European Journal of Human Genetics (2007) 15, 246–249

www.nature.com/ejhg

SHORT REPORT

Using HapMap data: a cautionary note

Nidhan K Biswas[1], Badal Dey[1] and Partha P Majumder*,[1]

[1]*Human Genetics Unit, Indian Statistical Institute, Kolkata, West Bengal, India*

The HapMap data are being widely used in human genetic studies. We show by direct resequencing of a ~6-kb region of chromosome 1 that the HapMap data are unreliable for this region. This region contains a recent mitochondrial (mt) DNA insertion. The HapMap data report the corresponding mtDNA variation and not the nuclear DNA variation. In view of mtDNA insertions of varying lengths throughout the human genome and considerable segmental duplications, it is necessary to use the HapMap data cautiously.
European Journal of Human Genetics (2007) **15,** 246–249. doi:10.1038/sj.ejhg.5201743; published online 29 November 2006

Keywords: HapMap; mitochondrial DNA; autosomal insertion; segmental duplication

Introduction

The multicountry effort to identify and catalog genetic similarities and differences in human beings has resulted in the creation of the public-domain genome database HapMap.[1] This database comprises data on human genomic polymorphisms, primarily single nucleotide polymorphisms (SNPs), on populations of African (Yoruba of Ibadan, Nigeria; $n=30$ trios), Asian (Japanese of Tokyo; $n=45$ unrelated individuals, and Han Chinese from Beijing area; $n=45$ unrelated individuals) and European ($n=30$ CEPH trios) ancestry. Another public-domain database, called dbSNP,[2] also catalogs SNP variation in the human genome based primarily on submissions by individual researchers. These public-domain databases have become valuable resources for designing human genetic studies, and for comparison of data and results generated from projects undertaken in different parts of the world. In the course of a population genetics project – undertaken to study genomic diversity among Indian ethnic groups – that involves resequencing of large regions of human chromosomes, we were astounded to find our data, pertaining to 45 phenotypically normal individuals (comprising approximately equal numbers of adult males and females) drawn from diverse ethnic groups of India, inhabiting various geographical zones, to be completely discordant with those provided in the HapMap database for a ~6 kb region of chromosome 1. We did not find such discrepancies in many other regions of chromosome 1 or other chromosomes (data not shown). We carried out investigations to identify the causes of the discrepancies found for chromosome 1. Our results show that the HapMap data need to be used with some caution, especially because of the presence of insertions of sequences in the nuclear genome from the mitochondrial genome[3] and segmental duplications[4] within the nuclear genome.

In the short arm of human chromosome 1 (nucleotide positions 604327–610167), there is an insertion[5] of 5841 nucleotides (~6 kb) from the human mitochondrial DNA (mtDNA). There are a large number of such insertions in the human nuclear DNA, termed as NUMTs.[6] These insertion events have been estimated to have taken place at different times.[7] The 5.8 kb NUMT in the nuclear DNA has 98.54% nucleotide identity with the corresponding segment of the mtDNA. Highly homologous partial copies of this 5.8 kb NUMT were also located in other parts of the nuclear genome. In the nuclear DNA, this NUMT is flanked on both sides by mammalian interspersed elements. We hypothesized that the data presented in the HapMap database for this region pertain to the mtDNA, not nuclear DNA.

***Correspondence: Professor PP Majumder, Human Genetics Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700108, West Bengal, India. Tel: +91 33 25753209 (Office), +91 33 25753212 (Lab); Fax: +91 33 25773049. E-mail: ppm@isical.ac.in**
Received 13 July 2006; revised 16 October 2006; accepted 17 October 2006; published online 29 November 2006

Table 1 Characteristics and allele frequencies of the 18 SNPs on the chromosome 1 region identified in the present study

