12640–12649 *Nucleic Acids Research, 2014, Vol. 42, No. 20* *Published online 27 October 2014*
doi: 10.1093/nar/gku1038

The genomic landscape of polymorphic human nuclear mitochondrial insertions

Gargi Dayama[1], Sarah B. Emery[2], Jeffrey M. Kidd[1,2] and Ryan E. Mills[1,2,*]

[1]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA and
[2]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA

Received September 05, 2014; Revised October 10, 2014; Accepted October 10, 2014

ABSTRACT

The transfer of mitochondrial genetic material into the nuclear genomes of eukaryotes is a well-established phenomenon that has been previously limited to the study of static reference genomes. The recent advancement of high throughput sequencing has enabled an expanded exploration into the diversity of polymorphic nuclear mitochondrial insertions (NumtS) within human populations. We have developed an approach to discover and genotype novel Numt insertions using whole genome, paired-end sequencing data. We have applied this method to a thousand individuals in 20 populations from the 1000 Genomes Project and other datasets and identified 141 new sites of Numt insertions, extending our current knowledge of existing NumtS by almost 20%. We find that recent Numt insertions are derived from throughout the mitochondrial genome, including the D-loop, and have integration biases that differ in some respects from previous studies on older, fixed NumtS in the reference genome. We determined the complete inserted sequence for a subset of these events and have identified a number of nearly full-length mitochondrial genome insertions into nuclear chromosomes. We further define their age and origin of insertion and present an analysis of their potential impact to ongoing studies of mitochondrial heteroplasmy and disease.

INTRODUCTION

The presence of mitochondrial DNA in the nuclear genomes of eukaryotes has been well established, and recent reports have shown that this transfer of genetic material is an ongoing evolutionary process (1–5). In humans, these nuclear insertions of mitochondrial origin (NumtS) have been estimated to occur at a rate of $\sim 5 \times 10^{-6}$ per germ cell per generation (6) and have been implicated directly in a number of genetic disorders (7–11) while also indirectly hindering studies of mitochondrial diseases (12). A total of 755 NumtS have been identified in version hg19 of the human reference genome (13), although some portion of these have likely arisen through the duplication of previously inserted NumtS. These fragments range in size from 39 bp to almost the entire mitochondrial sequence and are thought to integrate themselves through a non-homologous end joining mechanism during double-strand break repair based on studies in yeast (14,15). Over evolutionary time, many have been highly modified due to inversions, deletions, duplications and displaced sequences, but some remain very well conserved relative to their parent mitochondria genome. While these fragments appear to be randomly selected from different regions of the mitochondria, an under-representation of the D-loop region has been reported, though why this is observed is currently unknown (16).

Several studies have previously looked at the enrichment of Numt insertions found in the human reference genome assembly relative to different genomic features. Some reports have suggested that Numt insertions tend to co-localize with repetitive elements (16,17), while others have found them to be under-represented (18). Some groups have further shown an under-representation of repetitive elements nearby NumtS in humans but not flanking Numt insertions found in chimpanzees (19). In addition, there is evidence that numts preferably insert into open chromatin regions, typically near A + T oligomer sequences (16). As these studies are primarily based on older, fixed insertions in the human lineage, it is possible that they may have been confounded by evolutionary mutational processes that have occurred since the time of these insertions. As such, an investigation into more recent insertions is warranted in order to determine any insertion biases that may lead to a greater understanding of how this transfer of genetic material occurs.

Another important aspect of NumtS is their potential effect on studies of mitochondrial heteroplasmy, which are cell or tissue level differences in individual mitochondrial genomes due in part to the high rate of mutation within these sequences (20). Low levels of heteroplasmy are typical in healthy individuals, and recent reports have deter-

*To whom correspondence should be addressed. Tel: +1 734 647 9628; Fax: +1 734 615 6553; Email: remills@umich.edu

mined that each person carries between 1 to 14 heteroplasmies (4,16,21–26). However, higher levels of heteroplasmy have been implicated in aging and various diseases such as Leber's hereditary optic neuropathy, diabetes, deafness and even cancer (27–30). The presence of NumtS can confound the study and diagnosis of these diseases through the mistaken identification of nuclear-specific Numt mutations as heteroplasmy (12,31). Computational and molecular approaches have been developed to help reduce the effect of NumtS on these studies (25,26,32–34), but they only make use of known NumtS already present in the reference sequence and do not take into account recent insertions which may be still prevalent in a significant portion of the population.

In contrast to the many studies that have utilized NumtS present in the reference sequences of many species, there has been comparably little exploration into the landscape of polymorphic NumtS in humans (35,36) and the largest such investigation to date has identified only 14 segregating events through investigation of the 1000 Genomes Project INDEL catalog (37). While rigorous, this analysis was limited in its ability to find novel insertion polymorphisms not present in the human genome reference due to the size of the sequence reads in which the variants could be discovered, resulting in the identification of only four such events. Here, we describe a new method, *dinumt*, for identifying numt insertions in whole genomes sequenced using paired-end sequencing technology, thus allowing for a greater sensitivity in identifying Numt variants of all sizes. We applied this method to 999 individuals from the 1000 Genomes (38) and Human Genome Diversity Project (HGDP) (21,39) projects and conducted an updated enrichment analysis in humans using these polymorphic insertions. We further sequenced a subset of the polymorphic NumtS we discovered and examined them for their age, origin and sequence characteristics, and assessed their potential impact on ongoing studies of mitochondrial heteroplasmy.

