Alignment-Free Approaches Predict Novel Nuclear Mitochondrial Segments (NUMTs) in the Human Genome

Wentian Li, Jerome Freudenberg

The Robert S. Boas Center for Genomics and Human Genetics

The Feinstein Institute for Medical Research, Northwell Health, Manhasset, NY, USA

December 2017

Abstract

Regions of the nuclear genome may harbor sequences of mitochondrial origin, indicating an ancestral transfer of mitochondrial DNA into the nuclear genome. These Nuclear Mitochondrial Segments (NUMTs) are traditionally detected by sequence similarity search, as implemented in the Basic Local Alignment Search Tool (BLAST). Confidently detecting such NUMTs is important for the comprehensive annotation of the human genome. Here we explore the possibility of detecting NUMTs in the human genome by alignment-free approaches, such as k-mers (k-tuples, k-grams, oligos of length k) distributions. We find that when k=6 or larger, the k-mer approach and BLAST search produce almost identical results, e.g., detect the same set of NUMTs longer than 3kb. However, when k=5 or k=4, certain signals are only detected by the alignment-free approach, and these may indicate yet unrecognized, and potentially more ancestral NUMTs. We introduce a "Manhattan plot" style representation of NUMTs predictions across the genome, which are calculated based on the reciprocal of the Jensen-Shannon divergence between the nuclear and mitochondrial k-mer frequencies. An inspection of Genome Browser annotations shows that most of the k-mer specific NUMT predictions contain Long Terminal Repeat (LTR), whereas BLAST based NUMT predictions do not.

Abbreviations: BLAST: basic local alignment search tool; JS: Jensen-Shannon (divergence); kb: kilo (1000) bases; KL: Kullback-Leibler (divergence); LTR: long terminal repeat; Mb: million (1,000,000) bases; MDS: multi dimensional scaling; MT: mitochondrial; NUMT: nuclear mitochondrial DNA segments/sequences;

Introduction

Under the endosymbiosis hypothesis, mitochondria have a bacterial ancestry and co-evolve with their eukaryotic host cells, benefitting both entities by this co-existence (Margulis, 1970). Such a co-existence implies interaction between the two units. Early research had found hybridization of nuclear and mitochondrial DNA fragments in mouse (Du Buy and Riley, 1967). Another study, driven by the curiosity that the ATPases genes are encoded in mitochondrial DNA in *S. cerevisiae*, whereas the same gene is coded in nuclear DNA in a related species, *N. crassa*, discovered that the gene has both copies in mitochondrial and nuclear genome in *N. crassa* (Van Den Boogaart et al., 1982). A series of reports in 1983 confirmed widespread homology of DNA segments in the mitochondrial and nuclear genome (Farrelly and Butow, 1983; Gellissen et al., 1983; Wright et al., 1983; Jacob et al., 1983; Kemble et al., 1983; Hadler et al., 1983; Tsuzuki et al., 1983). Following (Lopez et al., 1994), we use the term "nuclear mitochondrial DNA segments/sequence" or NUMT, to refer the inserted piece of mitochondrial (MT) DNA into the nuclear DNA, without requiring any amplification event after the nuclear insertion (Lopez et al., 1994).

The presence of NUMTs in the human genome (Herrnstadt et al., 1999; Mourier et al., 2001; Ricchetti et al., 2004; Hazkani-Covo et al., 2010) is of interest because it is a type of repetitive sequence, which can contribute to the redundancy problem in sequencing, mapping, and genotyping (e.g., (Weber and Myers, 1997; Green, 1997; Derrien et al., 2012; Lee and Schatz, 2012; Li et al., 2014; Li and Fredenberg, 2014)). Variants within NUMTs are easily mistaken as MT variants in genetic disease mapping studies (Wallace et al., 1997; Parr et al., 2006; Yao et al., 2008). NUMTs may also be viewed analogously to structural changes of the nuclear genome called segmental duplication (ancestral), or copy number variation (CNV) (germline, recent and within species), or copy number alteration (CNA) (somatic, such as cancer cells): all dynamic processes though with different time scales. Accordingly, more recent MT insertions can either serve as genetic marker (Thomas et al., 1996; Caro et al., 2010; Lang et al., 2012; Dayama et al., 2014), or be functional (Willett-Brozick et al., 2001; Turner et al., 2003; Goldin et al., 2004; Schon et al., 2012). Somatic MT insertions in cancer cells can offer novel insight on genomic instability in cancer (Srinivasainagendra et al., 2017). The latter two

types might be called de novo NUMT and somatic NUMT, respectively. Furthermore, NUMTs also provide an opportunity to study the evolutionary history of genomes (e.g., (Zischler et al., 1995; Perna and Kocher, 1996; Mishmar et al., 2004; Gunbin et al., 2017)).

There have been several attempts to catalog all human NUMTs. After the initial human genome sequence became available, 296 NUMTs were identified through a combination of alignment (BLAST with the default setting) and co-linearity of conserved blocks, of sizes between 106 and 14kb (Mourier et al., 2001). Comparison of various early attempts of compiling human NUMTs was made in (Lascaro et al., 2008), with 190 consistent entries. The current standard list of human NUMT is based on two studies with very similar number of hits: 755 NUMTs in (Ramos et al., 2011) and 766 NUMTs in (Simone et al., 2011), with an update to 585 NUMTs in (Calabrese et al., 2012).

This list of computationally predicted (*in silico*) NUMTs is based on sequence alignment, in particular the BLAST alignment (Altschul et al., 1990). The BLAST algorithm first finds exactly matching k-mers (l-tuples, n-grams), with k usually small, in the two sequences. The matching k-mer is then extended to maximum length, which is identical in the two sequences. The series of exactly matching segments in the observed order, with possible gaps between segments, is then returned as a BLAST alignment. The biological assumption behind the design of this algorithm is that the divergence between two sequences is exclusively caused by small insertions, deletions, and point mutations. If the mutational dynamics of the sequences have been more complicated, alignment-based approaches like BLAST will have difficulty to detect a common ancestry of two sequences (greater dissimilarity due to greater evolutionary distance is another situation where alignment-based methods may fail).

To overcome these challenges, alignment-free methods have been proposed (Blaisdell, 1986; Vinga and Almeida, 2003), as these do not consider the ordering of the matching k-mers in two sequences. The justification of this alignment-free approach is that if two similar sequences diverge not only due to mutations, insertions, deletions, but also through a more complicated reshuffling, alignment-based methods would fail to detect the remote homology. Another potential situation where alignment-based methods fail are regions that are hot-spots of insertion of short DNA segments, where the insertion occurs at different time and at random insertion points. In these instances, even if the ancestry of the DNA segments is the same,

the lack of ordering of the inserted pieces would prevent an alignment-based to detect the homology.

In this paper, we ask the question whether large BLAST-alignment-detected NUMTs are equivalently detected by alignment-free methods, i.e., can be recognized by the similarity of k-mer/l-tuple distributions. Previously, it was found that repetitive sequences account for one-half of the human genome with the alignment-based approaches, whereas repetitive sequences account for two-third by alignment-free approaches (Gu et al., 2008; De Koning et al., 2011). Therefore, we may expect the possibility of certain difference between the k-mer-based and BLAST-alignment-based detection of NUMTs. The paper is organized as follows: we first describe our k-mer analysis design; we then establish that k-mer distributions between MT and nuclear genome are indeed distinguishable; k-mer based NUMTs detection is then carried out for k=6,7, and for k=4,5; the results are discussed in the context of our current knowledge about NUMTs as well as the broader biological context.

Results

Issues in designing a k-mer based NUMT detection method

We use a moving window to scan the nuclear genome sequence for regions with a similar k-mer frequency distribution as the mitochondrial genome sequence. There are several decisions to be made to have a practical detection approach:

1. **Window size and moving distance:** The choice of window size is closely related to the size of region that is similar to MT. We choose window size of 3kb, and moving distance of 1/8 the window size, targeting NUMTs in the range of 3kb size. In order to test robustness of the result, we also rerun the calculation at peak locations with a moving distance of 1 base. The size of the human mitochondrial genome is 16569 bp, and thus expect to detect NUMTs that comprise roughly one fifth of its size.

