertion of damaged mtDNA fragments. Fragments of mtDNA escape from mitochondria due to the presence of mutagenic agents or other forms of cellular stresses. The fragments reach the nucleus and during the repair of chromosomal breaks insert themselves into the nuclear DNA [1]. Close examination of the Nuclear mt Sequences (NumtS) loci reveals a lack of common features at integration sites. NumtS were discovered in 1967 by du Buy and Riley in the mouse liver nuclear genome. The presence of mtDNA in nuclear genome was later (1983) confirmed in several other organisms. In 1994 Lopez et al call these fragments NumtS (new mights). NumtS are not equally abundant in all species and are redundant and polymorphic in terms of number of copies For population genetists and clinicians studying mt diseases, it is important to have a complete overview of NumtS quantity and location. The reason for this is that they are a potential source of contamination when PCR is used to study mtDNA without prior purification of mtDNA and being located in the nucleus, evolve much more slowly than their functional counterparts thus, they can be used as outgroups in phylogenetic studies. Searching PubMed for NumtS or Mitochondrial pseudogenes, hundreds of papers are retrieved. Many of these papers report compilation of Human NumtS [2] mostly obtained by applying in silico approaches, while only a minority of them are derived from a wet-lab approach [3]. A comparison of the published compilations clearly shows significant discrepancies among data due to both an unwise application of Bioinformatics methods and to a not yet correctly assembled nuclear genome. Thus the data are still incomplete and imperfect. How to optimize quantification and localization of NumtS? By applying more bioinformatics approaches and verify the results through sequencing approaches. Here we report a consensus compilation of Human NumtS obtained by applying different bioinformatics approaches.

Methods

The in silico approach is based on the application of database similarity searching methods by comparing the rCRS sequence [4] with the complete Human Genome. Blast is the program used to this purpose, but there are many implementations of Blast and each implementation may produce different results depending on the parameters chosen and on the subject sequences. Thus, we have applied Blastn in different conditions, then we have applied MegaBlast and BLAT. By applying Blastn we have submitted different runs changing subject sequences (database), Limits by Entrez and Limits on Hits Number. The threshold E-value has always been fixed at 0.001. Among the runs we have selected results obtained by searching human chromosome database keeping high the Limits on Hits Number in order to be sure to avoid false negatives. We have applied Megablast on all the assemblies of the Human Genome at NCBI. Finally we have applied Blat at the UCSC site. The Human database at UCSC contains the Human golden sequences, and 4 different builds (hg15, hg16, hg17, hg18) can be searched, thus it is possible to eliminate false positives attributable to assembling artefacts.

Results

The comparison of the results obtained and the comparison of these results with published ones has allowed us to produce the RefHumanNumtS compilation reporting 188 Human NumtS. At present, the compilation is available through an Excel spreadsheet but soon it will be implemented in a database available through the HmtDB genomic resource. The following information is available for each NumtTS: the identifier, Chromosome and strand location, length, mt and nuclear coordinates, score of the alignment, Blat blocks number and similarity, orthologous NumtS in Chimpanzee, the sequenced NumtS and their polymorphic copy number, the genomic features of the NumtS locus, the isochore where the NumtS is located. The mapping with the published compilation is also annotated. Finally, for each gene and regulatory region of the human mt genome, the number of times it is duplicated along the human nuclear genome is reported. Future developments: checking for SNPs in the NumtS, analysing NumtS integration sites, developing NumtS database, applying our protocol to other species, checking for orthologous NumtS, sequencing the compiled NumtS. The latter task, however, would require a great effort and substantial funding.

References:

[1] Bensasson D.et al., J Mol Evol , 2003, 57:343354.
[2] Ricchetti M et al.,, PLoS Biol, 2004, 2(9): e273.
[3] Parr RL et al., BMC Genomics, 2006, 21:7-185.
[4]Andrews RM et al. Nat.Genet., 1999, 23:147.

Email: m.attimonelli@biologia.uniba.it