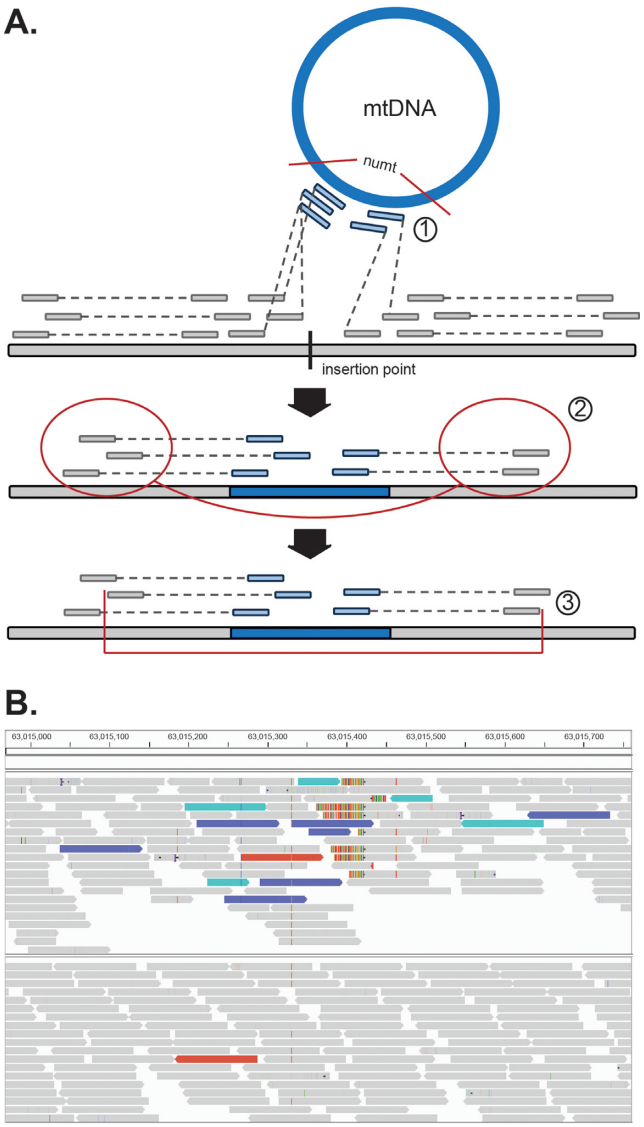

Figure 1. (**A**) Computational pipeline for Numt discovery: [1] Identification and filtration of paired reads with one read anchored to the mtDNA and another mapping to a nuclear chromosome; [2] Clustering and linking of nearby mapped nuclear reads together using insert size information; [3] Localization of insertion breakpoint using cluster distribution and truncated read alignments. (**B**) Example of Numt insertion (Poly_NumtS_2541) in sample HGDP00856 (top) compared to sample HGDP002222 (no insertion, bottom) on chromosome 8 as displayed in the IGV Browser. Sequences are represented by blocks and colored by the alignment of their mate sequence to the canonical location (gray), mitochondria genome (teal), or reference Numt homolog on chromosome 1 (blue). Multi-color bars represent split-reads whereby a portion of the sequence aligns to another location in the genome and are indicative of structural genomic breakpoints.

MATERIALS AND METHODS

Data sources

Whole genome sequences were generated as a part of Phase I of the 1000 Genomes Project (http://www.1000genomes.org) with an average 4–6X sequence coverage and from the CEPH-HGDP (SRA: SRP036155) (21,39) with a higher average coverage of 5–20X. Alignments to version GRCh37/hg19 of the human reference genome were provided in BAM format and optimized using the Genome Analysis Toolkit (GATK) (McKenna A, 2010) and Picard (http://picard.sourceforge.net/), as described elsewhere (38,39).

Detection of Numt insertions

Non-reference NumtS were discovered in paired-end, whole genome sequences using a newly developed software package named *dinumt*, as outlined in Figure 1. This approach first derives an empirical insert size distribution from the observed alignment positions of each read pair and then identifies sequences where one end aligns to either the mtDNA or a known reference Numt and the other maps elsewhere in the nuclear genome. Individual sequence reads that do not align uniquely to the nuclear genome (MAPQ < 10) are discarded, removing ambiguity from insertion sites but also limiting the ability of this approach to identify NumtS in highly repetitive sequences. These sequences are then clustered together based on their shared mapping orientation (forward or reverse) and whether or not they are within a distance of W_L from each other, where

W_L is calculated as the derived mean_insert_length + 3 x insert_standard_deviation. Clusters are further linked together if they are within a distance of 2 x W_L from each other and are in the correct orientation relative to each other (forward to reverse). Individual sequence reads are then examined within the clusters to identify soft-clipped reads with breaks at the same position to identify putative breakpoint locations. The likelihood of an insertion is then calculated as

$$L(g) = \frac{1}{m^k} \prod_{j=1}^{l} [(m-g)e_j + g(1-e_j)] \prod_{j=l+1}^{k} [(m-g)(1-e_j) + ge_j]$$

where m is the ploidy, l is the total number of reference supporting reads, k is the total number of insertion supporting reads, and e is the mapping error for read j, as modified from (40). Putative insertions were then filtered for quality (at least 50 using Phred scaled maximum insertion likelihood of non-reference allele), the number of total reads supporting the insertion (at least 4) and the depth of total coverage at the insertion point (at least 5).

Identified Numt insertions were then genotyped across the entire set of samples to identify sites which may have been previously missed or filtered in those individuals during the discovery step as well as to determine the copy number of the insertion. This was done by systematically examining each sample at an insertion location for clusters of reads supporting an insertion. In order to refine breakpoint positions, these clusters are used to search for positions where soft-clipped reads are consistently broken at the same location across samples and where the longest unaligned portion of such reads at that position map to the mtDNA reference genome. These refined positions are then utilized to determine how many sequences overlap the insertion location in an unbroken manner, supporting the reference allele, and how many contribute to the inserted Numt. These are then tabulated, combined with the read pair information and scored in a similar fashion as the discovery step above. However, the reported genotype here reflects the overall maximum likelihood of that genotype (0/0, 0/1, 1/1) and not just that of an insertion (as above). Global genotypes are then used to construct population level priors that are then applied in a Bayesian fashion for further likelihood calculations and iterated until convergence or a maximum of 10 iterations utilizing an Expectation-Maximization schema.

Validation and sequencing experiments

NumtS identified by computational analysis were validated by polymerase chain reaction (PCR) and Sanger sequencing of amplicon(s) that spanned 50–500 bp of gDNA flanking the insert, the breakpoint between the gDNA and the insert. Primer sets that hybridize to the gDNA flanking the insert were designed using Primer3 Software (http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi) and amplification was done with Platinum Taq (Invitrogen Life Technologies, Gaithersburg, MD), Picomaxx (Agilent Technologies, Palo Alto, CA, USA) or LongAmp (New England Biolabs, Beverly, MA, USA) products in a 20–50 μl reaction volume containing 50 ng of template DNA, 1-uM primer and 1.5-mM $MgCl_2$ if not supplied in the PCR buffer. Thermocycling was done for 30 cycles at 56–67°C annealing temperature and 1–15 min extension time. For inserts <3 kb, a PCR product of the predicted size was identified in individuals homozygous or heterozygous for the insert by agarose gel electrophoresis and the insert was sequenced in one individual. Amplicons of interest were purified from a PCR reaction for homozygous individuals (Qiaquick PCR purification kit, Qiagen, Valencia, CA, USA) or isolated from the gel for heterozygous individuals (Qiaquick Gel Extraction Kit, Qiagen) and sequenced at the University of Michigan Sequencing Core. For inserts larger than 3 kb, a PCR product of the predicted size was identified in individuals heterozygous for the insert by gel electrophoresis. For sequencing, two overlapping PCR products were made using primer sets designed as outlined above with one primer that binds in the gDNA flanking the insert and one primer that binds in the middle of the insert. Amplicons were purified from PCR reactions as outlined above and sequenced by primer walking at the University of Michigan Sequencing Core. Five loci failed initial validation efforts, likely due to a greater than predicted insertion size and uncertainty in the insertion breakpoints. For these loci, we performed a local assembly of the supporting reads using CAP3 (41) and designed additional PCR primers flanking the genome-insertion junction.

Enrichment analysis

We analyzed the genomic context of the regions flanking the Numt insertion positions for various characteristics, including genes, %GC content, open chromatin regions, repetitive elements, CpG Islands and microsatellites. With the exception of %GC and AT dinucleotide calculations, which were derived from the reference sequence itself using a custom PERL script, all datasets were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/) in BED format (see Supplementary Table S3 for the specific tables used). We then performed a two-tailed permutation test by resampling 1000 sets of random positions matched to our insertion set to determine whether they were significant enriched or depleted for each feature.

