A probabilistic approach for mtDNA/NUMTs reads separation in NGS data

Introduction

NUMTs (Nuclear MiTochondrial Sequences) are mitochondrial DNA (mtDNA) sequence colonized in the nuclear genome during evolution[1]. How and why mtDNA incorporate into nuclear genome are still unclear. Some suggested that in endosymbiotic theory, mitochondria were originally prokaryotic cells and became endosymbionts for eukaryotic cells. Special transfer system has formed for communication between mitochondria and nucleus, and the nucleus continuously receives DNA from mitochondria. It has been hypothesized that the embedding of NUMTs is associated with chromosomal repair system during cell repair double-strand breaks by recombination[1,2].

NUMTs are distributed across almost all chromosomes in nuclear DNA (nDNA), sharing the sequence similarity to their mitochondria analogs. There are 755 NUMTs identified in human hg19 reference genome[3], size ranging from 39bp to more than 14kb, suggesting more than 600 independent integration events. Once inserted into nuclear genome, NUMTs are undergoing evolutionary processes: insertion, deletion, duplication and nucleotide substitutions, making some of the NUMTs sequencing very different from their parental mtDNA while others are more conservative.

NUMTs as pseudo genes are thought to affect the expression of functional genes and cause alteration of genome structure. Another important aspect of NUMTs is their disturbance of studying mitochondrial heteroplasmy. Unlike only two copies of nuclear DNA, there are hundreds to thousands mtDNA presenting in a cell. Heteroplamsy is the co-existence of the wild type mtDNA and mutated mtDNA, where the frequency of mutated allele can vary from 0% to 100%. Next generation sequencing (NGS) technology enable accurately measure the frequency of heteroplasmy by counting the number of reads containing the mutated allele over the total number of reads aligned to a specific mitochondria genome location. The existence of NUMTs can confound estimation heteroplasmy frequency since sequencing reads generated from NUMTs can be mistakenly aligned to mitochondrial genome, and the mismatches would be false called as heteroplasmy. Reads from mtDNA can also be aligned to NUMTs regions, making the measurement of heteroplasmy frequency not accurate.

Computational and molecular approaches have been developed to help solve this issue. The most widely used approach to distinguish NUMTs reads is to use the number of mismatch between reads and reference genome as measure of alignment quality [4]. However, evaluating an alignment location just by the number of nucleotide mismatches alone is not optimal because it does not consider following aspects: i) different genome regions have biased mutation patterns ii) existing polymorphisms in NUMTs regions and mtDNA. iii) sequencing quality of mismatch bases. iii) cannot deal with alignment locations with same number of mismatch.

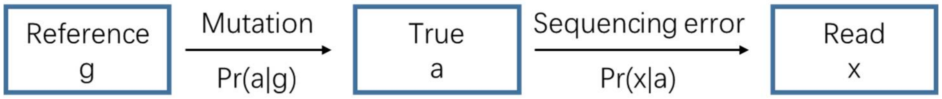
Evolution, mutation pattern and sequencing quality are naturally dealt with probability. In this project, I present a probabilistic approach to do mapping location assignment. All the three above aspects can be encoded in a probability model. Different alignments will have different model parameters, I can calculate the likelihood of each alignment under different models, and evaluate them by likelihood value.

Method

After sequencing, we can align the sequencing data to the reference genome. Some reads are uniquely mapped to a genome location, but some other reads can align to both mtDNA and NUMTs (denote them as mt reads and NUMT alignment, respectively), to evaluate which alignment is better, I will calculate the log likelihood of the read under NUMT model and mitochondrial model ($\Pr(\mathrm{read}|\mathrm{NUMT})$; $\Pr(\mathrm{read}|\mathrm{mt})$), and compare which one has higher likelihood.