2. **Measure of similarity between two k-mer frequency distributions:** This topic is extensively covered in (Vinga and Almeida, 2003). We use the Jensen-Shannon (JS) divergence (Burbea and Rao, 1982; Rao, 1982; Lin, 1991) based on an application of

Jensen's inequality for convex/concave functions. We denote $\{m_i = (p_i + q_i)/2\}$ for the high-dimensional mid-point between the two distributions, and JS divergence is defined as:

$$JS = \left(\sum_{i=1}^{4^k} p_i \log_2 \frac{p_i}{m_i} + \sum_{i=1}^{4^k} q_i \log_2 \frac{q_i}{m_i} \right) /2. \quad (1)$$

JS divergence is a symmetric version of the Kullback-Leibler distance (relative entropy). We previously used JS divergence to segment DNA sequences into relatively homogeneous halves (Bernaola-Galvan et al., 1996; Grosse et al., 2002; Li, 2001; Li et al., 2002) The alignment-free similarity signal is measured by the reciprocal of JS divergence: 1/JS. Other measures have been tried, such as the Euclidean distance, but JS divergence proved to be a choice that was more consistent with alignment-based approaches.

3. **The value(s) of k**: We have used the following k values: 3,4,5,6,7. Several procedures for choosing the value of k are used. The first is to find the k value that maximizes the number of more-than-once k-mer types (Sims et al., 2009). The idea behind this approach is that if a sequence is constructed with each k-mer type occurring only once (Fraenkel and Gillis, 1966) (a Hamiltonian cycle on the De Bruijn graph where each k-mer is represented by a node (e.g., (Pevzner et al., 2001))), no k-mer type could claim to represent the sequence. Only k-mer types which appear more than once have a chance to be part of the signature, and it is desirable to maximize the set of k-mer types. Fig.1(A) shows that for MT sequence, the number of k-mer types (with one or more counts) peaks at k=7.

   Another strategy for choosing the right value for k is based on whether a k-mer type frequency is well predicted by all (k-1)-mer and (k-2)-mer frequencies (Sims et al., 2009). If the predicted and observed k-mer type frequencies are close (measured by the Kullback-Leibler (KL) divergence), there is not much new information gained by increasing the oligo length to k. Fig.1(B) shows that the KL divergence is highest around k=7-8. (and reaches zero at k=14).

   Yet another consideration for choosing k is provided by the formula: k=$\log_4(N) + 1$ (Price et al., 2005; Campagna et al., 2004; Gu et al., 2008), where N is the size of

the region where k-mer frequency is counted, which can be the whole genome length, one chromosome length, or window length. This number is very similar to the formula k=$\log_4(2N) = \log_4(N) + 0.5$ when each k-mer type appears only once in the genome (Li et al., 2014). Based on these two formulas, for a 3kb window, the optimal choice of k is 6.28-6.78, and for 12000 base MT sequence, the choice of k is 7.28-7.78. Even longer k-mers are considered in (Wang et al., 2016), serving different purposes (and not discussed here).

4. **Not using strand symmetry to combine direct and reverse-complement k-mers**: Strand symmetry (or Chargaff's second parity rule) (Elson and Chargaff, 1952; Li, 1997; Forsdyke, 2016) is true for k-mers (Prabhu, 1993) for most long genomic sequences. If the strand symmetry holds, a k-mer (e.g. AACGT) has almost identical frequency as its reverse complement (e.g. ACGTT), then these two k-mers might be grouped into one unit. However, mitochondrial sequence is short, and many organellar sequences do violate strand symmetry (Nikolaou and Almirantis, 2006). The strand asymmetry in mitochondria is the basis for the distinction between heavy and light strands (e.g. (Reich and Luck, 1966)). Consequently, we use the frequency of all 4^k k-mer types, not $4^k/2$, in characterizing the MT sequence. However, a genomic sequence is scanned twice, once for the direct strand and another for the reverse complement strand.

5. **Threshold in an alignment-free peak call:** When 1/JS is high enough (or JS low enough), we make a NUMT prediction call. The threshold of the peak call was based on the known NUMTs that are longer than 3kb (same as the window size). The 1/JS values of all these large NUMT loci are sorted, and the second lowest 1/JS value is used as the threshold. Since there are 38 NUMTs in the most recent collection (Calabrese et al., 2012) that are larger than 3kb, missing 2 peak calls correspond to a false negative rate of 5%.

Note that we did not use other public domain, specially-designed, computationally-efficient k-mer count programs, because of the small k values chosen in this paper allowed for less optimal implementations. Many space/time efficient programs have been developed for dealing with much larger k-values (e.g. $k > 20$), and one can refer to these publications: (Kurtz et al.,

2008; Marcais and Kingsford , 2011; Melsted and Pritchard, 2011; Rizk et al., 2013; Deorowicz et al., 2013; Zhang et al., 2014; Roy et al., 2014; Patro et al., 2014; Audano and Vannberg, 2014; Melsted and Halldorsson, 2014; Deorowicz et al., 2015; Mamun et al., 2016; Sivadasan et al., 2016; Bray et al., 2016; Pandey et al., 2017; Marchet et al., 2017; Ebert et al., 2017; Kokot et al., 2017; Patro et al., 2017).

Nuclear DNA and mitochondrial DNA have overall different k-mer distributions

Different k-mer distributions of the nuclear genome sequence the MT sequence are already expected based on the difference in GC-content: the GC-content of human MT sequence is 55.6%, as compared to the GC-content in the nuclear genome of around 40% (Li, 2013). To further illustrate that the k-mer frequency of the nuclear and MT genome are different, we use a 3kb window to scan 5-mer frequencies in the first 10Mb region of human chromosome 1 (non-overlapping windows).

Fig.2(A) shows points representing the windows in the first two dimensions of multidimensional scaling (MDS). The MT is not the most distinct sequence in comparison to the genomic sequence: a low-complexity region (chr1:2.652Mb-2.772Mb, hg38) is the most obvious outlier. Windows from MT genome are near the outskirt of the main band. When windows with mostly (more than 99%) unique (non-repetitive) sequences and those with mostly (more than 99%) repetitive sequences are marked (red and green, respectively), MT windows are closer to repetitive sequences. Fig.2(B) shows the second and third dimension of MDS, and again MT windows are on the border of the main band.

To see whether the MDS dimensions (1,2,3) represents simple features of the sequence, we calculate the following parameters: percentage of unique (non-repetitive) sequence, GC-content, Jensen-Shannon divergence between the k-mer frequency and its reverse-complements. The latter is a measure of the strand asymmetry at the k-mer level: zero means strand symmetry, and large value for violation of strand symmetry. Table 1 shows that MDS dimension-1 closely follows the GC-content (Spearman correlation coefficient (cc) $= -0.98$, p-value $< 2 \times 10^{-16}$). The negative sign doesn't have any particular meaning, as the direction of

the MDS axis is arbitrary. MDS1 is also highly correlated with proportion of unique sequences (Spearman cc $= -0.66$, p-value $< 2 \times 10^{-16}$) MDS1 is not only correlated with these two quantities again (though with lesser strength), but also weakly correlated with the strand asymmetry (Spearman cc$= -0.1$, p-value $= 2 \times 10^{-8}$). MDS3 is weakly correlated with GC-content (Speaman cc$= 0.055$, p-value $= 0.002$).

The classic/metric MDS is a linear projection from high to low dimensional space. A newer dimension reduction and plotting technique, tSNE (t-distributed stochastic neighbor embedding), is a nonlinear projection (Van Der Maaten and Hinton, 2008). Previously, we applied tSNE to human genetic data (Li et al., 2017), and found it quite powerful in displaying both continental and sub-continental distributions of genetic variation. Fig.1(C)(D) shows the tSNE for 1st vs 2nd dimension, and 2nd vs 3rd dimension. Unlikely MDS, tSNE is able to display the substructure within the main band, even with the presence of the outliers (the low-complexity sequence). It is clear that MT points for a cluster separated from the rest of the genomics sequences in chromosome 1. A close inspection of the plots shows three chr1 windows within the MT cluster, which should be predicted as being NUMTs.

Manhattan plot of alignment-free peaks for NUMTs at k=6 and k=7

A common technique in genome-wide association studies (GWAS) of human genetically complex traits is the so called Manhattan plots, which plot genetic variant association signals (often measured by the $-\log$(p-value)) as a function of chromosome location across all chromosomes (e.g. (Wellcome Trust Case Control Consortium, 2007)). Here we introduce this idea to plotting the alignment-free signal of mitochondrial DNA insertion into nuclear genome. The signal strength used is the reciprocal of JS divergence, 1/JS, between the two k-mer distributions, with the target being the overlapping window of size 3kb along the chromosome (and its reverse complement), and with the query being the mitochondrial sequence (and its reverse complement).

