МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2012, том 46, № 1, с. 113–117

ГЕНОМИКА. ТРАНСКРИПТОМИКА

UDC 577.21;579.23"315

RARITY OF TagSNPs ACROSS TRANSFERRED mtDNA INSERTS IN HUMAN GENOME

© 2012 S. K. Behura*

Eck Institute for Global Health, Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556 USA

Received July 04, 2011
Accepted for publication August 10, 2011

The study deals with the single nucleotide polymorphism (SNPs, HapMap data) around the mtDNA insertions in human genome. The results obtained from this study suggest that application of tagSNP approach for large scale genotyping targeting NUMT integration sites may be difficult due to lack of informative mutations around these loci. This warrants development of new approaches to tag mtDNA inserts in genome-wide association studies.

***Keywords*: NUMTs, insertion sites, linkage disequilibrium, tagSNP, HapMap.**

Transfer of mitochondrial DNA(mtDNA) to nuclear genome is a natural phenomenon in many eukaryotes [1–5]. Although this promiscuous nature of organellar DNA has been known for long time [6–8], it is only recently that they are getting serious attention regarding their structure, evolution and significance in the nuclear DNA [9–14]. The transfer of mtDNA to nuclear genome represents a dynamic and continuous evolutionary process as evident from continued colonization of such DNAs in human chromosomes over time [15]. A number of loci (about 200) have been identified in human chromosomes based on BLASTN searches of the human genome sequence with mtDNA [16–19]. The lengths of these NUMTs are variable, some being exceptionally larger (at least five NUMTs are >10 kb) than majority of the inserts (<2 kb) in the genome [19]. It has been further suggested that about two- third of the human NUMTs are derived copies from pre-existing inserts by degradation and chromosomal rearrangements of the original inserts [20]. The current work highlights the patterns of linkage disequilibrium (LD), a process of nonrandom association of alleles of different loci, and explores identification of tagSNPs that can be useful in genome-wide association studies of NUMTs in human diseases. TagSNPs are select subset of all mutations that can be used as informative markers to tag the loci and are useful for minimizing cost and time of large scale targeted genotyping projects.

EXPERIMENTAL

A total of 54 NUMTs of varying length localized in different chromosomes (Supplementary table) were investigated in the present study. These inserts are of varying sizes (ranging from 591 bp to 18.604 bp) and are distributed in different (17) chromosomes in the human genome. These NUMTs were essentially same set of loci compiled by [19] as mtDNA inserts. The allelic associations of single nucleotide polymorphisms (SNPs) in the flanking regions (up to 100 kb) were studied for each NUMT. The flanking SNPs and their genotypes in different populations of human were obtained from HapMap data (Release 27 Phase II+III; February 2009; dbSNP b126) annotated on NCBI B36 assembly of human genome. The 54 NUMTs were selected based on quality genotype data in both the flanking regions (NUMTs were not included if more than 25% genotypes were undetermined). Based on these criteria, the ASW (African ancestry in Southwest USA) population was chosen to study the LD patterns across each NUMT.

However, a subset of NUMTs (table) were found qualify the above criteria and hence were stadied in additional populations such as CHD (Chinese in Metropolitan Denver, Colorado), GIH (Gujarati Indians in Houston, Texas), LWK (Luhya in Webuye, Kenya), MEX (Mexican ancestry in Los Angeles, California), MKK (Maasai in Kinyawa, Kenya) and TSI (Toscans in Italy). They were chosen based on the fact that the flanking SNPs of these specific NUMTs were relatively better covered for genotyping in these particular populations compared to the flanking SNPs of other NUMTs or other populations. Multiple sequence alignment and construction of Neighbor joining (NJ) phylogenetic tree of NUMT sequences were performed using Molecular Evolutionary Genetics Analysis (MEGA) Version 4 [21]. Bootstrap analysis of phylogeny was performed with 1000 replicates. The Kimura-2 parameter was used as a distance measure.