Phylogenetics and age estimation

An inferred ancestral mitochondria sequence was obtained from ENSEMBL Compara Release 71 based on alignment of six primate species (42). A profile of nucleotide changes was obtained by aligning this sequence to the mtDNA genome from current human reference hg19 using MEGA v5.2.2 with the Muscle algorithm (http://www.megasoftware.net). The age of each NumtS insertion and human-specific reference NumtS (>300 bp) was calculated by aligning each sequence to the previously aligned ancestral and modern mtDNA sequence. We tabulated the total number of sites in the aligned region where the ancestral and modern mitochondrial sequences differ and counted how often the NumtS sequence matched the modern human allele. We used the resulting allele matching ratio as an estimate of the point along the human lineage where the insertion occurred.

MEGA v5.2.2 (http://www.megasoftware.net) was used with our larger Numt insertions and previously reported human-specific NumtS present in the reference (13,19,37) to determine the evolutionary phylogeny of each sequence. Mitochondrial genomes were obtained for Human (Hg19), Chimpanzee (PanTro4), Gorilla (Gorgor3.1) and Orangutan (PonAbe2) from the UCSC Genome Browser, as well as mitochondrial sequences from *Neanderthals* (NCBI Accessions: KC879692, FM865411, FM865410, FM865408, FM865407), old European fossil (Homo heidelbergensis) (NCBI Accession: NC_023100) and Denisova (NCBI Accession: FN673705) that were downloaded from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov). The NumtS sequences were aligned to these mtDNA reference sequences using Muscle alignment tool (43) and the trees were built using Maximum-Likelihood method with bootstrap values.

Haplotyping and analysis of heteroplasmy

Single nucleotide polymorphism (SNP) profiles for each inserted sequence were generated using mtDNAprofiler (44). The identified variants were then annotated using Haplogrep (45) to identify potential haplotypes of origin. These SNP profiles were also used to assess potential heteroplasmy artifacts by direct comparison with reported events (22,24,25,33,46–48). To be considered a match, both the position and the Numt allele must match what was reported and sample specific information was also indicated where available (22).

Software and data availability

The genomic locations and sequences for the identified NumtS are provided as Supplementary Data to this manuscript. Sequences of mitochondria insertions and immediate flanking regions have been submitted to GenBank (Accession: KM281512–KM281534). The software package *dinumt* is available for download at https://bitbucket.org/remills/dinumt.

RESULTS

Detecting Numt insertions in whole genome sequences

Our discovery approach identifies clusters of read pairs mapping to both the nuclear and mitochondrial genomes and then examines these regions for potential insertion events (Figure 1). This is similar in principle to earlier strategies designed to identify insertions of novel genetic material by finding clusters of one-end anchored reads (23,49) and to discovering mobile element insertions (50–54), but requires that reads map to either the mitochondria or known reference NumtS to report a putative insertion (see 'Materials and Methods' section). Nearby clusters are grouped together based on their orientation and distance from each other and the surrounding region is examined for split reads that map partially to both the chromosome and mitochondria, indicating the precise molecular breakpoint of the insertion. These sites can then be systematically genotyped across the entire sample set using a statistical framework similar to that previously developed for SNPs (40) to determine the insertion copy number in each individual.

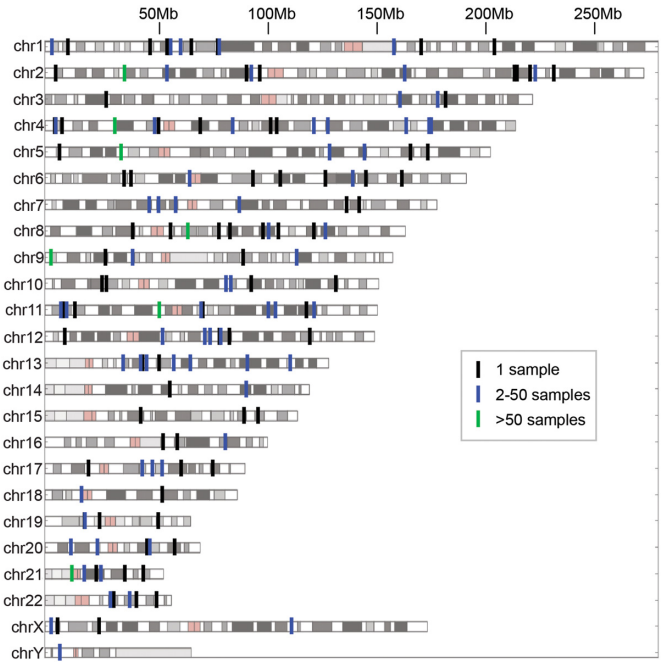

Figure 2. Chromosomal locations of individual Numt insertions, with color indicating its initial discovery in one sample (black), 2–50 samples (blue) and more than 50 samples (green).

We applied our method to 946 low coverage, whole genomes that were sequenced in Phase 1 of the 1000 Genomes Project (38) as well as 53 additional genomes sequenced to higher coverage from the HGDP (21,39) and were able to identify 141 polymorphic nuclear insertions of mitochondrial origin among all nuclear chromosomes except chrY, (Figure 2), including three which had been previously characterized (37). On average, ~1.5 non-reference NumtS were seen in each sample, which showed a statistically significant but modest correlation with coverage (r^2 = 0.12, Supplementary Figure S1) suggesting that although some NumtS may be missed due to low coverage, they are likely minimal. These insertions were fairly evenly distributed among all 20 different populations assessed (Supplementary Figures S2 and S3).

We next assessed the overall accuracy of our approach. Using PCR, we were able to verify 23/24 of the predicted Numt insertion sites in the HGDP samples in which they were discovered and a further 17/18 from a subset of the lower coverage 1000 Genomes samples (Figure 3A, Supplementary Table S1), with the events that we did not validate occurring either in segmental duplications or having uncertain breakpoints with potentially large insertion sizes, making them difficult to conclusively verify. Additional validation with PCR panels across multiple samples showed a concordance for 713/748 (95.3%) of our predicted allele genotypes (Figure 3B). We further validated that these were indeed of mitochondrial origin and not post-insertion duplications by Sanger sequencing through the breakpoints for 23 events (Supplementary Table S2. These results sug-