Fig.3 shows the Manhattan plot for k=6 (only windows with high enough signal are included). First of all, regions with the highest peak heights are easily identified (chr1, chr5, chr17), indicating a high level of similarity with the MT sequence. Second, all peaks (deter-

mined by 1/JS > 3.346) can be explained by the known NUMTs in (Calabrese et al., 2012), with sizes larger than 3kb (pink), 2kb (blue), or 1kb (green). In other words, at k=6, alignment-free approach does not reveal any new potential NUMTs which were not already identified by the alignment-based BLAST method.

We note that some peaks are very close to either telomere (chr1) or centromere (chr17). It has been observed that somatic large-scale structural changes in cancer cells can be grouped into telomere/centromere-bounded and not-telomere/centromere-bounded categories (Zack et al., 2013). The two categories affect different sizes of chromosome regions, and may be caused by different mechanisms. In Fig.3, two out of the four highest peaks are close to telomere or centromere.

Our observations that for k=6 alignment-free peaks are fully explained by alignment-based NUMTs at 3kb or larger size, as well as some NUMTs with 1kb-3kb sizes, is robust to our choice of peak calling. Initially, the peak called is based on the second lowest 1/JS value for the 38 NUMTs larger than 3k, when 3kb window moves 375 bases at the time (that 1/JS is 3.305). We do not come to different conclusions, when we move up the threshold for peak calling at 1/JS > 3.346, when the window moving step is changed from 375bp to 1 base. To test the robustness, we also lowered the threshold for peak calling by the lowest (instead of the second lowest) 1/JS value for the 38 NUMTs larger than 3kb (1/JS > 3.2945): again, all peaks are explained by known NUMTs 1kb or larger.

Similar conclusion are also reached for k=7 (plot not shown). If we use 1/JS > 7.7336 as the peak calling criterion (which is the second lowest 1/JS value for +3kb NUMTs with moving step of 1 base), all peaks can be explained by BLAST-based NUMTs 1kb or larger. When the peak calling is less stringent, at 1/JS > 1.711 (lowest 1/JS value for +3kb NUMTs with moving step of 1 base), 1.6845 (second lowest 1/JS value with moving step of 375 bases), 1.659 (lowest 1/JS value with moving step of 375 bases), we either observe the same conclusion or only a few isolated probably false-positives.

New alignment-free peaks at k=5 and k=4

New alignment-free peaks start to appear when k is reduced to 5 and below. This is shown

in Fig.4 which has 9 loci with peaks (1/JS > 8.666, which is the second lowest peaks for known NUMTs larger than 3kb, with moving distance of 1 base) that do not have an underlying known NUMTs of 1kb or longer. Two of the loci (chr3:89.587Mb, chrX:126.471Mb) contain known NUMTs of length shorter than 1kb. The rest of the 7 loci are listed in Table 2.

The new peak near chr1:9.093Mb is the highest and widest one among those in Table 2. We plot the 1/JS in the region separately in Fig.5(A). The 3kb window size is marked for each point as a reminder of the length scale. If the peak height is about 9, and the genome-wide average is 5, the half-height points can be used to measure the peak width. This leads to the peak region from 9087.75kb to 9096.75kb, or, of width of 9kb. Other peaks listed in Table 2 have narrower widths, as well as lower heights.

At k=4, the relative peak heights of +3kb NUMTs span a much bigger range. If we still use the second lowest peak value of the 38 +3kb NUMTs, many new loci will pass the peak call. To be more conservative, we use the 4th lowest value among the +3kb NUMTs, which is 1/JS = 22.89, as the threshold. With this criterion, 18 loci are called: 5 of them already appear for k=5; 1 with known underlying NUMTs of less than 1kb size. This leads to 12 new alignment-free signals at k=4 (see Table 2).

We arbitrarily picked one of the new alignment-free peaks for k=4 at chr3:141.699Mb to be examined in more detail (Fig.5(B)). If we consider the peak height as 1/JS=24, baseline as 1/JS=10, the half-height value is 1/JS=17. Using this half-height, the width of the peak is roughly 7.1kb (extending both ends of the 3kb window).

The alignment-free similarity signal for k=5/chr1 and k=4/chr3 are further analyzed by the pairwise BLAST program comparisons to the mitochondrial genome. As expected, there are no hit if megablast (highly similar) and discontigous megablast (more dissimilar) options are used. However, blastn (somewhat similar sequences at least seven bases long) produce 9 (8) matches between the chr1 (chr3) peak and MT sequence.

Annotation tracks of the new alignment-free peaks

We finally used the Genome Browser to examine all +3kb BLAST-based NUMTs and new alignment-free peak regions listed in Table 2. Both the BLAST-based large NUMTs and

new alignment-free peak regions overlap with genes: RP5-857K21.4 (chr1, HSA_NumtS_001), MTRNR2L11 (chr1, HSA_NumtS_043), ARHGAP15 (chr2,HSA_NumtS_090), JAK2 (chr9, HSA_NumtS_329), DNAJC3 (chr13, HSA_NumtS_470), AQP9 (chr15, HSA_NumtS_490), RAE1 (chr20, HSA_NumtS_547), SMARCA2 (chrX, HSA_NumtS_565), and ZNF678 (chr1), BIRC6 (chr2), CPA4 (chr7), MOK (chr14), AC002310.13 (chr16), LOC100294362/CTD-2047H16.4 (chr17) (Table 2). There is no indication of a difference in the tendency to co-localize with genes in the two groups.

Multiple large BLAST-based NUMTs overlap with DNase hypersensitive site clusters, inlcuding HSA_NumtS_001, 042, 043 (on chr1), 219, 222, 228 (on chr5), 508 (on chr17). For our alignment-free peak regions, only one (on chr2) overlaps with DNase hypersensitive site clusters. There appear to be isolated appearance of SINE and LINE, but these are not prominent in these regions.

The biggest annotation difference between the BLAST-based and alignment-free peaks is the co-localization with Long Terminal Repeat (LTR). All alignment-free peak regions in Table 2, with one exception, cover a LTR. On the other hand, only one BLAST-based NUMTs (HSA_NumtS_565, chrX) covers a LTR. The difference is statistically very significant (χ^2-test $p-$value - 3×10^{-11}) indicating that biological mechanism of NUMT predictions as derived from the two types of signals might be different.

Discussion

We use a novel alignment-free strategy for detecting NUMTs through sequence similarity, which points to several genome regions that may have yet unrecognized ancestry in the mitochondrial genome. The principal validity of our alignment-free approach is supported by the fact that all of our detected NUMTs with parameters at k=6 and k=7 (and window size 3kb) have also been identified with earlier alignment-based methods. This confirms a previous conclusion that alignment-free methods could be as powerful as (or even outperform) Smith-Waterman alignment (Zielezinski et al., 2017) for the detection of homologous sequences. Also, this result indicates that potential MT insertions followed by complicated dynamics may only be detected in a certain parameter range (e.g., lower k's). The principal question is very sim-

ilar to the situation of detecting repetitive elements in the human genome (De Koning et al., 2011), where the result depends on parameter values chosen, such as proportion of bases in a window that match signature k-mers (k=14-15 in that application).

The motivation of using alignment-free methods in detecting homologous sequences has been well summarized in a recent review (Zielezinski et al., 2017). Our approach mainly aims at the problem (case no.1 in (Zielezinski et al., 2017)) that alignment-based approaches may miss MT insertions, if the insertion event was followed by a more complicated evolutionary dynamics. Although we follow the basic playbook in k-mer distribution approach (Fig.1 in (Zielezinski et al., 2017)), there are some subtle differences: we compare the k-mer distribution from one strand of MT with that from the nuclear DNA of the standard strand, as well as that of the reverse complement strand; we use the JS distance instead of Euclidean distance, because we found that the former leads to a more consistent result with the alignment-based method.