SNP (rs # in dbSNP Build 126) identified in the present study	*Alleles (major/minor)*	*mtDNA allele at the corresponding position*	*Minor allele frequency*						
			All geographical regions combined ($n = 45$)[a]	*Geographical region*					
				North ($n = 9$)	*East ($n = 9$)*	*South ($n = 9$)*	*West ($n = 9$)*	*Central ($n = 9$)*	*Flanking sequence (SNP position is capitalized)*
SNP1 (rs7416152)	C/T	T	0.089	0.056	0.111	0	0.167	0.111	aataaCagcag
SNP2	A/G	A	0.033	0	0	0.167	0	0	accgcAttcct
SNP3	C/T	C	0.078	0.056	0.111	0	0.111	0.111	tactcCctata
SNP4 (rs9283151)	C/T	T	0.044	0.056	0.056	0	0.111	0	gaacaCacaaa
SNP5 (rs6421780)	G/A	A	0.044	0.056	0.056	0	0.111	0	tttatGctaat
SNP6	T/C	C	0.022	0	0.111	0	0	0	gaggcTtaacc
SNP7	G/A	G	0.011	0	0.056	0	0	0	tccccGcataa
SNP8	T/C	T	0.011	0	0	0	0.056	0	cacccTggagc
SNP9	G/T	G	0.056	0.056	0.056	0.167	0	0	ttttcGtctga
SNP10	C/A	C	0.022	0.056	0	0	0.056	0	ctgagCcctag
SNP11	G/A	G	0.033	0	0.056	0	0.056	0.056	gccccGacgtt
SNP12	A/G	A	0.011	0	0	0.056	0	0	ggactAtcccg
SNP13	C/T	C	0.044	0.056	0.056	0	0.111	0	gatcaCgccct
SNP14	CA insertion (−/+)	CA	0.144	0.167	0.167	0	0.222	0.167	Aac(+CA) cagtt
SNP15	C/T	C	0.011	0	0	0.056	0	0	gggccCgtatt
SNP16	C/T	C	0.044	0.056	0.056	0	0.111	0	gctttCgctct
SNP17	G/A	G	0.011	0	0	0.056	0	0	tggccGtacgc
SNP18	G/A	G	0.044	0.056	0.056	0	0.111	0	ggcgcGatgta

[a]n denotes number of individuals.

Methods

To test this hypothesis, we carried out experiments using a strategy (Supplementary Information) of amplifying only the NUMT and not the corresponding homologous segment of mtDNA. It may be noted that standard experimental approaches of DNA amplification using overlapping PCR primers designed to amplify 0.5–1 kb of DNA sequence preferentially amplify homologous segments present in mtDNA because of the presence of large number of copies of mtDNA in a human cell. We used the ABI 3100 DNA sequencer to carry out bidirectional resequencing of overlapping PCR fragments. Variant sites were detected using the Polyphred, Ver. 4.0.[8] Human nuclear reference sequence was taken from NCBI Build 35[9] and mtDNA reference sequence from the Mitomap[10] database.

Results and discussion

We detected 18 polymorphic or variant loci in this 5.8 kb nuclear region, of which 17 are single nucleotide changes, and one is a dinucleotide insertion (Table 1). The insertion was detected by visual inspection of bidirectional DNA sequences and standard approaches of offsetting and overlaying chromatograms obtained from resequencing from opposite directions. Only three of these loci are reported in dbSNP (Build 126). One possible reason for this maybe that the 15 SNPs unreported in dbSNP are specific to Indian ethnic groups. It is also possible that this is a reflection of incomplete submissions of data to dbSNP. For this genomic region, 120 SNPs are listed in dbSNP, of which 29 were selected and genotyped in the International HapMap Project.[11] We note that 71 of these 120 SNPs are reported in Mitomap[10] or mtDB[12] databases as mtDNA variants in the corresponding homologous positions. Of the 18 variant sites detected by us, none coincides to the 29 sites chosen in the HapMap study. (We note that there is variation in allele frequencies at these sites across geographical regions within India. In fact, some of the sites are monomorphic in one or more geographical regions. The sample sizes within geographical regions are too small to draw valid inferences regarding the statistical significance of variation in allele frequencies across geographical regions, which in any case was not the focus of this study.) The relevant HapMap data are given in Table 2. These data show that (a) if a site in the HapMap database is monomorphic, then irrespective of the nucleotide in the Reference Sequence, the nucleotide reported in HapMap is the nucleotide (major allele) that is present at the corresponding position in the mtDNA (as assessed from the Mitomap database,[10] which is a compendium of polymorphisms and mutations of the human mtDNA and mtDNA resequencing of all the 45 individuals included in this study) and (b) if the site in the HapMap database is polymorphic, then the major and minor alleles at this site are exactly the same as those in mtDNA, irrespective of the nucleotide in the Reference Sequence. These features clearly support our hypothesis that for this ~6 kb region, the data reported in HapMap pertain to mtDNA, not nuclear DNA.