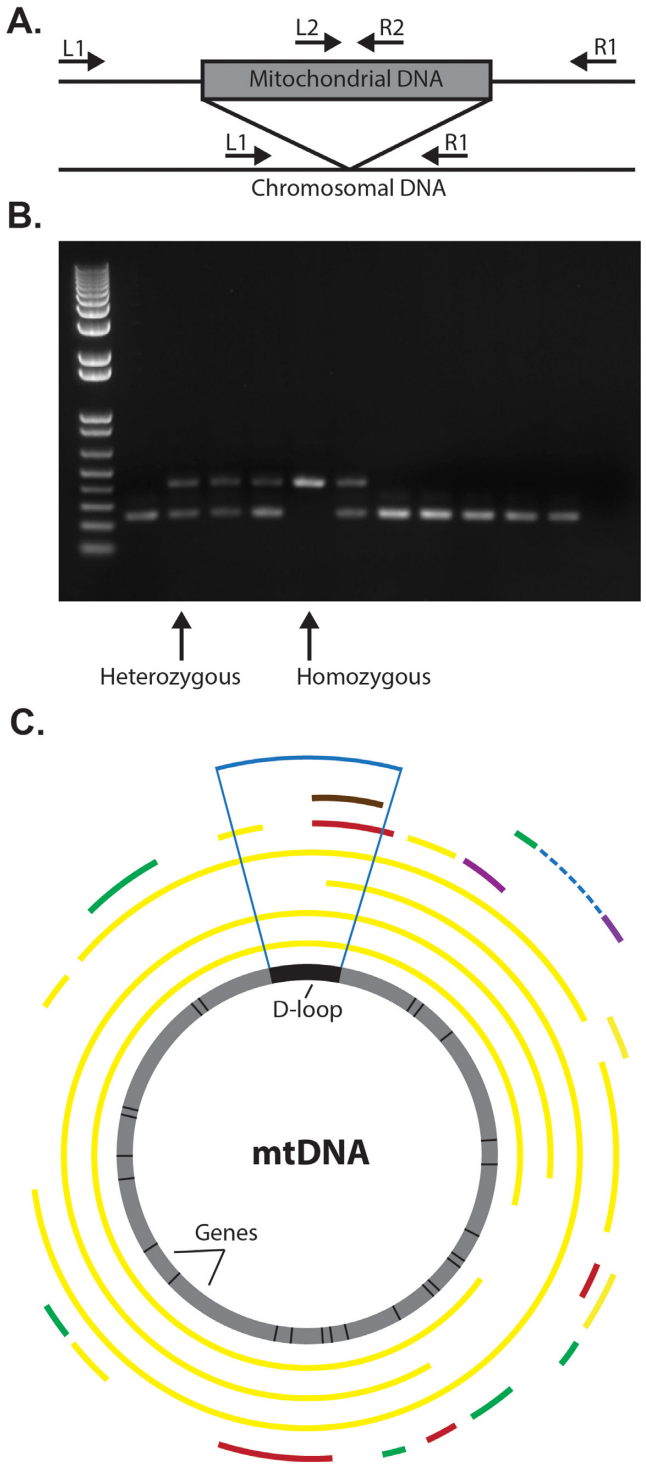

Figure 3. (**A**) PCR strategy for validating Numt polymorphisms. Short insertions can be directly assessed using outer primers (L1, R1), while larger insertions require the additional use of internal primers (L1,R2; L2,R1). (**B**) Representative PCR validation panel for numt Poly_NumtS_1843 located on chromosome 4, with heterozygous and homozygous polymorphisms labeled with arrows. (**C**) Genomic context of sequenced polymorphic Numt insertions with respect to their Mitochondrial origin. Sequence identity of each Numt to the consensus mtDNA is indicated by color: 100% (green), 99% (yellow), 98% (purple), 97% (red) and 96% (brown). The D-loop region is denoted with vertical blue lines.

gest that *dinumt* is able to accurately discover Numt insertions in whole genome sequence data.

Characteristics and enrichment of Numt insertions

We conducted an analysis to confirm whether the insertion positions of these recent NumtS co-localized with specific genomic features, as has previously been assessed with the older, fixed events present in the reference genome (Supplementary Table S3). Using a series of permutation tests, we found no enrichment in regions containing CpG islands, microsatellites and other types of structural variants ($P > 0.05$). Most Numt insertions were in intronic (42%) and intergenic regions (43%), consistent with expectations from random sampling, and we observed no insertions into coding exons. NumtS were also found in the 5′ and 3′ UTR's as well as promoter and terminator regions (5 kbp up and downstream, respectively), albeit at a much lower frequency. Consistent with previous reports, we found a significant enrichment near repetitive regions ($P \leq 0.004$) (16) and an insertion preference for slightly higher%GC regions overall (Supplementary Figure S4).

Interestingly, we neither observed enrichment for A + T oligomers immediately adjacent to the polymorphic insertions (Supplementary Figure S5) nor a preference for open chromatin regions in the cell lines we investigated ($P > 0.05$), as had been previously reported (16). This enrichment was prevalent even when limited to those NumtS for which we had validated breakpoints. To verify the consistency of our permutation approach, we applied the same analysis to the 610 reference NumtS described in Tsuji *et al.* as well as the human-specific NumtS described in Lang *et al.* (37) and were able to replicate their results. This difference at the insertion sites could represent a *bona fide* change in the integration mechanism for recent NumtS, but may also be an artifact from the way reference NumtS have been annotated relative to the mitochondria genome sequence. It is also possible that this is reflective of a technical difference between the identification of NumtS from short-read, next generation sequences and those found in longer capillary sequences from which the reference genome was constructed, as the latter would have fewer potential biases due to alignment artifacts in repetitive regions.

Analysis of polymorphic Numt sequences

While the precise insertion location for each Numt is informative, the underlying sequence itself can provide additional information. Using either direct Sanger sequencing of PCR products or subsequent primer walking and assembly for larger insertions (see 'Materails and Methods' section, Supplementary Table S1), we were able to determine the sequence for 23-Numt insertions (Supplementary Table S2). Most NumtS sequences were small (<500 bp), however we did identify a number of larger events including an almost complete mitochondrial genome insertion of 16 106 bp in sample HGDP01275. We observed fragments originating from all parts of the mitochondrial genome, including multiple sequences overlapping the D-loop region (Figure 2C) which had been previously reported as underrepresented in the human reference (16). Interestingly, these

polymorphic sequences exhibited a higher%GC than that of their parent mitochondria genome (47 versus 44%) and also showed similar characteristics to other human-specific NumtS in the reference genome (Supplementary Figure S6). In contrast, non-human specific NumtS in the reference showed a markedly lower%GC that is more consistent with the average nuclear genome%GC of 41.5%.

Using these sequences, we estimated when these insertions occurred in the human lineage by comparing the human mtDNA reference sequence to an inferred ancestral mitochondrial sequence and identifying diagnostic mutations that matched specific positions in each Numt sequence (see 'Materials and Methods' section). We then used the fraction of alleles that matched those from the modern human mitochondrial sequence to derive an approximate age for each insertion, relative to an estimated human-chimpanzee divergence time of 6 million years (Table 1). We observed that most of the polymorphic insertions occurred within the past 1 million years, however there were six NumtS that were markedly older, including two that appear to have inserted over 2.5 million years in the past. We next constructed maximum-likelihood trees to compare fixed human-specific NumtS present in the reference genome (Figure 4A) with the discovered polymorphic NumtS (Figure 4B). As expected, the ongoing polymorphisms co-localized with the human lineage while the fixed events were likely inserted further back in time.

Assessment of Numt impact on heteroplasmy

Although many recent studies of mitochondrial heteroplasmy have taken NumtS into consideration (22,24,25,33,46–48), they have all been limited to those insertions present in the reference sequence. To assess the potential impact of more recent insertions, we compared our set of sequenced polymorphic NumtS to these studies by identifying single nucleotide differences in the Numt insertions relative to the mtDNA reference and comparing the allelic changes to those reported (Table 2). We identified 59 positions of possible Numt confounding, most of which occur in polymorphic insertions common in the general human population (MAF $>$ 0.01). The samples used in most of these studies differ from those analyzed here, making direct inferences regarding these effects of NumtS difficult. However, one study (22) had an intersecting set of individuals with our analysis, and we were able to determine that there were eight positions within NumtS that were genotyped in those samples and had alleles matching the reported heteroplasmy.