We have chosen the peak calling criterion carefully, by using the peak heights from the known (i.e., alignment-based) NUMTs and allowing for a 5% false negative rate, in order to avoid high false positive rates. Although we do not have a theoretical foundation for our choices (which are mostly made by trial-and-error), there are multiple reasons to believe that chance events do not cause these k-mer (at k=5 or 4) similarity peaks, that are not listed among the previously known NUMTs. First of all, when the sequence is reversed (but not complementary), 1/JS values in these peak regions are much reduced. This is another way to say that an insertion followed by an inversion is very rare at the 3kb range or longer. Secondly, when k=3, all peaks observed at k=5 or k=4 spanning more than one window (see Table 2) are still peaks. This consistency for reduced k indicates the robustness of the result.

A previous application of alignment-free methods to mitochondrial genome has been to distinguish different MT haplogroups (with close to 5000 of them existing) (Navarro-Gomez et al., 2014). This is a very different goal from our analysis: all MT haplogroups have similar k-mer distributions when k is low (e.g., k=4-7), so in order to identify one haplogroup among many, the signature k-mers have to be unique to that haplogroup, which can only be achieved with a large k. Accordingly, the choice in (Navarro-Gomez et al., 2014) was to use k=12.

Alignment-free approaches further have the advantage of greater computational speed (Zielezinski et al., 2017), which makes it easy to use our approach to scan NUMTs in the genomes of

other species. We easily confirmed the 270kb MT insertion in chromosome 2 of plant Arabidopsis thaliana (plot not shown) (Lin et al., 1999) (another experiment found the insertion size is 620kb (Stupar et al., 2001)). A similar check for chloroplastic DNA showed none of such peaks (data not shown).

To visualize the regions of nuclear DNA that are predicted NUMTs, we use the so-called Manhattan plot. Manhattan plots are standard tools in genome-wide association studies (GWAS) for complex disease, but have been used less in studies that evaluate the evolutionary history and functional content of the human genome. Here we plot the reciprocal of JS over chromosome positions, such that peaks indicate nuclear regions that are similar to the mitochondrial sequences. On the other hand, when JS is plotted instead of 1/JS, the peaks usually indicate centromeres, telomeres, and other low-complexity regions (plot not shown).

In conclusion we find that current NUMTs predictions are equally well detected by for certain parameter range (e.g. k¿5), but additional NUMTs are detected with smaller choices of k (k=5 or k=4). The precise nature of these novel NUMT predictions remains to be determined. Mutually non-exclusive possibilities would be a even earlier ancestral origin of these novel predictions, or a more complex mutational history. We note that most of these novel predictions overlap with LTR annotations in the human genome, which may indicate interactions of NUMTs with LTRs, or greater sequence similarity of parts of the mitochondrial genome with LTRs. Future studies may also look into the parts of the mitochondrial genome where these NUMTs did originate, or consider the hypothesis that certain nuclear sequences have been transferred into the mitochondrial genome over the course of more than a billion years of co-evolution of the two genome.

Data and Methods

The list of previously established NUMTs is obtained from the supplementary material of (Calabrese et al., 2012): https://static-content.springer.com/esm/art%3A10.1186%2F1471-2105-13-S4-S15/ MediaObjects/12859_2012_5112_MOESM1_ESM.xls The chromosome positions in this list are in hg19, and we applied the liftOver program (*https://genome.ucsc.edu/cgi-*

bin/hgLiftOver) to convert the NUMT coordinates to hg38.

The k-mer count programs are implemented as custom scripts in Python and Perl by the authors. The plot and analysis was carried out in R (*https://www.r-project.org/*), including the classic multi-dimensional scaling cmdscale() and t-distributed stochastic neighboring embedding Rtsne() (*https://github.com/jkrijthe/Rtsne*).

We used the BLAST web interface at *https://blast.ncbi.nlm.nih.gov/Blast.cgi* to obtain pairwise alignments, choosing "Nucleotide BLAST" $\rightarrow$ "Align two or more sequences" $\rightarrow$ (parameter setting) "Somewhat similar sequences" (blastn).

We used the UCSC Genome Browser *https://genome.ucsc.edu* to obtain annotation at queried chromosomal regions, choosing "(Genomes) Human GRCh38/hg38". Both mitochondrial and nuclear DNA sequences were downloaded from *http://hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/* (chr1.fa.gz, chr2.fa.gz, $\cdots$ chrM.fa.gz).

Authors' contributions

WL designed the overall study, performed the final analysis, and wrote the paper; JF designed and performed the initial analysis, and wrote an initial analysis report.

Acknowledgment

JF would like to thank the financial and other support from the Summer Internship Program of the Feinstein Institute for Medical Research. We thank Jan Freudenberg, Susana Marquez Renteria, Andrew Shih, Ilya Korsunsky, Jane Cerise, Daniel Miller, Dimitris Thanos, Astero Provata, Pedro Miramontes, and Pedro Bernaola for discussions.

References

SF Altschul, W Gish, W Miller, EW Myers, DJ Lipman (1990). Basic local alignment search tool, *J. Mol. Biol.*, 215:403-410.

P Audano and F Vannberg (2014), KAnalyze: a fast versatile pipelined K-mer toolkit, *Bioinformatics*, 30:2070-2072.

P Bernaola-Galván, R Ramán-Roldán, JL Oliver (1996), Compositional segmentation and long-range fractal correlations in DNA sequences, *Phys. Rev. E*, 53:5181-5189.

BE Blaisdell (1986), A measure of the similarity of sets of sequences not requiring sequence alignment, *Proc. Natl. Acad. Sci.*, 83:5155-5159.

NL Bray, H Pimentel P Melsted, L Pachter (2016), Near-optimal probabilistic RNA-seq quantification, *Nature Biotech.*, 34:525-527.

J Burbea, CR Rao (1982), On the convexity of some divergence measures based on entropy functions, *IEEE Trans. Info. Theory*, 28:489-495.

FM Calabrese, D Simone, M Attimonelli (2012), Primates and mouse NumtS in the UCSC Genome Browser, *BMC Bioinformatics*, 13(Suppl. 4):S15.

D Campagna, C Romualdi, N Vitulo, M del Favero, M Lexa, N Cannata, G Valle (2005), RAP: a new computer program for de novo identification of repeated sequences in whole genomes, *Bioinformatics*, 21:582-588.

P Caro, J Gomez, A Arduini, M Gonzalez-Sanchez, M Gonzalez-Garcia, C Borras, J Vina, MJ Puertas, J Sastre, G Barja (2010), Mitochondrial DNA sequences are present inside nuclear DNA in rat tissues and increase with age, *Mitochondrion*, 10:479-486.

G Dayama, SB Emery, JM Kidd, RE Mills (2014), The genomic landscape of polymorphic human nuclear mitochondrial insertions, *Nucl. Acids Res.*, 42:12640-12649.

APJ De Koning, W Gu, TA Castoe, MA Batzer, DD Pollock (2011), Repetitive elements may comprise over two-thirds of the human genome, *PLoS Genet.*, 7:e1002384.

S Deorowicz, A Debudaj-Grabysz, S Grabowski (2013), Disk-based k-mer counting on a PC, *BMC Bioinf.*, 14:160.

S Deorowicz, M Kokot, S Grabowski, A Debudaj-Grabysz (2015), KMC 2: fast and resource-frugal k-mer counting, *Bioinformatics*, 31:1569-1576.

T Derrien, J Estelle, SM Sola, DG Knowles, E Raineri, R Guigo, P Ribeca (2012), Fast computation and applications of genome mappability, *PLoS ONE*,7:e30377.

HG Du Buy and FL Riley (1967), Hybridization between the nuclear and kinetoplast DNA's of Leishmania Enriettii and between nuclear and mitochondrial DNA's of mouse liver, *Proc. Natl. Acad. Sci.*, 57:790-797.

D Elson and E Chargaff (1952), On the deoxyribonucleic acid content of sea urchin gametes, *Experientia*, 8:143145.

M Erbert, S Rechner, M Muller-Hannemann (2017), Gerbil: a fast and memory-efficient k-mer counter with GPU-support, *Algorithms for Mol. Biol*, 12:9.

F Farrelly, RA Butow (1983), Rearranged mitochondrial genes in the yeast nuclear genome, *Nature*, 301:296-301.

DR Forsdyke (2016), *Evolutionary Bioinformatics* (Springer).

AS Fraenkel and J Gillis (1966), Appendix II. Proof that sequences of A, C, G, and T can be assembled to produce chains of ultimate length avoiding repetitions everywhere, *Prog. Nucl. Acids Res. and Mol. Biol.*, 5:343-348.

