The human NumtS revised compilation, RHNumtS.2: custom tracks, polymorphisms and validation by amplification and sequencing

Simone D[1], Calabrese FM[1], Mineccia G[1], Lang M[2], Gasparre G[2], Attimonelli M[1]

Motivation

In our group in Bari we have a long tradition in the study of human mtDNA variability. We have thus developed the HmtDB database [1] (www.hmtdb.uniba.it) storing the published and unpublished human mitochondrial genomes (about 7000) annotated with adding values data concerning population samples and DNA variability. In the last years we have moved our focus on human nuclear DNA regions where traces of human mitochondrial DNA can be detected: the NumtS [2,3]. By comparing the reference human mt DNA vs the reference human nuclear DNA through database similarity searching tools (Blastn+MegaBlast+Blat), we have produced and published the RHNumtS compilation [4], reporting 190 different NumtS; to validate the in silico results, some of the NumtS reported in the compilation showing higher risk to be false postives, have been sequenced from a European sample. In order to complete the validation, a more systematic protocol to search primers and hence to amplify and sequence all the NumtS in RHNumtS has been set up. By the way, in the step of the protocol aimed to design the primers, by applying Primer-Blast software, we have detected in the flanking regions of the annotated NumtS, DNA regions that, in the application of Blastn, MegaBlast and BLAT, had not shown any similarity with the human mtDNA, while Primer BLAST sensibility has been successful in this. These results have suggested us to fully revise the RHNumtS compilation by applying Blastn with more relaxed parameters, thus allowing the detection of NumtS regions less conserved if compared to their mitochondrial counterpart although showing traces of their mt origin. Here we present the revised RHNumtS compilation (RHNumtS.2), supporting tools and variability data concerning human NumtS.

Methods

RHNumtS.2 annotates the following categories of NumtS: % NumtS already present in the first release (numeric code + A, B, C); % NumtS flanking the ones from the previous category, detected with primer designing. They are identified with the “nn” suffix; % NumtS obtained by applying BLASTn (query: J01415.2, the ref human mt genome) with more relaxed parameters (gap opening penalty = -5, gap

[1] Dipartimento di Biochimica e Biologia molecolare “Ernesto Quagliariello”, University of Bari, Italy [2] Unità di Genetica Medica, Policlinico Universitario S. Orsola-Malpighi, Bologna, Italy

extend penalty = -2, match reward = 2, mismatch penalty = -3, e-value = 1e-04). They are identified with the “r” prefix. The UCSC custom tracks function has been used to develop both the human nuclear and mitochondrial custom tracks allowing to map each RHNumtS sequence on both the human nuclear and mt genomes. The revised RHNumtS sequences, selected on the custom tracks among those not falling in repeated regions have been amplified and sequenced after the selection of specific and unique primer pairs by applying Primer-Blast.

Results

The Blastn returned 766 hits; nearly spaced hits were concatenated with criteria described by Graur [5]. Thus the RHNumtS.2 compilation reports 624 NumtS. Primer designing was performed on 265 NumtS. 223 (84%) of the selected primers have been amplified and sequenced from a European sample, thus showing the efficiency of our protocol and the good quality of the data annotated in RHNumtS. However, most of the not sequenced NumtS were amplified but, due to the presence of insertions in heterozygosis, it was difficult to produce the sequence. The produced sequences have been multialigned with the mtDNA counterpart from the same sample, with the hg18 NumtS, and the rCRS (J01415.2) corresponding fragment. The application of a Java script to each NumtS multialignment extracted variant sites due to mt/nu SNPs or nu/nu SNPs or mt/mtSNPs. The resulting data are under check in dbSNPs the nu/nu and in Phylotree and HmtDB the mt/nu SNP with the aim to contribute to NumtS datation (see F.M.Calabrese et al in this book). Finally, the browsing of NumtS custom track allows the user to characterize each NumtS to know in which genomic components the NumtS is located: whether in a genic or intergenic region, which SNPs have been mapped inside, if the NumtS is internal to a repetitive region, the synthenic region in other organisms and so on. The full extended RHNumtS custom track are available upon request and can be implemented on the UCSC genome browser as personal custom tracks.

Contact e-mail

dome.simone@gmail.com

Supplementary information

References

[1] Attimonelli, M. et al., BMC Bioinformatics, 2005, 6:S4. [2] Lopez JV et. al, JME, 1997, 39 (2) :174-190 [3] Parr RL et al., 2006, BMC Genomics, 7:185 [4] Lascaro D. et al. 2008, BMC genomics 9, no. 1: 267. [5] Hazkani-Covo, E., e D. Graur. 2007. Mol. Biol. and Evolution 24, no. 1: 13.