DISCUSSION

Almost every eukaryotic species that has had their genome fully sequenced to date has exhibited evidence for the transfer of organelle DNA to the nuclear genome. These NumtS occur in both animals and plants, and show a strong correlation with genome size and the total number of NumtS observed (1). In humans, there are ∼755 annotated NumtS in the reference genome (13), though this number is variable depending on the methods and parameters used to identify their presence. However, very few of these are due to recent

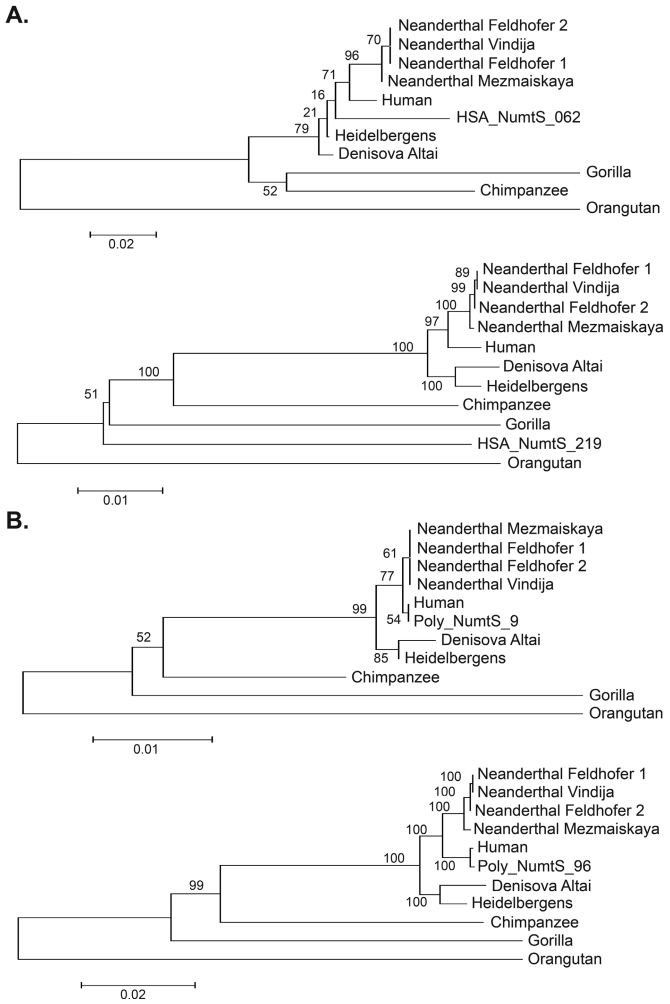

Figure 4. Phylogenetic trees for select (**A**) fixed and (**B**) polymorphic Numt insertion sequences relative to various species along the human and other primate lineages using distances based on nucleotide substitution rates. Fixed NumtS were previously identified as human-specific (19) and are present in the human reference sequence (hg19) as well as every 1000 Genomes sample assessed (38). Polymorphic sequences were chosen from among the longest insertions that were identified. Bootstrap values are indicated at branch locations.

insertions and almost all NumtS that have been identified are present in every human genome. Indeed, only 14 events differentially present in human populations have been previously reported (37).

Here, we present a large-scale analysis of polymorphic mitochondrial insertions into the nuclear genome of humans. These recent Numt insertions share many characteristics with previously identified human-specific NumtS that are fixed in the genome, including their patterns of integration within the genome and sequence composition. We identified many NumtS that contain the mitochondrial D-loop, a noncoding region of the mitochondria that controls the synthesis of DNA and RNA within the organelle and typically exhibits a higher mutation rate than the rest of the genome. This region is often used in forensic and population genetics due to its two hypervariable regions that provide distinguishing polymorphisms between individuals

Table 1. Age estimations of sequenced Numt insertions using a consensus ancestral mitochondrial sequence

Numt ID	Length of insertion	# of Diagnostic positions	# of positions matching human	% of human mtDNA	[a]Estimated age of insertion (MYA)
Poly_NumtS_139	391	20	20	100	<0.1
Poly_NumtS_1239	246	8	7	88	0.72
Poly_NumtS_1259*	547	40	37	92	0.48
Poly_NumtS_1440	144	9	9	100	<0.1
Poly_NumtS_1843	163	11	10	91	0.54
Poly_NumtS_1900	166	5	3	60	2.4
Poly_NumtS_1929	229	8	7	88	0.72
Poly_NumtS_2010	153/147[c]	16	15	94	0.36
Poly_NumtS_2186*	245	0	[b]NA		
Poly_NumtS_2289*	11840	577	426	86	0.84
Poly_NumtS_2377	196	8	8	100	<0.1
Poly_NumtS_2541	1412	68	50	74	1.56
Poly_NumtS_2578*	16091	678	664	98	0.12
Poly_NumtS_2611	119	6	6	100	<0.1
Poly_NumtS_2653	1665	91	69	76	1.44
Poly_NumtS_316*	655	2	2	100	<0.1
Poly_NumtS_430*	482	39	29	74	1.56
Poly_NumtS_445	145	3	3	100	<0.1
Poly_NumtS_531	269	17	16	94	0.36
Poly_NumtS_709	140	7	3	43	3.42
Poly_NumtS_1465*	6775	261	256	98	0.12
Poly_NumtS_1480*	13783	580	447	77	1.38
Poly_NumtS_1583	126	6	5	83	1.02

[a]Based on approximate divergence between human and chimpanzee of 6 million years (20).
[b]Poly_NumtS_2186 did not overlap any of the identified diagnostic differences between the human and ancestral mitochondria.
[c]Poly_NumtS_2010 had two noncontiguous fragments of lengths 153 and 147.
*Contains hypervariable D-loop region.

(55,56) and has been previously found to be depleted among NumtS present in the human reference sequence (16). Previous studies have found little effect of existing, older NumtS on these types of assays (57), but have not taken into account these more recent insertions that are more likely to cause off-target amplification and erroneous conclusions. Our methods and datasets should thus provide a useful resource in these types of analyses.

There are a number of limitations to our approach, many of which are shared with other general strategies for identifying genetic variation using next generation sequencing. The depth to which an individual genome is sequenced can have an effect on variant detection accuracy (58), leading to potentially missed polymorphisms. Our results showed a modest correlation between sequence coverage and the number of insertions identified, so it is likely that a small number of NumtS were missed due to the use of low coverage genomes in our analysis. It is also very challenging to identify variation in highly repetitive sequences, and our analytical filters preclude the identification of NumtS in such regions of the genome. As such, it is possible that the results of our enrichment analysis are biased due to their omission. This is a technical obstacle, however, and should be resolved as future sequencing technologies are established.