G Gellissen, JY Bradfield, BN White, GR Wyatt (1983), Mitochondrial DNA sequences in the nuclear genome of a locust, *Nature*, 301:631-634.

E Goldin, S Stahl, AM Cooney, CR Kaneski, S Gupta, RO Brady, JR Ellis, R Schiffmann (2004), Transfer of a mitochondrial DNA fragment to MCOLN1 causes an inherited case of mucolipidosis IV, *Hum. Mut.*, 24:460-465.

P Green (1997), Against a whole-genomeshotgun, *Genome Res.*, 7:410-417.

I Grosse, P Bernaola-Galván, P Carpena, R Román-Roldán, J Oliver, HE Stanley (2001), Analysis of symbolic sequences using the Jensen-Shannon divergence, *Phys. Rev. E*, 65:041905.

W Gu, TA Castoe, DJ Hedges, MA Batzer, DD Pollock (2008), Identification of repeat structure in large genomes using repeat probability clouds, *Analytical Biochem.*, 380:77-83.

K Gunbin, L Peshkin, K Popadin, S Annis, RR Ackermann, K Khrapko (2017), Integration of mtDNA pseudogenes into the nuclear genome coincides with speciation of the human genus. A hypothesis, *Mitochondrion*, 34:20-23.

HI Hadler, B Dimitrijevic, R Mahalingam (1983), Mitochondrial DNA and nuclear DNA from normal rat liver have a common sequence, *Proc. Natl. Acad. Sci.*, 80:6495-6499.

E Hazkani-Covo, RM Zeller, W Martin (2010), Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes,

C Herrnstadt, W Clevenger, SS Ghosh, C Anderson, E Fahy, S Miller, N Howell, RE Davis (1999) A novel mitochondrial DNA-like sequence in the human nuclear genome, *Genomics*, 60:67-77.

HT Jacobs, JW Posakony, JW Grula, JW Roberts, JH Xin, RJ Britten (1983), Mitochondrial DNA sequences in the nuclear genome of Strongylocentrotus pupuratus, *J. Mol. Evol.*, 165:609-632.

RJ Kemble, RJ Mans, S Gabay-Laughnan, JR Laughnan (1983), Sequences homologous to episomal mitochondrial DNAs in the maize nuclear genome, *Nature*, 304:744-747.

M Kokot, M Dlugosz, S Deorowicz (2017), KMC 3: counting and manipulating k-mer statistics, *Bioinformatics*, 33:2759-2761.

S Kurtz, A Narechania, JC Stein, D Ware (2008), A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes, *BMC Genomics*, 9:517.

M Lang, M Sazzini, FM Calabrese, D Simone, A Boattini, G Romeo, D Luiselli, M Attimonelli, G Gasparre (2012), Polymorphic NumtS trace human population relationships, *Hum. Genet.*, 131:757-771.

D Lascaro, S Castellana, G Gasparre, G Romeo, C Saccone, M Attimonelli (2008), The RHNumtS compilation: Features and bioinformatics approaches to locate and quantify Human NumtS, *BMC Genomics*, 9:267.

H Lee, MC Schatz (2012), Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score, *Bioinformatics*, 28:2097-2105.

W Li (1997), Study of correlation structure in DNA sequences: a critical review, *Comp. & Chem.* , 21:257-272.

W Li (2001), New stopping criteria for segmenting DNA sequences, *Phys. Rev. Lett.*, 86:5815-5818.

W Li (2013), G+C content evolution in the human genome, *eLS*, DOI: 10.1002/9780470015902.a0021751.

W Li, P Bernaola-Galván, F Haghighi, I Grosse (2002), Applications of recursive segmentation to the analysis of DNA sequences, *Comp. and Chem.*, 26:491-510.

W Li, JE Cerise, Y Yang, H Han (2017), Application of t-SNE to human genetic data, *J. Bioinf. Comp. Biol.*, 15:1750017.

W Li and J Freudenberg (2014), Characterizing regions in the human genome unmappable by next-generation-sequencing at reads length of 1000 bases, *Comp. Biol. Chem.*, 53:108-117.

W Li, J Freudenberg, P Miramontes (2014), Diminishing return for increased mappability with longer sequencing reads: implications of the k-mer distributions in the human genome, *BMC Bioinfo.*, 15:2.

J Lin (1991), Divergence measures based on the Shannon entropy, *IEEE Trans. Info. Theo.*, 37:145-151.

X Lin, S Kaul, S Rounsley, TP Shea, MI Benito, CD Town, CY Fujii, T Mason, CL Bowman, M Barnstead, TV Feldblyum, CR Buell, KA Ketchum, J Lee, CM Ronning, HL Koo, KS Moffat, LA Cronin, M Shen, G Pai, S Van Aken, L Umayam, LJ Tallon, JE Gill, MD Adams, AJ Carrera, TH Creasy, HM Goodman, CR Somerville, GP Copenhaver, D Preuss, WC Nierman, O White, JA Eisen, SL Salzberg, CM Fraser, JC Venter (1999), Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana, *Nature*, 402:761-768.

JV Lopez, N Yuhki, R Masuda, W Modi, SJ O'Brien (1994), Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat, *J. Mol. Evol.*, 39:174-190.

AA Mamun, S Pal, S Rajasekaran (2016), KCMBT: a k-mer Counter based on Multiple Burst Trees, *Bioinformatics*, 32:2783-2790.

G Marcais and C Kingsford (2011), A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, *Bioinformatics*, 27:764-770.

C Marchet, L Lecompte, A Limasset, L Bittner, P Peterlongo (2017), A resource-frugal probabilistic dictionary and applications in bioinformatics, *arXiv* preprint: arXiv:1703.00667 [cs.DS].

L Marguilis (1970), *Origin of Eukaryotic Cells* (Yale Univ. Press).

P Melsted and BV Halldorsson (2014), KmerStream: streaming algorithms for k -mer abundance estimation, *Bioinformatics*, 30:3541-3547.

P Melsted and JK Pritchard (2011), Efficient counting of k-mers in DNA sequences using a bloom filter, *BMC Bioinf.*, 12:333.

D Mishmar, E Ruiz-Pesini, M Brandon, DC Wallace (2004), Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration, *Hum. Mut.*, 23:125-133.

T Mourier, AJ Hansen, E Willerslev, P Arctander (2001), The human genome project reveals a continuous transfer of large mitochondrial fragments to the nucleus, *Mol. Biol. and Evo.*, 18:1833-1837.

D Navarro-Gomez, J Leipzig, L Shen, M Lott, APM Stassen, DA Wallace, JL Wiggs, MJ Falk, M van Oven, X Gai (2014), Phy-Mer: a novel alignment-free and reference-indepdent mitochondrial haplogroup classifier, *Bioinformatics*, 31:1310-1312.

C Nikolaou and Y Almirantis (2006), Deviations from Chargaff's second parity rule in organellar DNA Insights into the evolution of organellar genomes, *Gene*, 381:34-41.

P Pandey, MA Bender, R Johnson, R Patro (2017), Squeakr: An Exact and Approximate k-mer Counting System, *bioRxiv* preprint. DOI: doi.org/10.1101/122077

RL Parr, GD Dakubo, KA Crandall, J Maki, B Reguly, A Aguirre, R Wittock, K Robinson, JS Alexander, MA Birch-Machin, M Abdel-Malak, MK Froberg, EP Diamandis, RE Thayer (2006), Somatic mitochondrial DNA mutations in prostate cancer and normal appearing adjacent glands in comparison to age-matched prostate samples without malignant histology, *J. Mol. Diagn.*, 8:312-319.

R Patro, G Duggal, MI Love, RA Irizarry, C Kingsford (2017), Salmon provides fast and bias-aware quantification of transcript expression, *Nature Meth.*, 14:417-419.

R Patro, SM Mount, C Kingsford (2014), Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms, *Nature Biotech.*, 32:462-464.

NT Perna, TD Kocher (1996), Mitochondrial DNA: Molecular fossils in the nucleus, *Current Biol.*, 6:128-129.

PA Pevzner, H Tang, MS Waterman (2001), An Eulerian path approach to DNA fragment assembly, *Proc. Natl. Acad. Sci.*, 98:9748 9753.

VV Prabhu (1993), Symmetry observations in long nucleotide sequences, *Nucl. Acids Res.*, 21:2797-2800.