* E-mail: sbehura@nd.edu

The linkage disequilibrium among the flanking SNPs was determined separately for each NUMT by HAPLOVIEW software [22]. The 'Tagger' program included in this software was then used to identify tagSNPs by pair-wise tagging method [23]. The minimum threshold limit of correlation coefficient of linkage disequilibrium (r^2) was kept at 0.8 for selecting such SNPs.

RESULTS

The pair-wise tagging of flanking single nucleotide polymorphisms (SNPs: HapMap Project) within 100 kb on either side of each of the 54 NUMT loci (fig. 1, Supplementary table (http://www.molecbio.com/downloads/2012/1/supp_susanta.k.behura_en.pdf)) shows striking contraction in the patterns of linkage disequilibrium (LD) around these transferred loci. Using tagSNP prediction approach [23], it was found that no (or only few) such tagSNP pairs were identifiable that revealed a strong allelic association [with coefficient (r^2) of LD > 0.8] between the left and right flanking chromosomal DNAs of any NUMT. On the other hand, the tagSNP pairs were found in excess either within the 100 kb-left or within the 100 kb-right flanking region of each NUMT (fig. 1). The data in table shows that about 50% of the NUMT loci showed no LD extension (with r^2 of LD > 0.8) across the insertions. In the contrary to this, SNPs within the flanking DNAs showed strong LD among each side. Although extended LD across the mtDNA inserts were observed for the other NUMT loci (Supplementary table (http://www.molecbio.com/downloads/2012/1/supp_susanta.k.behura_en.pdf)), the SNPs showing such LD extension between the left and right flanking DNAs was much less in number than such SNPs within the left or within the right flanking DNA.

It was further found that the observation of rarity of tagSNPs across the NUMTs was not specific to the genome within the ASW population (African ancestry in Southwest USA). These SNPs were consistently rare across the NUMTs in different human populations (table). This was based on analysis of a subset of NUMTs (ID# 6, 7, 9, 11, 14 and 25 see Supplementary table for genome position) in additional five human populations (see Experimental) where a total of 299 common SNP markers were compared among the left- and right-100 kb of these NUMT loci in these specific populations (the common markers were chosen based on availability of quality genotype data in different populations for the NUMT flanking regions as well as based on the captured SNP sets by the Tagger program in predicting tagSNPs).

Furthermore, our analysis also indicates that the length of NUMT seems to be a good indication of pattern of LD across them (fig. 2). This was evident from

Consistent pattern of rarity in the number of tagSNPs across NUMTs among human populations*

NUMT	L/R	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
6	L	19	19	19	19	19	19	19	19	19	19	19
6	R	19	18	19	19	19	19	19	19	19	19	19
6	L..R	0	0	0	0	0	0	0	0	0	0	0
7	L	12	12	10	11	11	11	12	11	12	11	12
7	R	26	26	26	10	26	10	25	26	26	26	26
7	L..R	0	0	1	16	1	16	1	0	0	0	0
9	L	3	2	0	0	3	0	3	3	3	3	3
9	R	26	26	26	26	26	25	26	26	26	26	26
9	L..R	0	0	3	3	0	3	0	0	0	0	0
11	L	29	27	26	26	29	27	29	27	29	29	26
11	R	44	44	41	41	41	43	46	46	46	41	45
11	L..R	2	4	8	8	5	5	0	2	0	5	0
14	L	45	45	45	45	45	43	45	45	45	45	44
14	R	40	41	40	38	41	39	41	41	41	40	40
14	L..R	1	0	0	3	0	2	0	0	0	1	1
25	L	9	9	6	6	9	7	9	9	9	9	6
25	R	23	23	23	23	23	23	23	23	21	23	23
25	L..R	0	0	3	3	0	2	0	0	0	0	1

*A randomly selected six NUMTs (NUMT IDs # 6, 7, 9, 11, 14 and 25, see table for genome position of these NUMTs) were analyzed for tagSNPs within left (L) and right (R) 100 kb flanking sequences as well as between them (L-R; i.e. across NUMT). The headings on 1st row shows abbreviation of different populations.

