LETTER

Reply to Just et al.: Mitochondrial DNA heteroplasmy could be reliably detected with massively parallel sequencing technologies

In their comment on our report (1), Just et al. suggest that sample contamination explained mitochondrial DNA (mtDNA) heteroplasmy identified in some individuals (2). The authors further question the validity of our conclusions and the reliability of using massively parallel sequencing (MPS) to detect low-frequency heteroplasmy. We systematically evaluated the presence and impact of contamination and found that it only affects a small fraction of all individuals, leaving our original conclusions unchanged.

First, if contamination is common, the number of heteroplasmy per individual should approximate the number of mtDNA differences between randomly chosen individuals. However, the average number of pairwise mtDNA differences in the 1000 Genomes Project is 37, with a range of 20~51 for within-population comparisons (Fig. 1*A*). These results are much higher than the observed heteroplasmy number. Additionally, an elevated heteroplasmy number was not observed in Africans, who have the highest interindividual mtDNA differences (figure S11 in ref. 1).

Second, for each individual, we constructed two consensus sequences, covering the major and minor alleles at heteroplasmic sites, respectively, and we defined the haplogroup for both sequences based on PhyloTree (3). Although mutations creating a new haplogroup or erasing original haplogroup-defining alleles could occur, to be conservative we considered all individuals with a secondary haplogroup as being possibly contaminated. Overall, only 5.8% individuals are impacted (Table 1).

Third, the remaining 1,022 individuals could be argued to be contaminated by a same-haplogroup sample, with private mutations contributing to heteroplasmy. However, among all within-population pairwise comparisons, only 2.2% have the same haplogroup. Other sources of contamination are unlikely to have a higher chance of having the same haplogroup as the sequenced sample than a random within-population individual in the 1000 Genomes Project. The chance of having contamination multiplying by that of being in the same haplogroup yields minuscule probability.

Fourth, some of the 63 possibly contaminated individuals still carry real heteroplasmy. Because contaminated heteroplasmic sites should exhibit similar minor allele frequency (MAF), approximating the contamination fraction (for example, Fig. 1*B*), real heteroplasmy could be detected by its clear deviation (for example, Fig. 1*C*). Furthermore, enrichment of heteroplasmy on haplogroup-defining sites is not necessarily an indication of contamination because haplogroup-defining sites have much higher mutation rates than others (Fig. 1*D*).

Our original conclusions remain unchanged even after excluding all 63 possibly contaminated individuals, mainly because 910 of 1,022 individuals carry heteroplasmy, yielding a prevalence estimate of 89.04% (originally reported as 89.68%). We still observed significant correlation between the relative mutation rate and heteroplasmy rate, even after applying an MAF cut-off of 15% ($R^2 = 0.3$, $P < 2.2e\text{-}16$). We do not understand why Just et al. (2) restricted their analysis to coding regions. We recognize the concerns of false positives in detecting heteroplasmy. However, 22 heteroplasmies identified in nine individuals with Illumina data were all observed with LS454 data at comparable frequencies (figure S2 in ref. 1), including those with very low MAF. We believe MPS is suitable for heteroplasmy study, given careful quality control to fend off errors from experiment and analysis, as also demonstrated in other studies (4–6).

Kaixiong Ye[a,1], Jian Lu[a,b], Fei Ma[c], Alon Keinan[d], and Zhenglong Gu[a,1]

[a]Division of Nutritional Sciences and [d]Department of Biostatistics and Computational Biology, Cornell University, Ithaca, NY 14853; [b]State Key Laboratory of Protein and Plant Gene Research, Center for Bioinformatics, College of Life Sciences, Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China; and [c]Department of Medical Oncology, Cancer Hospital, Chinese Academy of Medical Sciences, Beijing 100021, China

1 Ye K, Lu J, Ma F, Keinan A, Gu Z (2014) Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proc Natl Acad Sci USA* 111(29):10654–10659.
2 Just RS, Irwin JA, Parson W (2014) Questioning the prevalence and reliability of human mitochondrial DNA heteroplasmy from massively parallel sequencing data. *Proc Natl Acad Sci USA* 111:E4546–E4547.
3 van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* 30(2):E386–E394.
4 Goto H, et al. (2011) Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. *Genome Biol* 12(6):R59.
5 Payne BA, et al. (2013) Universal heteroplasmy of human mitochondrial DNA. *Hum Mol Genet* 22(2):384–390.
6 Gardner K, Payne BA, Horvath R, Chinnery PF (2014) Use of stereotypical mutational motifs to define resolution limits for the ultra-deep resequencing of mitochondrial DNA. *Eur J Hum Genet*, 10.1038/ejhg.2014.96.

Author contributions: K.Y., J.L., A.K., and Z.G. designed research; K.Y. and Z.G. performed research; K.Y. and Z.G. analyzed data; and K.Y., J.L., F.M., A.K., and Z.G. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. Email: ky279@cornell.edu or zg27@cornell.edu.