For a given alignment, let's take one base at a time to calculate the likelihood under the proposed models. Briefly, for a specific genome site i, let's denote the nucleotide base in reference is g, the true base in the genome of interest is a, and the base shown in the sequencing read is x. There are two processes that can make x be different from g: 1. A mutation event that the reference g changes to a 2. A sequencing error. These two process can be presented as a bayesian network (Figure 1).

Figure 1



Therefore, the alignment model should contain two sub-models, one for the mutation process (**evolutionary model**) and the other one for the sequencing error (**sequencing error model**). If we assume mutation and sequencing error are independent, we can calculate the probability of observing x given the reference base is g in a sequencing read by eq 1, where $\Pr(\mathrm{a}|\mathrm{g})$ is the evolutionary model and $\Pr(\mathrm{x}|\mathrm{a})$ is the sequencing error model.

$$\Pr(\mathrm{x}|\mathrm{g}) = \sum_{\mathrm{a}} \Pr(x|a)\Pr(a|g). \quad a \in A,T,G,C \quad (1)$$

If bases in a read are independent from each other, the likelihood of a sequencing read coming from a certain genome location would be:

$$\Pr(\text{read}|\mathbf{g}) = \prod_{i=1}^{l} \Pr(x_i|g_i)$$

Where **g** is a genome location, *l* is the read length.

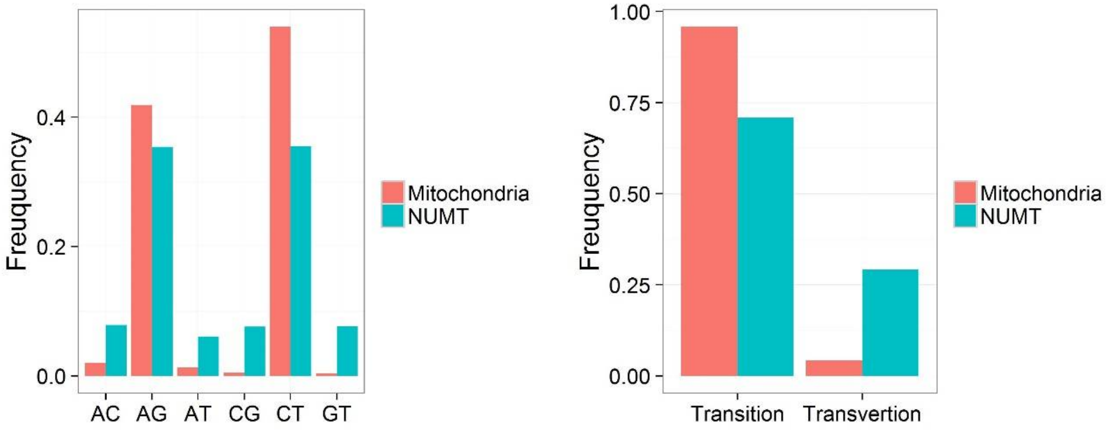
Evolutionary model:

After NUMTs inserting into nuclear genome, they are under different evolutionary selection pressures to mtDNA, as a consequence, mutation patterns are very different for NUMTs and mtDNA. I calculate the frequency of 6 different kind of nucleotide substitutions in using population variants data in 1000 genome project. And find these two have dramatic difference (Figure 2). In both scenarios, there are more transitions versus transversions, but the ratios are quite different: Ts/Tv for NUMTs is 2.43, which is close to typical nuclear genome Ts/Ts ratio (~2.1), while, Ts/Ts for mitochondria is 23. These distinct mutation patterns can help distinguish reads from different locations. We can incorporate it into a simple model:

$$\Pr(a|g) = \begin{cases} 1 - p_s & if\ a = g \\ p_s p_{ts} & if\ g\ to\ a\ is\ a\ transition \\ 1/2(p_s - p_s p_{ts}) & if\ g\ to\ a\ is\ a\ transvertion \end{cases} \quad (2)$$

In this model, $\mathrm{p_s}$ is the probability that a substitution happens, $\mathrm{p_{ts}}$ is the probability of a transition happens given that a substitution have happened. By this Ts/Tv model we can calculate the base probability of a given genome site.