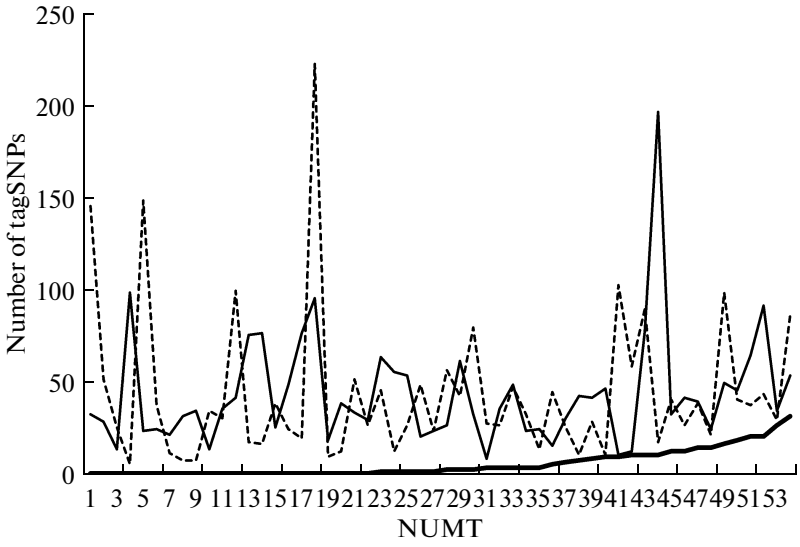

Fig. 1. Comparison of number of tagSNPs (y-axis) in the flanking sides of NUMTs (x-axis) in human genome. The dotted line represents the number of tagSNP pairs within the left 100 kb region, the solid thin line shows the number of tagSNP pairs within the right 100 kb region and the solid thick line shows the number of tagSNP across (between left and right flanking region) of each NUMT.

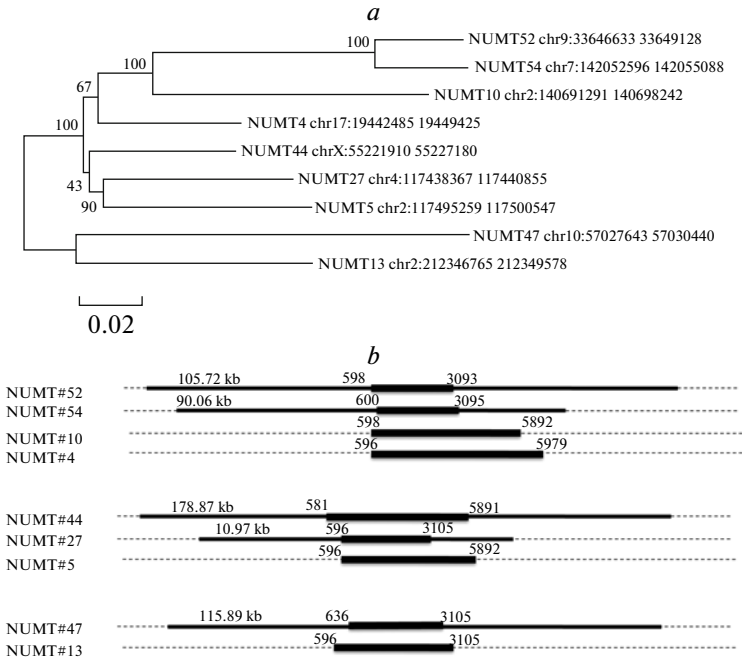

Fig. 2. Relationship between LD extension with the length of NUMTs. *a* – Neighbor-joining phylogenetic tree of NUMTs those originated from same region of mtDNA. The three grouping in the tree are illustrated in *b*. The dotted line in *b* shows the chromosomal DNA, the thicker line shows the extent of LD ($r^2 > 0.8$) and the thickest line shows the NUMT length (with I.D. on the left). The numbers on both ends of NUMTs show the positions in mtDNA to which the NUMT show sequence homologies. The numbers on the LD extension show the length (in kb) between the farthest tagSNP pairs across the mtDNA insertion. Note that the drawing is not in scale.