Table 2 Characteristics of loci investigated in the International HapMap Project for a 5.8 kb region on chromosome 1

		Major/minor allele[a]		
Sl. No.	*rs #*	*HapMap Release 21 (nuclear DNA)*	*mtDNA*	*Nucleotide in Genbank reference sequence (NCBI Build 35)*
1	rs6650104	A	A	A
2	rs11240781	G	G	G
3	rs6594028	G/A	G/A	A
4	rs10458597	C/T	C/T	C
5	rs9701661	A	A	A
6	rs9701055	C/T	C/T	C
7	rs7349153	C	C	T
8	rs7417504	C	C/T	T
9	rs7340022	C	C	C
10	rs2185539	C	C	C
11	rs2185540	C	C	T
12	rs9651229	C	C	C
13	rs9699555	A	A	A
14	rs4098611	C	C	T
15	rs10159005	C	C	C
16	rs9699599	A/G	A/G	A
17	rs4098613	C	C	T
18	rs8179287	A	A	A
19	rs3905037	C	C	T
20	rs11497407	G/A	G	G
21	rs9326624	C	C	T
22	rs11510104	C	C	C
23	rs9645429	G	G	G
24	rs6594035	C	C	T
25	rs2096045	A/G	A/G	G
26	rs9700408	G	G	G
27	rs2096047	A	A	G
28	rs9326626	A	A	G
29	rs10449773	T	T	T

[a]Loci with only one allele are monomorphic.

NUMTs are dispersed on all human chromosomes. It has been reported[6] that ~200 kb of the human nuclear genomic sequence shows significant levels of similarity to the human mtDNA. Further, segmental duplications cover 5.3% of the human genome.[13] The duplicated regions have diverged to different degrees. Standard methods of PCR amplification and DNA resequencing can lead to detection of a SNP in one of these regions when in fact the SNP belongs to the homologous duplicated segment of this region. For these inserted (NUMTs) or duplicated regions, our results indicate the need for exercising caution while using the HapMap data.

References

1 http://www.hapmap.org.
2 http://www.ncbi.nlm.nih.gov/SNP/.
3 Ricchetti M, Tekaia F, Dujon B: Continued colonization of the human genome by mitochondrial DNA. *PloS Biol* 2004; **2**: 1313–1324.
4 Fredman D, White SJ, Potter S *et al*: Complex SNP-related sequence variation in segmental genome duplications. *Nat Genet* 2004; **36**: 861–866.
5 Herrnstadt C, Clevenger W, Ghosh SS *et al*: A novel mitochondrial DNA-like sequence in the human nuclear genome. *Genomics* 1999; **60**: 67–77.
6 Hazkani-Covo E, Sorek R, Graur D: Evolutionary dynamics of large *Numts* in the human genome: Rarity of independent insertions and abundance of post-insertion duplications. *J Mol Evol* 2003; **56**: 169–174.
7 Bensasson D, Feldman MW, Petrov DA: Rates of DNA duplication and mitochondrial DNA insertion in the human genome. *J Mol Evol* 2003; **57**: 343–354.
8 Nickerson DA, Tobe VO, Taylor SL: PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res* 1997; **25**: 2745–2751.
9 http://www.ncbi.nlm.nih.gov/entrez/.
10 http://www.mitomap.org.
11 The International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
12 http://www.genpat.uu.se/mtDB/.
13 International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. *Nature* 2004; **431**: 931–945.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)