Figure 2



This Ts/Tv model assumes substitutions happen uniformly across the genome, which is not true in real biology data. Thus we could also collect nucleotide base frequencies using existing genome sequencing data. If a genome position is a polymorphic site in population, our expectation to observe an alternative nucleotide at this position is higher than a non-polymorphic site. Thus except for the above transition/transvertion model, we can also represent the bases probabilities at a specific genome locus by the frequency in population.

The final evolutionary model will be a combination of these two. At a site, if a non-reference base is observed as alternative nucleotide in population, its probabilities in the model will be its frequency in population data, otherwise, its probability will follow the Ts/Tv model. Population variants probabilities are calculated from 1000 genome project[5] vcf file for NUMTs regions and MITOMAP[6] for mitochondrial genome.

Sequencing error model:

Sequencing error model could be same for both mt model and NUMT model. Illumina sequencing platforms provide sequencing quality for each sequenced base using phred score, which derived from the probability that the nucleotide is wrongly assigned in the base-calling step. Illumina sequencer has an error rate range from 0.1% to several percent, and the error probability is not uniform across the reads. Generally, the error rate is higher at the 5’ end and lower down toward 3’ end, and error probability is higher for Read 2 comparing to Read 1 in pair end sequencing. Obviously, a mismatch happened at high quality base is more reliable than a mismatch happened at low quality base. Thus it’s necessary to include the sequencing error sub-model.

According to phred score definition, the error probability can be calculated from quality score Q by$\Pr(\mathrm{error}) = 10^{\frac{-Q}{10}}$ $\Pr(\mathrm{error}) = 10^{\frac{-Q}{10}}$. From this the sequencing error model can be presented as

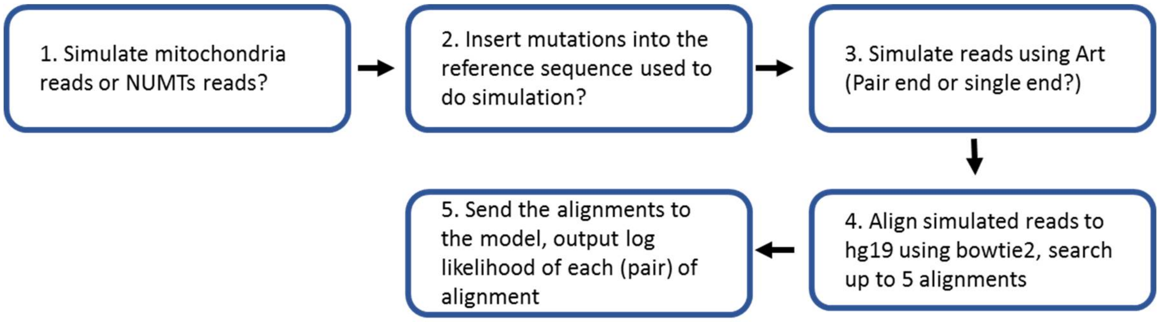
$$\Pr(a|x) = \begin{cases} 1 - Pr_{error} & if\ a = x \\ \dfrac{Pr_{error}}{3} & if\ a \neq x \end{cases}$$

Result:

To test the performance of this method, I simulated some sequencing data from mitochondria genome and NUMTs separately and run the pipeline to see whether the simulated reads are correctly assigned as mitochondrial reads or NUMT reads.

The simulation workflow is shown as Figure 3. In this pipeline, I would first choose to simulate mitochondrial reference genome or NUMTs reference sequences. Next I could choose whether to insert mutation into the reference for simulation. Here I could force to mutate all the polymorphic site or randomly mutate them according to their allele frequency using binomial distribution. After preparing the reference sequences, I used ART[7] to simulate sequencing reads, in which I could choose the sequencing platform to mimic the sequencing error, I could also choose simulate single end reads or pair end reads. The sequenced reads were then aligned to hg19 reference using bowtie2[8]. Here I allowed the software to output up to 5 alignments for each (pair) of read(s). For reads that can align to multiple locations, I would put them into the proposed model to see which alignment has the highest likelihood, and assign the reads based on these likelihood value.