analyzing NUMTs those are of different lengths but show sequence homologies among each other and to an overlapping origin in the mtDNA. Among these NUMTs, extension of strong LD was observed in the case of smaller NUMTs and no such LD pattern was observed across the larger NUMTs. The negative correlation was statistically significant ($p < 0.05$) among the NUMTs that showed sequence homologies to same

region of the mtDNA. However, the relationship was not significant when unrelated NUMTS (sequence homologies to different regions of mtDNA) were considered together (data not shown). The absence of strong LD extension across the flaking regions of the larger NUMTs suggest that these inserts are associated with chromosomal sites those may be relatively more prone to recombination. Low LD and high recombination across these large inserts are necessary for disintegration to smaller NUMTs that are present in large excess not only in human genome but other eukaryotes as well. The disintegrated smaller NUMTs are associated with regions with extended LD where recombination events are less active. This indicates that strong linkage disequilibrium between the flanking regions of mtDNA insertion sites may be associated with the stability of NUMT sequences in the chromosomal DNA.

DISCUSSION

The linkage disequilibrium patterns around mtDNA insertions are investigated in the human genome. The strong LD structure around NUMT insertion sites may be common in humans as revealed from genotype data of flanking SNPs in different ethnic populations. It further indicates that such LD structure may be ancestral in human genome and population structuring has little or no effect in breaking these patterns. Recombination events across these loci are responsible for breaking up ancestral LD that may eventually restore equilibrium between flanking chromosomal regions. The presence of extended LDs within the left and right flanking DNAs suggests that these flanking regions may be less prone to the recombination events. This allows the inserted mtDNA to sustain for a long evolutionary time in nuclear genome. This pattern of LD structures most likely contributes to the extent of 'fossilization' of mtDNA inserts in the human chromosomes. The result of this investigation signifies the role of linkage disequilibrium in the evolutionary process of parasitization of human chromosomes by such transferred organellar genetic materials. Apart from that, we think the important message from this study should be in relevance to genetically tagging of mtDNA inserts using HapMap data. The analysis clearly shows that HapMap genotyping data is not adequate or appropriate to generate tagSNPs to carry out large scale genotyping experiments of these loci. Even unreliability of HapMap data, particularly at the NUMT loci, can't be ruled out [24]. This study [24], by re-sequencing a mtDNA insertion site in chromosome 1, showed that the HapMap data for that region was probably misassembled. It was suggested that the SNP data was actually representation of the mtDNA rather than the nuclear DNA. Nevertheless, in the current study, we deal with SNPs in the flanking regions that are non-mtDNA sequences. However, the current study as well as study presented in [24] highlights the problems of using HapMap data particularly in connection to genotyping of NUMT insertion sites by tagSNP approach. While studies have been pointing association of NUMTs in some human diseases [25, 26], efforts are necessary to develop sizable number of tagSNPs in order to carry out affordable and reliable genome-wide association studies targeted towards these loci.

ACKNOWLEDGEMENTS

This work would have not been possible without free availability of HapMap data. I am also thankful to David W. Severson and other members of Eck Institute for Global Health for help, support and encouragement.