AL Price, NC Jones, PA Pevzner (2005), De novo identification of repeat families in large genomes, *Bioinformatics*, 21:i351-i358.

A Ramos, E Barbena, L Mateiu, M Del Mar Gonzalez, Q Mairal, M Lima, R Montiel, MP Aluja, C Santos (2011), Nuclear insertions of mitochondrial origin: Database updating and usefulness in cancer studies, *Mitochondrion*, 11:946-953.

CR Rao (1982), Diversity and dissimilarity coefficients: a unified approach, *J. Theoret. Pop. Biol.*, 21:24-43.

E Reich, DJL Luck (1966), Replication and inheritance of mitochondrial DNA, *Biochemistry*, 55:1600-1608.

M Ricchetti, F Tekaia, B Dujon (2004), Continued colonization of the human genome by mitochondrial DNA, *PLoS Biol.*, 2:E273.

G Rizk, D Lavenier, R Chikhi (2013), Dsk: k-mer counting with very low memory usage, *Bioinformatics*, 29:652653.

RS Roy, D Bhattacharya, A Schliep (2014), Turtle: Identifying frequent k -mers with cache-efficient algorithms, *Bioinformatics*, 30:1950-1957.

EA Schon, S DiMauro, M Hirano (2012), Human mitochondrial DNA: roles of inherited and somatic mutations, *Nature Rev. Genet.*, 13:878-890.

D Simone, FM Calabrese, M Lang, G Gasparre, M Attimonelli (2011), The reference human nuclear mitochondrial sequences compilation validated and implemented on the UCSC genome browser, *BMC Genomics*, 12:517.

GE Sims, SR Jun, GA Wu, SH Kim (2009), Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolution, *Proc. Natl. Acad. Sci.*, 106:2677-2682.

N Sivadasan, R Srinivasan, K Goyal (2016), Kmerlight: fast and accurate k-mer abundance estimation, *arXiv* preprint: arXiv:1609.05626 [cs.DS].

V Srinivasainagendra, MW Sandel, B Singh, A Sundaresan, VP Mooga, P Bajpai, HK Tiwari, KK Singh (2017), Migration of mitochondrial DNA in the nuclear genome of colorectal adenocarcinoma, *Genome Med.*, 9:31.

RM Stupar, JW Lilly, CD Town, Z Cheng, S Kaul, CR Buell, J Jiang (2001), Complex mtDNA constitutes an approximate 620-kb insertion on Arabidopsis thaliana chromosome 2: Implication of potential sequencing errors caused by large-unit repeats, *Proc. Natl. Acad. Sci.*, 98:5099-5103.

R Thomas R, H Zischler, S Paabo, M Stoneking (1996), Novel mitochondrial DNA insertion polymorphism and its usefulness for human population studies, *Hum. Biol.*, 68:847-854.

T Tsuzuki, H Nomiyama, C Setoyama, S Maeda, K Shimada (1983) Presence of mitochondrial-DNA-like sequences in the human nuclear DNA, *Gene*, 25:223-229.

C Turner, C Killoran, NS Thomas, M Rosenberg, NA Chuzhanova, J Johnston, Y Kemel, DN Cooper, LG Biesecker (2003), Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer, *Hum. Genet.*, 112:303309.

P Van Deb Boogaart, J Samallo, E Agsteribbe (1982), Similar genes for a mitochondrial ATPase subunit in the nuclear and mitochondrial genomes of Neurospora crassa, *Nature*, 298:187-189.

LJP Van Der Maaten and GE Hinton (2008), Visualizing High-Dimensional Data Using t-SNE, *J. Machine Learning Res.*, 9:2579-2605.

S Vinga, J Almeida (2003), Alignment-free sequence comparison – a review, *Bioinformatics*, 19:513-523.

DC Wallace, C Stugard, D Murdock, T Schurr, MD Brown (1997), Ancient mtDNA sequences in the human nuclear genome: A potential source of errors in identifying pathogenicmutations, *Proc. Natl. Acad. Sci.*, 94:14900-14905.

Y Wang, X Lei, S Wang, Z Wang, N Song, F Zeng, T Chen (2016), Effect of k-tuple length on sample-comparison with high-throughput sequencing data, *Biochem. Biophy. Res. Comm.*, 469:1021-1027.

JL Weber, EW Myers (1997), Human whole-genome shotgun sequencing, *Genome Res.*, 7:401-409.

Wellcome Trust Case Control Consortium (2007), Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, *Nature*, 447:661-678.

JE Willett-Brozick, SA Savul, LE Richey, BE Baysal (2001), Germ line insertion of mtDNA at the breakpoint junction of a reciprocal constitutional translocation, *Hum. Genet.*, 109:216223.

RM Wright, DJ Cummings (1983), Integration of mitochondrial gene sequences within the nuclear genome during senescence in a fungus, *Nature*, 302:86-88.

YG Yao, QP Kong, A Salas, HJ Bandelt (2008), Pseudomitochondrial genome haunts disease studies, *J. Med. Genet.*, 45:769-772.

TI Zack, SE Schumacher, SL Carter, AD Cherniack, G Saksena, B Tabak, MS Lawrence, CZ Zhang, J Wala, CH Mermel, C Sougnez, SB Gabriel, B Hernandez, H Shen, PW Laird, G Getz, M Meyerson, R Beroukhim (2013), Pan-cancer patterns of somatic copy number alteration, *Nature Genet.*, 45:1134-1140.

Q Zhang, J Pell, R Canino-Koning, AC Howe, CT Brown (2014), These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure, *PLoS ONE*,9:e101271.

A Zielezinski, S Vinga, J Almeida, WM Karlowski (2017), Alignment-free sequence comparison: benefits, applications, and tools, *Genome Biol.*, 18:186.

H Zischler, H Geisert, A von Haeseler, S Paabo (1995), A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans, *Nature*, 378:489-492.

	MDS1	MDS2	MDS3
% unique	-0.66 (pv=0)	0.32 (pv=0)	0.02 (pv=0.24)
GC	-0.98 (pv=0)	0.58 (pv=0)	0.055 (pv=0.002)
strand asym	0.04 (pv=0.02)	-0.098 (pv=2E-8)	0.017 (pv=0.3)

Table 1: Spearman's correlation coefficient between simple summary statistics of sequence windows (proportion of unique sequences, GC content, Jensen-Shannon divergence between 5-mer frequencies and those of its reverse complements) and the first three dimensions from the multi dimensional scaling (MDS) analysis across all windows.

k	chr	pos(Mb,hg38)	1/JS	nw	annotation
5,4	1	9.090375-9.094875	9.001, 24.938	5	LTR
5,4	1	227.635875-227.638875	8.767, 23.124	1	LTR, gene ZNF678
5,4	2	32.357625-32.360625	8.803, 25.392	1	gene BIRC6, DNase cluster
5,4	2	238.206375-238.209750	8.718, 23.702	2	LTR
5	7	29.789625-29.792625	8.849	1	LTR
5,4	7	130.314375-130.317375	8.986, 26.431	1	LTR, SINE, gene CPA4
5	X	64.846875-64.850250	8.682	2	LTR
4	3	141.696375-141.702250	24.198	6	LTR
4	4	145.835625-145.838625	23.346	1	LTR
4	6	17.319000-17.323125	23.942	2	(LTR)
4	8	47.988750-47.992500	23.869	3	LTR
4	11	122.828625-122.831625	23.214	1	LTR, (LINE)
4	13	54.939375-54.942375	23.938	1	LTR
4	13	83.908125-83.911125	23.268	1	LTR
4	14	102.244125-102.248625	26.094	5	LTR, gene MOK
4	16	30.541875-30.545250	24.845	2	LTR, gene AC002310.13
4	17	80.401500-80.404500	23.731	1	LTR, gene LOC100294362/CTD-2047H16.4
4	18	28.209375-28.212375	23.205	1	LTR
4	X	47.286375-47.289375	23.002	1	LTR

Table 2: List of NUMT predictions without a corresponding BLAST-based NUMT as called from similarity peaks for k=5 and 4. Column headers. k: the k value at which the peaks are detected; chr: chromosome; pos: chromosome position in Mb, hg38; 1/JS: similarity of k-mer distributions as measured by the reciprocal of Jensen-Shannon entropy; nw: number of 3kb windows that pass the threshold; annotation: other information obtained from the UCSC Genome Browser.