Figure 3



1. Simulate mitochondria reads or NUMTs reads?

2. Insert mutations into the reference sequence used to do simulation?

3. Simulate reads using Art (Pair end or single end?)

4. Align simulated reads to hg19 using bowtie2, search up to 5 alignments

5. Send the alignments to the model, output log likelihood of each (pair) of alignment

I tried several parameters to do the simulation and evaluated the performance of this method, summarized in Table 1:

Table 1

	1	2	3	4	5	6	7	8
Mt/NUMT	Mt	NUMT	MT	NUMT	MT	NUMT	MT	NUMT
Mutation	Insert all	Insert all	Insert by frequency	Insert by frequency	Insert all	Insert all	Insert by frequency	Insert by frequency
PE/SE	SE	SE	SE	SE	PE	PE	PE	PE
Simulated reads (pair) number	159986	239098	162956	249050	157054	238626	163084	249049
# of reads (pair) with multiple alignments (% of total reads)	27302 (17.1%)	2741 (1.1%)	26812 (16.5%)	1249 (0.5%)	123343 (78.5%)	5200 (2.2%)	123627 (75.8%)	7854
Overall accuracy	74.5%	79.8%	54.6%	66.9%	98.9%	96.7%	96.7%	98.5%
Accuracy for multiple alignments with same mismatches	79.6%	79.2%	71.0%	82.4%	92.4%	98.4%	52.3%	97.0%

I first tried simulate single end reads with forced inserted mutations or binomial inserted mutations. In these cases, I simply classify the reads to the alignments with highest log likelihood. We can evaluate the classification of reads aligned to multiple location by several criteria: 1. Overall classification accuracy 2. Classification accuracy for reads aligned to multiple locations with same number of mismatches (# of mismatch > 0).

From table 1 we could see, for single end there are about 17% mitochondrial reads can align to multiple locations, the overall accuracy for classification are 74.5% and 54.6% for different mutation insertion method since more empirical mutations insertions can help separate the reads. In those reads with multiple alignments, we are more curious about the alignments with same number of mismatches (mismatches > 0), for example, if a read can align to both mtDNA and NUMTs with 2 mismatches. In these scenarios, the method has 79.6% and 71.0% accuracy. There are less NUMTs reads can align to multiple locations, and has similar accuracy comparing to mtDNA reads.

Next I simulated pair end reads with similar mutation insertions. In pair end mode, I would classify the reads by the sum of log likelihood of the read pair. Similarly, I evaluated the results by overall accuracy and accuracy of reads with same number of mismatches.

There are more reads pairs aligned to multiple locations in pair end setting, but the classification accuracies are actually higher because we have the information from both end of a DNA fragment. In most cases, the classification can achieve accuracy > 90%. But for mtDNA reads, when classifying the alignments with same mismatches to NUMTs, the accuracy is only 52.3%. This low accuracy may due to the read pairs have only 1 mismatch for both end, in this situation, it will be very difficult to assign the alignments.

Real data application:

I next test the method on real sequencing data. Human mitochondrial genome is ~16kb long. To target sequence this small genome, researcher will first use PCR to amplify it, (usually perform 2 overlap PCR, each about 9kb long to cover the entire mitochondrial genome), then sequencing the PCR product. If we assume the PCR primer can specifically amply mtDNA, the sequencing reads should all be mtDNA reads. We can use this kind of sequencing data to evaluate whether mtDNA reads can be correctly classified.

I used 3 real sequencing samples to evaluate the method. Results are summarized in Table 2. The results are comparable to simulation result 7. The overall accuracy for all 3 samples are over 90%. But for alignments with same number of mismatches, the accuracy is still low, ie, many mtDNA reads are mistakenly classified as NUMTs reads.