REFERENCES

1. Tsuzuki T., Nomiyama H., Setoyama C., Maeda S., Shimada K. 1983. Presence of mitochondrial-DNA-like sequences in the human nuclear DNA. *Gene.* **25**, 223–229.
2. Perna N.T., Kocher T.D. 1996. Mitochondrial DNA: molecular fossils in the nucleus. *Curr. Biol.* **6**, 128–129.
3. Blanchard J.L., Schmidt G.W. 1996. Mitochondrial DNA migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. *Mol. Biol. Evol.* **13**, 537–548.
4. Bensasson D., Zhang D.X., Hartl D., Hewitt G. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol. Evol.* **16**, 314–321.
5. Behura S.K. 2007. Analysis of nuclear copies of mitochondrial sequences in honeybee (*Apis mellifera*) genome. *Mol. Biol. Evol.* **24**, 1492–1505.
6. Dubuy H.G., Riley F.L. 1967. Hybridization between the nuclear and kinetoplast DNA's of *Leishmania enrietii* and between nuclear and mitochondrial DNA's of mouse liver. *Proc. Natl. Acad. Sci. USA.* **57**, 790–797.
7. Ellis J. 1982. Promiscuous DNA chloroplast genes inside plant mitochondria. *Nature.* **299**, 678–679.
8. Lewin R. 1983. Promiscuous DNA leaps all barriers. *Science.* **219**, 478–479.
9. Richly E., Leister D. 2004. NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* **21**, 1081.–1084.
10. Leister D. 2005. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends Genet.* **21**, 655–663.
11. Hazkani-Covo E., Covo S. 2008. Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS Genet.* **4**, e1000237.
12. Hazkani-Covo E. 2009. Mitochondrial insertions into primate nuclear genomes suggest the use of numts as a tool for phylogeny. *Mol. Biol. Evol.* **26**, 2175–2179.
13. Hazkani-Covo E., Zeller R.M., Martin W. 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet.* **6**, e1000834.
14. Viljakainen L., Oliveira D.C.S.G., Werren J.H., Behura S.K. 2010. Transfers of mitochondrial DNA to the nuclear genome in the wasp *Nasonia vitripennis*. *Insect Mol. Biol.* **19**, 27–35.

15. Ricchetti M., Tekaia F., Dujon B. 2004. Continued colonization of the human genome by mitochondrial DNA. *PLoS Biol*. **2**, E273.
16. Mourier T., Hansen A.J., Willerslev E., Arctander P. 2001. The human genome project reveals a continuous transfer of large mitochondrial fragments to the nucleus. *Mol. Biol. Evol*. **18**, 1833–1837.
17. Tourmen Y., Baris O., Dessen P., Jacques C., Malthièry Y., Reynier P. 2002. Structure and chromosomal distribution of human mitochondrial pseudogenes. *Genomics*. **80**, 71–77.
18. Mishmar D., Ruiz-Pesini E., Brandon M., Wallace D.C. 2004. Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. *Hum. Mutat*. **23**, 125–133.
19. Lascaro D., Castellana S., Gasparre G., Romeo G., Saccone C., Attimonelli M. 2008. The RHNumtS compilation: features and bioinformatics approaches to locate and quantify Human NumtS. *BMC Genomics*. **9**, 267.
20. Hazkani-Covo E., Sorek R., Graur D. 2003. Evolutionary dynamics of large numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications. *J. Mol. Evol*. **56**, 169–174.
21. Tamura K., Dudley J., Nei M., Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol*. 2007. **24**, 1596–1599.
22. Barrett J.C., Fry B., Maller J., Daly M.J. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. **21**, 263–265.
23. de Bakker P.I., Yelensky R., Pe'er I., Gabriel S.B., Daly M.J., Altshuler D. 2005. Efficiency and power in genetic association studies. *Nat. Genet*. **37**, 1217–1223.
24. Biswas N.K., Dey B., Majumder P.P. 2007. Using HapMap data: a cautionary note. *Eur. J. Hum. Genet*. **15**, 246–249.
25. Turner C., Killoran C., Thomas N.S., Rosenberg M., Chuzhanova N.A., Johnston J., Kemel Y., Cooper D.N., Biesecker L.G. 2003. Human genetic disease caused by *de novo* mitochondrial-nuclear DNA transfer. *Hum. Genet*. **112**, 303–309.
26. Goldin E., Stahl S., Cooney A.M., Kaneski C.R., Gupta S., Brady R.O., Ellis J.R., Schiffmann R. 2004. Transfer of a mitochondrial DNA fragment to MCOLN1 causes an inherited case of mucolipidosis IV. *Hum. Mutat*. **24**, 460–465.