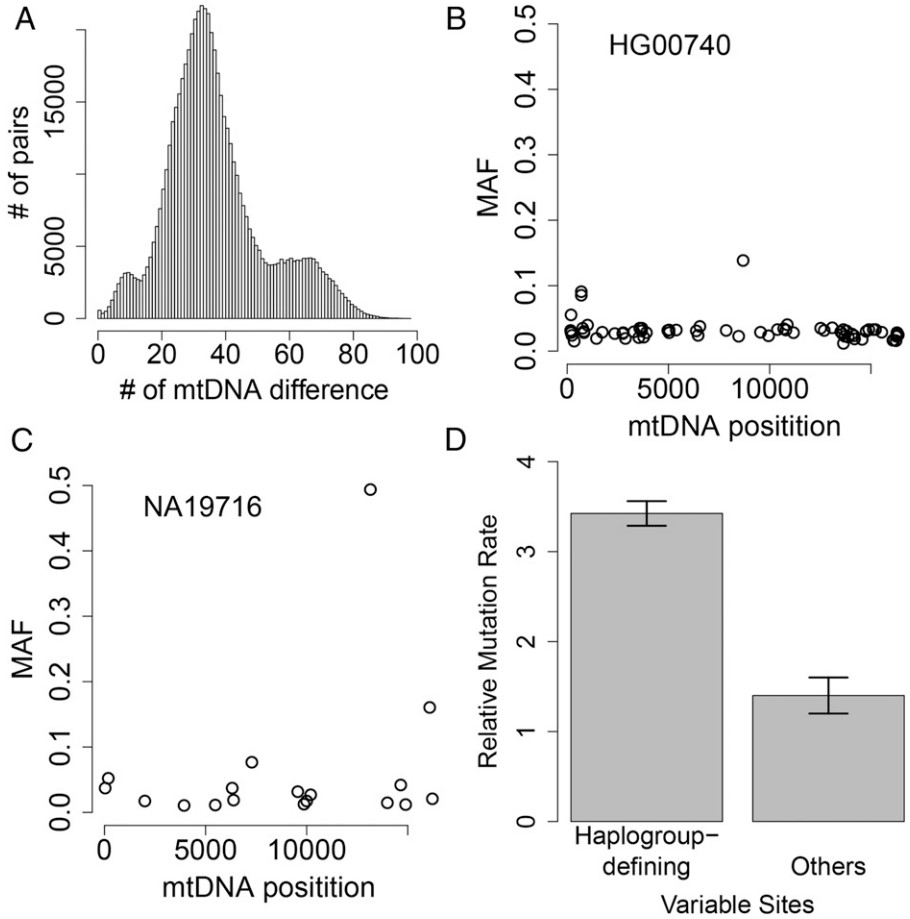

Fig. 1. (*A*) The pairwise mtDNA differences among all 1,085 individuals. (*B*) The MAF for heteroplasmic sites identified in individual HG00740. Most heteroplasmic sites exhibit a similar MAF, which should approximate the contamination fraction. (*C*) The MAF for heteroplasmic sites identified in individual NA19716. Clear deviation of MAF indicates real heteroplasmy. (*D*) The mean relative mutation rates for the haplogroup-defining and other variable sites. Error bar represents one SE. Two groups show significant differences ($P < 2.2e\text{-}16$, Wilcoxon rank sum test).

Table 1. Individuals with a second mtDNA haplogroup

ID	Major	Minor	No. of heteroplasmic sites
HG01104	A2k1a	A2k1	1
HG01171	A2ah	A2	1
HG00146	U5a2a1	U5a2a	2
NA18933	L1b1a15	L1b1a15a	2
NA18973	N9a2a1	N9a2	2
NA19657	C1b12	C1	3
HG00553	A2	A2k	4
HG01518	J1c3a2	J1c3	4
NA19351	L2a1a2	L2a1a	4
HG00120	J1c5	J1c	5
HG01259	A2q	A2+(64)	5
NA11920	H1a1a1	H1	5
NA18630	D5a2a1	D5a2a	5
NA19707	L3e2a	L3e2	5
NA20774	U5a1b1	U5a1b1b	5
HG00124	I2c	I2	6
HG00179	H10e	H10	6
HG00188	U5b1b1a	U5b1b1+@16192	6
HG00318	H3b+16129	H3b	6
NA18545	F1b1b	F1b1	6
NA19654	B2	B2+16278	6
HG00119	T1a1	T1a1+@152	7
HG00264	H5s	H	7
NA12144	H1e1a	H1e	7
NA12155	H8c	H8	7
NA18856	L2a1b3	L2a1b+143	7
NA19309	L3e2b	L3e	7
NA18538	M8a2	M	8
NA19084	Z3	M8	8
NA19160	L3e2b2	L3e	8
NA19678	A2+(64)	A2+(64)+16129	8
NA20754	H14b+152	H2+152	8
HG00671	D4i	D4+195	9
NA19774	A2	A2+(64)+@153	9
HG00245	H3g1b	H54	10
HG00536	R11b1b	R	10
HG01489	H6a1a7	H	10
NA19020	L3e2a	L3	10
NA20814	HV9c	H3an	10
HG00258	J1c2q	J1c2	11
HG00262	T2b13	T	11
NA19704	L3e2a1a	L3e2a	13
HG00155	U4b1b2	U4a2a	14
NA12044	J1c3c	R	17
NA19085	D4b1b1a1	D4b1b	18
NA19716	D1d2	D1d	18
NA07048	H2a1	J1c2c	19
NA19452	L2a1a2	L2a1a	19
HG00367	U5a1b1	H5a1e	20
NA18960	N9a4a	H1c	20
NA07000	T2b	R	23
NA12282	J1b1a1a	H+73	27
NA18990	G1a1a2	N	29
HG00377	J1c2n	H1a8	30
NA19012	A3a	D4b2b	31
HG00244	X2b5	T	33
NA18574	M7c1b2b	B2a2	42
NA18941	N9b1a	D4b1b1a1	43
NA18539	K3	D4	45
HG01626	T2a1b1a1b	U1a1a	48
NA19703	L0a1b1a	L3e'i'k'x	53
HG01108	L0a1a2	M7c1b	69
HG00740	L1b1a7a	B2b3a	71