Table 2

	1	2	3
PE/SE	PE	PE	PE
Simulated reads (pair) number	68626	68396	65590
# of reads (pair) with multiple alignments (% of total reads)	49475 (72.1%)	49619 (72.5%)	50363 (76.8%)
Overall accuracy	91.7%	96.4%	96.4%
Accuracy for multiple alignments with same mismatches	36.2%	47.4%	48.7%

Discussion

In this project I implemented a probabilistic method to classify sequencing reads that can align to both mitochondrial genome and NUMTs regions in nuclear genome, and evaluated the performance using simulated sequencing reads and real mtDNA sequencing data.

Overall, this method has good accuracy of assigning reads to their correction origins, especially for pair end sequencing reads. When we zoom into subgroup of reads, if a pair of reads can align to multiple regions with same number of mismatch, mtDNA originated reads has low classification accuracy (52.3% in simulation data, and ~47% in real sequencing data). That means half of these mtDNA reads are misclassified as NUMTs reads, in real applications, they will be removed from downstream analysis, which can a waste of sequencing data. However, as I mentioned in introduction part, the main aim of this method is to remove NUMTs reads to inhibit them to affect mtDNA heteroplasmy identification. Thus we are more curious about whether the NUMTs reads can be correctly classified as NUMTs. In the simulation data we can achieve 98.4% accuracy for NUMTs reads (with same mismatches to mtDNA). Given that total number of NUMTs reads is much smaller than mtDNA reads in real biology data (because mtDNA copy number is from hundreds to thousands while nuclear DNA is only 2), this method can effective eliminate the contamination of NUMTs when analyze mtDNA sequence.

This method can also be improved from several aspects: 1. Current implementation only considered SNPs in population, we can also include insertion and deletions in the future versions since indels can be stronger evidence for sequencing difference between mtDNA and NUMTs. I have also implement the code to capture all the alignments indels in each read, the method can be adapted easily. 2. An assumption of this method is that each site in a sequencing read is independent, which is not true. There can be some linkage between the variants, adding this information can also help increase the accuracy.

There are also some scenarios that current method cannot deal with, which may need further improvements: 1. When sequencing reads align to mtDNA and NUMTs with 0 mismatches. Without any difference between alignments, this method will not work properly. 2. This method also heavily rely on reference genome and empirical population variants. If there are de nova NUMTs in a sequenced individual, it will be difficult to distinguish these NUMTs reads, especially the NUMTs are inserted recently.

Reference

1 Gaziev, A. I. & Shaikhaev, G. O. Nuclear mitochondrial pseudogenes. *Molecular Biology* **44**, 358-368, doi:10.1134/s0026893310030027 (2010).

2 Hazkani-Covo, E. & Covo, S. Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS Genet* **4**, e1000237 (2008).

3 Calabrese, F. M., Simone, D. & Attimonelli, M. Primates and mouse NumtS in the UCSC Genome Browser. *BMC Bioinformatics* **13**, S15, doi:10.1186/1471-2105-13-s4-s15 (2012).

4 Ye, K., Lu, J., Ma, F., Keinan, A. & Gu, Z. Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proceedings of the National Academy of Sciences* **111**, 10654-10659, doi:10.1073/pnas.1403521111 (2014).

5 The Genomes Project, C. A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393

http://www.nature.com/nature/journal/v526/n7571/abs/nature15393.html#supplementary-information (2015).

6 Ruiz-Pesini, E. *et al.* An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic acids research* **35**, D823-D828 (2007).

7 Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)* **28**, 593-594, doi:10.1093/bioinformatics/btr708 (2012).

8 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9**, 357-359, doi:10.1038/nmeth.1923

http://www.nature.com/nmeth/journal/v9/n4/abs/nmeth.1923.html#supplementary-information (2012).