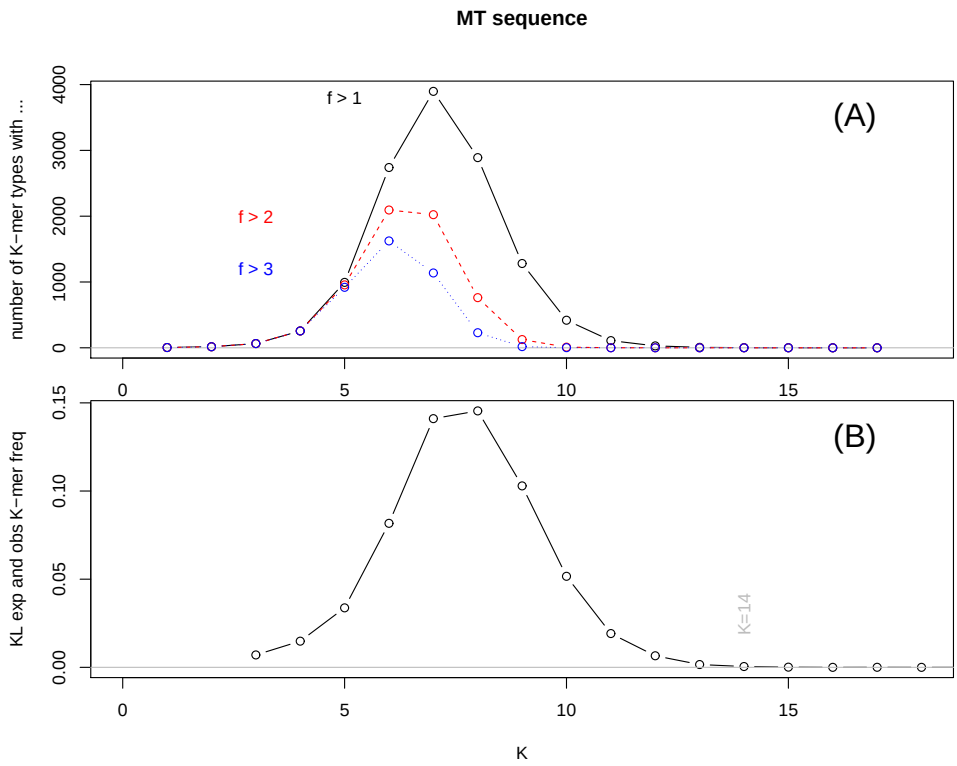

Figure 1: (A) The number of k-mer types in human MT sequence that appear at least once, as a function of k. The red and blue points denote the number of k-mer types that appear at least two and three times in the MT sequence. (B) The Kullback-Leibler divergence between the observed k-mer frequencies and those predicted by (k-1) and (k-2)-mer frequencies (Sims et al., 2009).

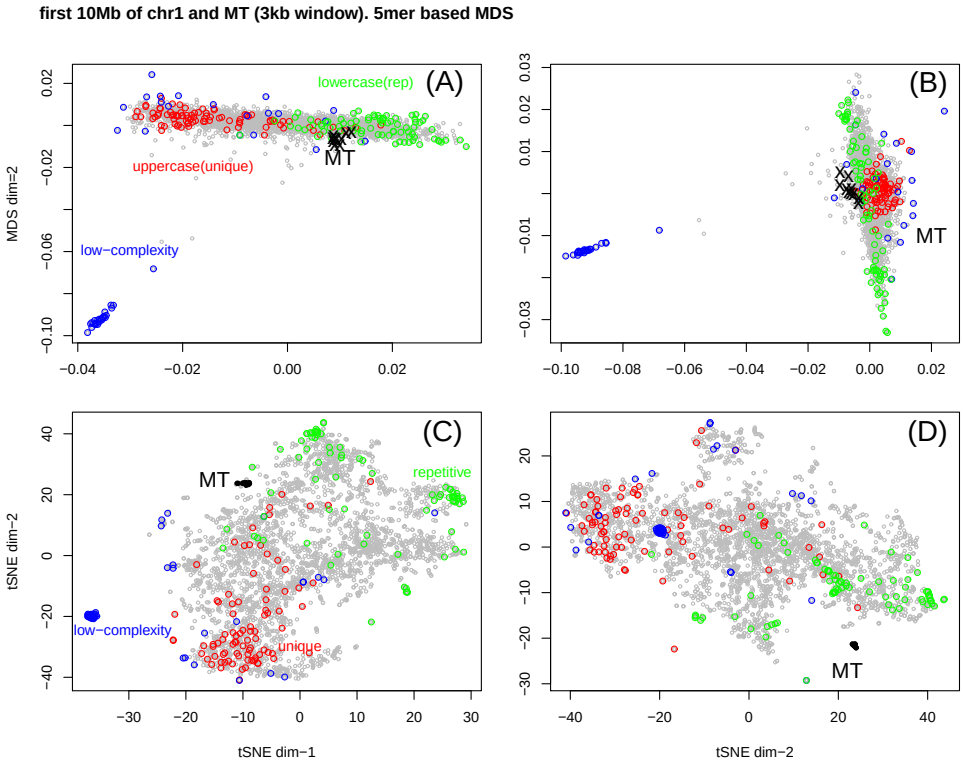

Figure 2: Multi-dimensional scaling (MDS) for (A) the 1st and 2nd dimension; and (B) 2nd and 3rd dimension. Each point represents a 5-mer frequency distribution of a 3kb window (non-overlapping window in human chromosome 1, overlapping in human mitochondrial DNA). Low complexity windows (E ¡ 8.1) are marked in blue. Windows with high proportion of non-repetitive (unique, not filtered by RepeatMask, uppercase) bases are marked in red, and those with high proportion of repetitive sequences (lowercase) are marked in green. The MT windows are marked in black. (C) and (D) are equivalent tSNE plot of the 1st vs 2nd dimension, and 2nd vs 3rd dimension.

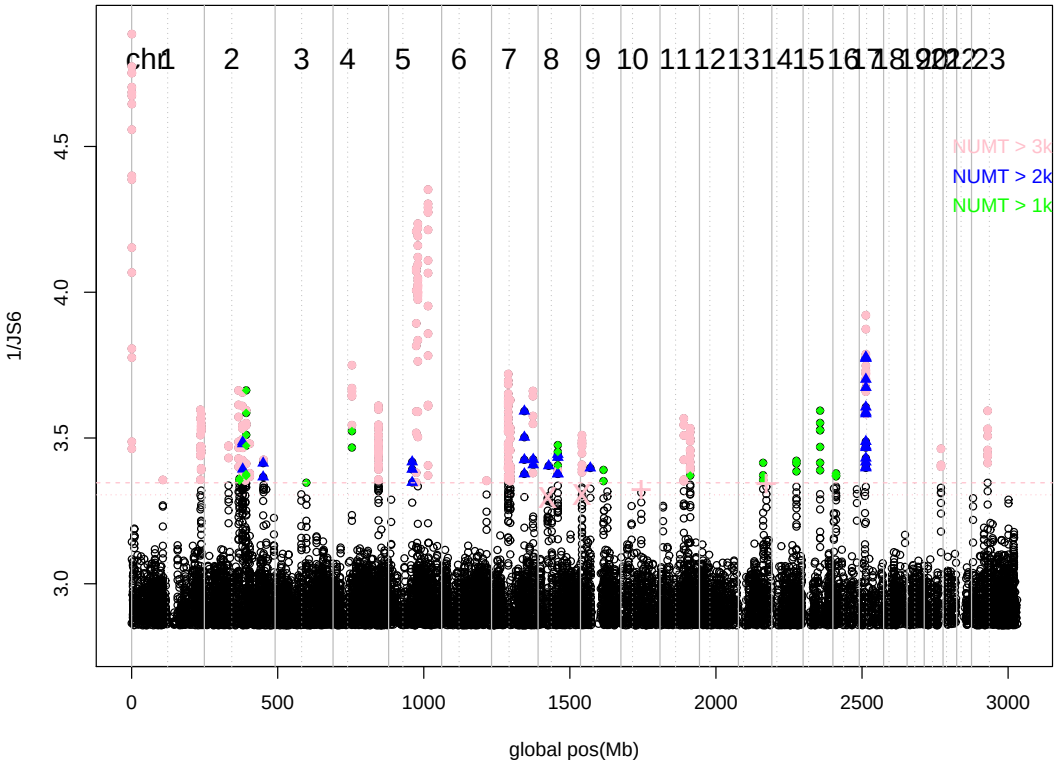

Figure 3: Manhattan plot of alignment-free NUMT detection signals, as obtained by 1/JS for k=6, and plotted as a function of the global genomic location (Mb). Each point represents a 3kb window, which moves by 1/8 of the window size (375 bases). Windows with 1/JS being less than 1/0.35=2.857 are not plotted. Chromosome start/end positions are marked by vertical solid lines, and centromeres marked by vertical dashed lines. The horizontal lines represent the threshold for peak calling: the threshold value for calling a NUMT prediction is chosen to be the second lowest 1/JS values among 38 previously known NUMTs (with size larger than 3kb). The difference between the two horizonal lines is that one was obtained from moving windows 375 bases (for 1/JS=3.346), and another from moving windows by 1 base (for 1/JS=3.305). All observed peaks above these lines can be explained by the known NUMTs as listed in (Calabrese et al., 2012): those with sizes larger than 3kb are colored in pink, those with size between 2kb and 3kb in blue, and those between 1kb and 2kb in green.