The polymorphic nature of these insertions within the human lineage indicates that they have likely occurred since the most recent common ancestor between humans and chimpanzees, however it is unknown where they are relative to other species of humans such as *Neanderthals* or *Denisovans*. We thus attempted to date our insertions both through direct comparison with a consensus ancestral mitochondrial as well as through phylogenic analysis and found that most were integrated into the nuclear genome within the past 100 000 years, although the small number of substitutions in many fragments makes their exact age difficult to be precisely determined. Over half of the NumtS we discovered were present in very low frequencies across the samples we interrogated (MAF < 0.1%), suggesting that they were likely integrated even more recently than the resolution of our analysis would allow. This supports the theory that mitochondria gene transfer to the human nuclear genome is ongoing and prevalent (1,3).

NumtS have been previously implicated in a number of sporadic disease cases through their integration into functional regions of the genome (8–10). While we did not identify any NumtS that would directly affect the coding region of a gene, we did identify a Numt insertion in a single individual that represented an almost entire insertion of a mitochondrial genome into chromosomal DNA (Figure 5). This insertion was 16 106 bp in size and integrated into a potential regulatory region in the first intron of the SDC2 gene (59), a member of the syndecan family that encodes an integral membrane protein and has been associated with cell proliferation and migration, including altered expression in several cancer cells (60,61). We investigated whether this insertion may have had an effect on the expression of this gene by looking at recently published RNA-Seq data over these same samples (39), however this gene is not expressed in the tissues which were used in that analysis and so we are unable to draw any conclusions regarding its potential impact. It is tempting to speculate, however, that an insertion that is highly enriched for functional regions could indeed affect

Table 2. Examples of Numt insertion alleles matching identified mitochondrial heteroplasmic positions

Position	mtDNA allele	Numt allele	Numt ID(s)	[a]Max allele frequency	[b]Study ref.
73	A	G	Poly_NumtS_1465;Poly_NumtS_1480	0.026	1,5,7
263	A	G	Poly_NumtS_1465;Poly_NumtS_1480	0.026	7
489	T	C	Poly_NumtS_1465	0.007	5,7
750	A	G	Poly_NumtS_1465;Poly_NumtS_1480	0.026	7
1438	A	G	Poly_NumtS_1465;Poly_NumtS_1480	0.026	7
2706	A	G	Poly_NumtS_1465;Poly_NumtS_1480	0.026	5,7
4769	A	G	Poly_NumtS_2289	0.003	7
5460	G	A	Poly_NumtS_2289	0.003	2,5
7028	C	T	Poly_NumtS_2289	0.003	5,7
7220	T	C	Poly_NumtS_2541	0.154	5
7256	C	T	Poly_NumtS_2289	0.003	7
7521	G	A	Poly_NumtS_2289;Poly_NumtS_2541	0.154	7
7861	T	C	Poly_NumtS_2541	0.154	6,7
7912	G	A	Poly_NumtS_2377	0.016	3
7927	C	T	Poly_NumtS_2541	0.154	5
8021	A	G	Poly_NumtS_2377	0.016	1
8122	A	G	Poly_NumtS_2541	0.154	5
8152	G	A	Poly_NumtS_2377	0.016	2
8206	G	A	Poly_NumtS_2541	0.154	7
8251	G	A	Poly_NumtS_2541	0.154	5
11176	G	A	Poly_NumtS_1900	0.023	7*
12612	A	G	Poly_NumtS_2653	0.003	5,6,7
12630	G	A	Poly_NumtS_2653	0.003	5
12705	C	T	Poly_NumtS_2653;Poly_NumtS_1480	0.026	5,7
13506	C	T	Poly_NumtS_2653	0.026	5,7
13650	C	T	Poly_NumtS_2653	0.026	7
14766	C	T	Poly_NumtS_1465;Poly_NumtS_1480	0.026	5,7
15043	G	A	Poly_NumtS_1465	0.007	7
15301	G	A	Poly_NumtS_1465	0.007	5,6,7
15326	A	G	Poly_NumtS_1465;Poly_NumtS_1480	0.026	7
15575	G	C	Poly_NumtS_1480	0.026	5
16093	T	C	Poly_NumtS_1259	0.013	1,2,4,5,7*
16129	G	A	Poly_NumtS_430	0.568	4,7*
16189	T	A	Poly_NumtS_430	0.568	4
16209	T	C	Poly_NumtS_1259	0.013	6,7
16218	C	T	Poly_NumtS_430	0.568	4,7*
16223	C	T	Poly_NumtS_1259;Poly_NumtS_1465;Poly_NumtS_1480	0.026	2,4,5,6,7
16230	A	G	Poly_NumtS_1259;Poly_NumtS_430	0.568	4,7
16234	C	T	Poly_NumtS_1259	0.013	5
16249	T	C	Poly_NumtS_430	0.568	4,7
16259	C	A	Poly_NumtS_430	0.568	4
16263	T	C	Poly_NumtS_430	0.568	4
16264	C	T	Poly_NumtS_430	0.568	4
16274	G	A	Poly_NumtS_430	0.568	4
16278	C	T	Poly_NumtS_1259;Poly_NumtS_430	0.568	4,5,7*
16284	A	G	Poly_NumtS_430	0.568	4
16288	T	C	Poly_NumtS_430	0.568	4
16290	C	T	Poly_NumtS_430	0.568	4
16293	A	C	Poly_NumtS_430	0.568	4,5
16301	C	T	Poly_NumtS_430	0.568	4
16311	T	C	Poly_NumtS_1259;Poly_NumtS_430	0.568	4,5,6,7*
16355	C	T	Poly_NumtS_1259;Poly_NumtS_430	0.568	4,5
16356	T	C	Poly_NumtS_1259;Poly_NumtS_430	0.568	4
16362	T	C	Poly_NumtS_1259	0.013	2,5,7
16368	T	C	Poly_NumtS_430	0.568	4
16390	G	A	Poly_NumtS_430	0.568	4,7*
16519	T	C	Poly_NumtS_1259;Poly_NumtS_430	0.568	5,7*
16527	C	T	Poly_NumtS_1259;Poly_NumtS_430	0.568	5

[a]Largest Numt allele frequency is listed from among sites with multiple overlapping insertions.
[b]Studies included are 1:(24), 2:(48), 3:(25), 4:(46), 5:(47), 6:(33), and 7:(22). It should be noted that (5) identified heteroplasmy in RNA sequences while the remaining studies were from DNA.
*Matches sample specific heteroplasmy allele as reported in (22).

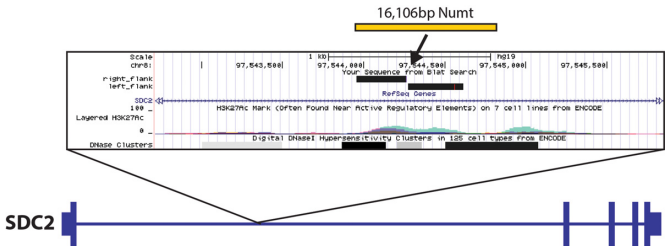

Figure 5. Insertion of an almost full-length mitochondrial insertion (Poly_NumtS_2578) into the first intron of the SDC2 gene in sample HGDP01275. Zoomed panel shows UCSC Genome Browser (http://genome.ucsc.edu) view of 3 kbp surrounding region of insertion, with sequenced flanking regions of the insertion breakpoint indicated by solid black rectangles.

canonical gene structure and expression, and ongoing studies in individual tissues from projects such as GTEx (62) and others may provide they keys for further investigating these types of events.