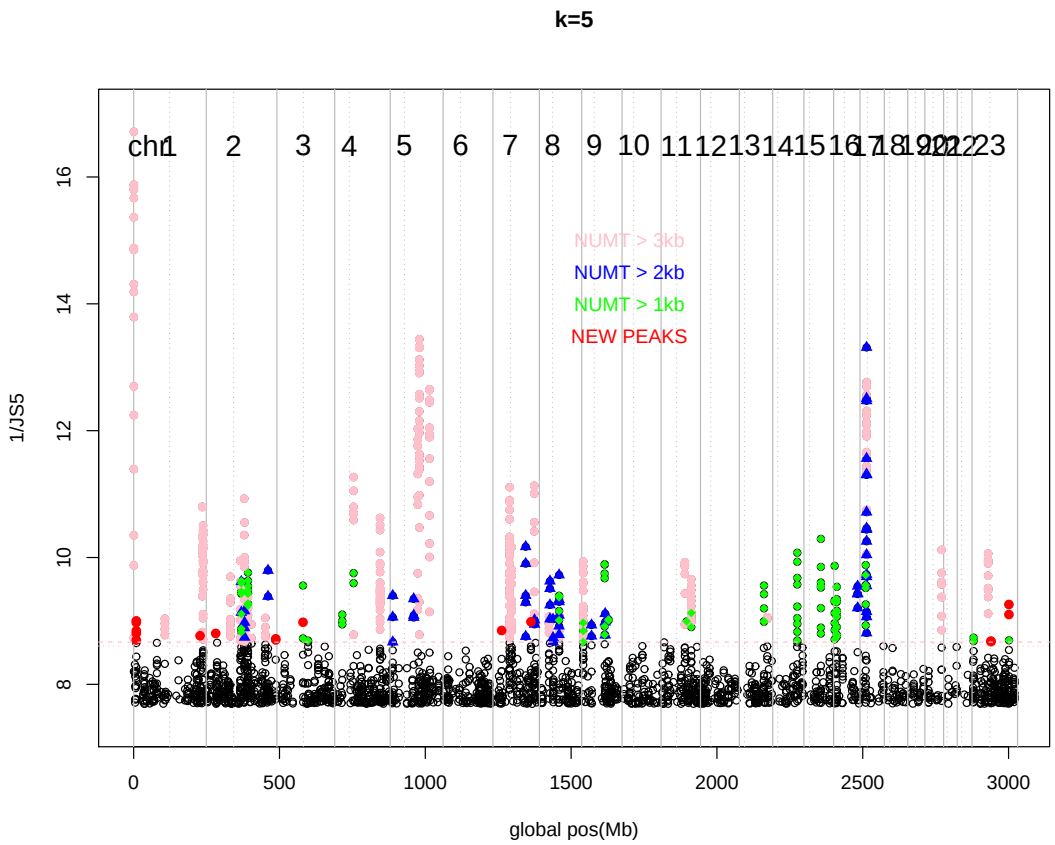

Figure 4: Manhattan plot of alignment-free NUMT detection signals, as obtained by 1/JS for k=5, again plotted as a function of the global genomic location (Mb). Each point again represents a 3kb window, which moves by 1/8 of the window size (375 bases). Windows with 1/JS less than 1/0.14 =7.143 are not plotted. The horizontal lines represent the threshold for calling NUMTs from peak signals: the second lowest 1/JS values among 38 known NUMTs with size larger than 3kb, 1/JS=8.604 for step size 375 and 1/JS=8.666 for step size 1. Peaks which are explained by known NUMTs listed in (Calabrese et al., 2012) of size larger than 3kb are colored in pink, those with size between 2kb and 3kb in blue, and those between 1kb and 2kb in green. New NUMT prediction peaks are marked in red.

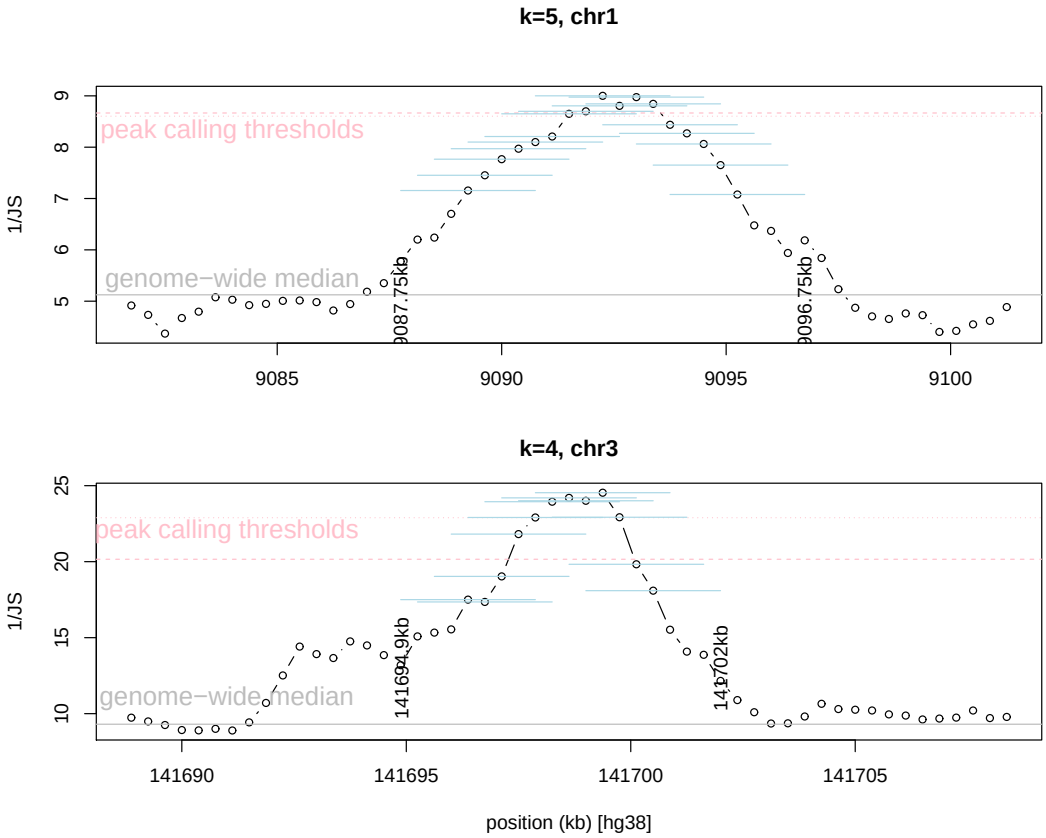

Figure 5: (A) Distribution of 1/JS for k=5 over a signal peak region at chromosome 1. The peak calling threshold is marked by pink, and genome-wide median of 1/JS is marked by grey. The peak width is estimated from the left end of the 3kb windows reaching 1/JS=7, to the right end of the 3kb window falling off from 1/JS=7, which is from 9087.75kb to 9096.75kb, or 9kb. (B) Distribution of 1/JS for k=4 over a signal peak region at chromosome 3. The width of the peak is estimated to be 7.1kb from 141694.9kb to 141702kb.