Finally, we explored the potential effect of our NumtS on studies of mitochondrial heteroplasmy and identified a number of positions within the mitochondrial genome that could be erroneously attributed to mutations in NumtS. It is possible that these heteroplasmies are prevalent and the allelic changes in the NumtS occurred prior to their insertion, and indeed a recent study using RNA-Seq expression data to examine heteroplasmy identified a number of these same positions (47). While it is possible that NumtS may be expressed at some low level in the nucleus, it is more likely that the reported heteroplasmies are *bona fide* mitochondrial differences. However we believe that our set of genotypes and insertion sequences will be a useful resource for future studies into mitochondrial heterogeneity.

ACCESSION NUMBERS

KM281512, KM281513, KM281514, KM281515, KM281516, KM281517, KM281518, KM281519, KM281520, KM281521, KM281522, KM281523, KM281524, KM281525, KM281526, KM281527, KM281528, KM281529, KM281530, KM281531, KM281532, KM281533, KM281534.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENT

We would like to thank Xuefang Zhao for her help with figure construction.

FUNDING

National Institute of Health [1R01-HG007068-01A1 to R.E.M. and DP5OD009154 to J.M.K.]; Funding for open access charge: National Institutes of Health [1R01-HG007068-01A1].
Conflict of interest statement. None declared.

REFERENCES

1. Hazkani-Covo,E., Zeller,R.M. and Martin,W. (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet.*, **6**, e1000834.
2. Soto-Calderon,I.D., Lee,E.J., Jensen-Seaman,M.I. and Anthony,N.M. (2012) Factors affecting the relative abundance of nuclear copies of mitochondrial DNA (numts) in hominoids. *J. Mol. Evol.*, **75**, 102–111.
3. Ricchetti,M., Tekaia,F. and Dujon,B. (2004) Continued colonization of the human genome by mitochondrial DNA. *PLoS Biol.*, **2**, E273.
4. Hazkani-Covo,E. and Covo,S. (2008) Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS Genet.*, **4**, e1000237.
5. Mourier,T., Hansen,A.J., Willerslev,E. and Arctander,P. (2001) The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus. *Mol. Biol. Evol.*, **18**, 1833–1837.
6. Leister,D. (2005) Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends Genet.*, **21**, 655–663.
7. Willett-Brozick,J.E., Savul,S.A., Richey,L.E. and Baysal,B.E. (2001) Germ line insertion of mtDNA at the breakpoint junction of a reciprocal constitutional translocation. *Hum. Genet.*, **109**, 216–223.
8. Borensztajn,K., Chafa,O., Alhenc-Gelas,M., Salha,S., Reghis,A., Fischer,A.M. and Tapon-Bretaudiere,J. (2002) Characterization of two novel splice site mutations in human factor VII gene causing severe plasma factor VII deficiency and bleeding diathesis. *Br. J. Haematol.*, **117**, 168–171.
9. Turner,C., Killoran,C., Thomas,N.S., Rosenberg,M., Chuzhanova,N.A., Johnston,J., Kemel,Y., Cooper,D.N. and Biesecker,L.G. (2003) Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. *Hum. Genet.*, **112**, 303–309.
10. Goldin,E., Stahl,S., Cooney,A.M., Kaneski,C.R., Gupta,S., Brady,R.O., Ellis,J.R. and Schiffmann,R. (2004) Transfer of a mitochondrial DNA fragment to MCOLN1 causes an inherited case of mucolipidosis IV. *Hum. Mutat.*, **24**, 460–465.
11. Ahmed,Z.M., Smith,T.N., Riazuddin,S., Makishima,T., Ghosh,M., Bokhari,S., Menon,P.S., Deshmukh,D., Griffith,A.J., Riazuddin,S. *et al.* (2002) Nonsyndromic recessive deafness DFNB18 and Usher syndrome type IC are allelic mutations of USHIC. *Hum. Genet.*, **110**, 527–531.
12. Yao,Y.G., Kong,Q.P., Salas,A. and Bandelt,H.J. (2008) Pseudomitochondrial genome haunts disease studies. *J. Med. Genet.*, **45**, 769–772.
13. Calabrese,F.M., Simone,D. and Attimonelli,M. (2012) Primates and mouse NumtS in the UCSC Genome Browser. *BMC Bioinformatics*, **13(Suppl. 4)**, S15.
14. Blanchard,J.L. and 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
15. Ricchetti,M., Fairhead,C. and Dujon,B. (1999) Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature*, **402**, 96–100.
16. Tsuji,J., Frith,M.C., Tomii,K. and Horton,P. (2012) Mammalian NUMT insertion is non-random. *Nucleic Acids Res.*, **40**, 9073–9088.
17. Mishmar,D., Ruiz-Pesini,E., Brandon,M. and 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.
18. Gherman,A., Chen,P.E., Teslovich,T.M., Stankiewicz,P., Withers,M., Kashuk,C.S., Chakravarti,A., Lupski,J.R., Cutler,D.J. and Katsanis,N. (2007) Population bottlenecks as a potential major shaping force of human genome architecture. *PLoS Genet.*, **3**, e119.
19. Jensen-Seaman,M.I., Wildschutte,J.H., Soto-Calderon,I.D. and Anthony,N.M. (2009) A comparative approach shows differences in patterns of numt insertion during hominoid evolution. *J. Mol. Evol.*, **68**, 688–699.
20. Meyer,M., Fu,Q., Aximu-Petri,A., Glocke,I., Nickel,B., Arsuaga,J.L., Martinez,I., Gracia,A., de Castro,J.M., Carbonell,E. *et al.* (2014) A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature*, **505**, 403–406.
21. Cann,H.M., de Toma,C., Cazes,L., Legrand,M.F., Morel,V., Piouffre,L., Bodmer,J., Bodmer,W.F., Bonne-Tamir,B.,

Cambon-Thomsen,A. *et al.* (2002) A human genome diversity cell line panel. *Science*, **296**, 261–262.
22. Diroma,M.A., Calabrese,C., Simone,D., Santorsola,M., Calabrese,F.M., Gasparre,G. and Attimonelli,M. (2014) Extraction and annotation of human mitochondrial genomes from 1000 Genomes Whole Exome Sequencing data. *BMC Genomics*, **15**, S2.
23. Hajirasouliha,I., Hormozdiari,F., Alkan,C., Kidd,J.M., Birol,I., Eichler,E.E. and Sahinalp,S.C. (2010) Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics*, **26**, 1277–1283.
24. He,Y., Wu,J., Dressman,D.C., Iacobuzio-Donahue,C., Markowitz,S.D., Velculescu,V.E., Diaz,L.A. Jr, Kinzler,K.W., Vogelstein,B. and Papadopoulos,N. (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature*, **464**, 610–614.
25. Ramos,A., Santos,C., Mateiu,L., Mdel,M., Alvarez,L., Azevedo,L., Amorim,A. and Aluja,M.P. (2013) Frequency and pattern of heteroplasmy in the complete human mitochondrial genome. *PLoS One*, **8**, e74636.
26. Ye,K., Lu,J., Ma,F., Keinan,A. and Gu,Z. (2014) Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proc. Natl. Acad. Sci. U.S.A.*, **111** , 10654–10659.
27. Ross,J.M., Stewart,J.B., Hagstrom,E., Brene,S., Mourier,A., Coppotelli,G., Freyer,C., Lagouge,M., Hoffer,B.J., Olson,L. *et al.* (2013) Germline mitochondrial DNA mutations aggravate ageing and can impair brain development. *Nature*, **501**, 412–415.
28. Wallace,D.C. (1994) Mitochondrial DNA sequence variation in human evolution and disease. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 8739–8746.
29. Gasparre,G., Porcelli,A.M., Lenaz,G. and Romeo,G. (2013) Relevance of mitochondrial genetics and metabolism in cancer development. *Cold Spring Harb. Perspect. Biol.*, **5**.
30. Avital,G., Buchshtav,M., Zhidkov,I., Feder,J., Dadon,S., Rubin,E., Glass,D., Spector,T.D. and Mishmar,D. (2012) Mitochondrial DNA heteroplasmy in diabetes and normal adults: role of acquired and inherited mutational patterns in twins. *Hum. Mol. Genet.*, **21**, 4214–4224.
31. Song,H., Buhay,J.E., Whiting,M.F. and Crandall,K.A. (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 13486–13491.
32. Goto,H., Dickins,B., Afgan,E., Paul,I.M., Taylor,J., Makova,K.D. and Nekrutenko,A. (2011) Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. *Genome Biol.*, **12**, R59.
33. Jayaprakash,A.D., Benson,E., Liang,R., Shim,J., Lambertini,L., Wigler,M., Aaronson,S. and Sachidanandam,R. (2014) Mito-seek enables deep analysis of mitochondrial DNA, revealing ubiquitous, stable heteroplasmy maintained by intercellular exchange. *bioRxiv*, doi:10.1101/007005.
34. Wolff,J.N. (2014) Targeted and robust amplification of mitochondrial DNA in the presence of nuclear-encoded mitochondrial pseudogenes using Phi29 DNA polymerases. *Methods Mol. Biol.*, **1167**, 255–263.
35. Zischler,H., Geisert,H., von Haeseler,A. and Paabo,S. (1995) A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans. *Nature*, **378**, 489–492.
36. Thomas,R., Zischler,H., Paabo,S. and Stoneking,M. (1996) Novel mitochondrial DNA insertion polymorphism and its usefulness for human population studies. *Hum. Biol.*, **68**, 847–854.
37. Lang,M., Sazzini,M., Calabrese,F.M., Simone,D., Boattini,A., Romeo,G., Luiselli,D., Attimonelli,M. and Gasparre,G. (2012) Polymorphic NumtS trace human population relationships. *Hum. Genet.*, **131**, 757–771.
38. Genomes Project,C., Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
39. Martin,A.R., Costa,H.A., Lappalainen,T., Henn,B.M., Kidd,J.M., Yee,M.C., Grubert,F., Cann,H.M., Snyder,M., Montgomery,S.B. *et al.* (2014) Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. *PLoS Genet.*, **10**, e1004549.
40. Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.
41. Huang,X. and Madan,A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
42. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–D755.
43. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
44. Yang,I.S., Lee,H.Y., Yang,W.I. and Shin,K.J. (2013) mtDNAprofiler: a Web application for the nomenclature and comparison of human mitochondrial DNA sequences. *J. Forensic Sci.*, **58**, 972–980.
45. Kloss-Brandstatter,A., Pacher,D., Schonherr,S., Weissensteiner,H., Binna,R., Specht,G. and Kronenberg,F. (2011) HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.*, **32**, 25–32.
46. Bintz,B.J., Dixon,G.B. and Wilson,M.R. (2014) Simultaneous detection of human mitochondrial DNA and nuclear-inserted mitochondrial-origin sequences (NumtS) using forensic mtDNA amplification strategies and pyrosequencing technology. *J. Forensic Sci.*, **59**, 1064–1073.
47. Hodgkinson,A., Idaghdour,Y., Gbeha,E., Grenier,J.C., Hip-Ki,E., Bruat,V., Goulet,J.P., de Malliard,T. and Awadalla,P. (2014) High-resolution genomic analysis of human mitochondrial RNA sequence variation. *Science*, **344**, 413–415.
48. Li,M., Schonberg,A., Schaefer,M., Schroeder,R., Nasidze,I. and Stoneking,M. (2010) Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. *Am. J. Hum. Genet.*, **87**, 237–249.
49. Kidd,J.M., Cooper,G.M., Donahue,W.F., Hayden,H.S., Sampas,N., Graves,T., Hansen,N., Teague,B., Alkan,C., Antonacci,F. *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.
50. Ewing,A.D. and Kazazian,H.H. Jr (2011) Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res.*, **21**, 985–990.
51. Stewart,C., Kural,D., Stromberg,M.P., Walker,J.A., Konkel,M.K., Stutz,A.M., Urban,A.E., Grubert,F., Lam,H.Y., Lee,W.P. *et al.* (2011) A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.*, **7**, e1002236.
52. Keane,T.M., Wong,K. and Adams,D.J. (2013) RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics*, **29**, 389–390.
53. Quinlan,A.R., Clark,R.A., Sokolova,S., Leibowitz,M.L., Zhang,Y., Hurles,M.E., Mell,J.C. and Hall,I.M. (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.*, **20**, 623–635.
54. Hormozdiari,F., Hajirasouliha,I., Dao,P., Hach,F., Yorukoglu,D., Alkan,C., Eichler,E.E. and Sahinalp,S.C. (2010) Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, **26**, i350–i357.
55. Szibor,R., Michael,M., Plate,I. and Krause,D. (2000) Efficiency of forensic mtDNA analysis. Case examples demonstrating the identification of traces. *Forensic Sci. Int.*, **113**, 71–78.
56. Budowle,B., Wilson,M.R., DiZinno,J.A., Stauffer,C., Fasano,M.A., Holland,M.M. and Monson,K.L. (1999) Mitochondrial DNA regions HVI and HVII population data. *Forensic Sci. Int.*, **103**, 23–35.
57. Goios,A., Amorim,A. and Pereira,L. (2006) Mitochondrial DNA pseudogenes in the nuclear genome as possible sources of contamination. *Int. Congr. Ser.*, **1288**, 697–699.
58. Sims,D., Sudbery,I., Ilott,N.E., Heger,A. and Ponting,C.P. (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.*, **15**, 121–132.
59. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
60. De Oliveira,T., Abiatari,I., Raulefs,S., Sauliunaite,D., Erkan,M., Kong,B., Friess,H., Michalski,C.W. and Kleeff,J. (2012) Syndecan-2 promotes perineural invasion and cooperates with K-ras to induce an invasive pancreatic cancer cell phenotype. *Mol. Cancer*, **11**, 19.
61. Oh,T., Kim,N., Moon,Y., Kim,M.S., Hoehn,B.D., Park,C.H., Kim,T.S., Kim,N.K., Chung,H.C. and An,S. (2013) Genome-wide identification and validation of a novel methylation biomarker, SDC2, for blood-based detection of colorectal cancer. *J. Mol. Diagn.*, **15**, 498–507.
62. G. TEx Consortium. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.