

Abstract

DNA barcodes have achieved prominence as a tool for species-level identifications. Consequently, there is a rapidly growing database of these short sequences from a wide variety of taxa. In this study, we have analyzed the correlation between the nucleotide content of the short DNA barcode sequences and the genomes from which they are derived. Our results show that such short sequences can yield important, and surprisingly accurate, information about the composition of the entire genome. In other words, for unsequenced genomes, the DNA barcodes can provide a quick preview of the whole genome composition.

Questions?

- Could the nucleotide composition of a DNA barcode reflect the overall nucleotide composition of the parent genome?
- Could the nucleotide skews of a DNA barcode reflect the overall nucleotide skews of the parent genome?
- Could the attributes of a DNA barcode reflect the overall attributes of all protein-coding DNA sequences?

Data and Methods

- A total of 849 complete mitochondrial genomes (mtDNA) in metazoa were collected.
- Standard barcodes of ~648 nt from cytochrome c oxidase subunit I (COI) were retrieved from each genome.
- Nucleotide frequencies of mtDNA, total coding DNA, and DNA barcode sequences were analyzed.
- The GC- and AT- skews were measured according to the following formulae:

$$\text{GC-skew} = \frac{(G-C)}{(G+C)}$$

$$\text{AT-skew} = \frac{(A-T)}{(A+T)}$$

Results

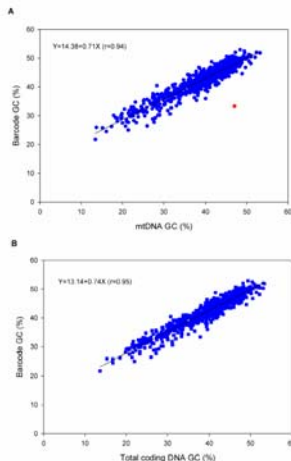


Figure 1. Correlation between nucleotide content of DNA barcodes and the nucleotide content of the mitochondrial genome. Panel A shows the GC content of the DNA barcodes plotted against the GC content of the entire mitochondrial genome (including protein-coding and non-coding sequences). The outlier point (shown in red) is from *Trichoplax adhaerens* (NC_008151). Panel B shows the DNA barcode GC content plotted against the GC content of the combined protein-coding sequences only.

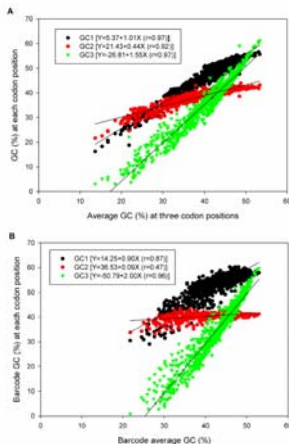


Figure 2. Correlation between GC content at each codon position and the average GC content. Panel A shows the results for all protein coding sequences in the mitochondrial genome. Panel B shows the results based on the DNA barcode region alone.

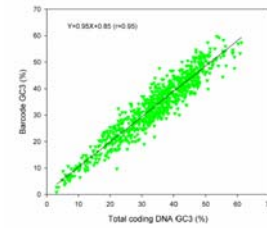


Figure 3. Correlation between GC content of the entire mitochondrial genome and the DNA barcode sequences (the nucleotide contents were calculated for the third codon position only).

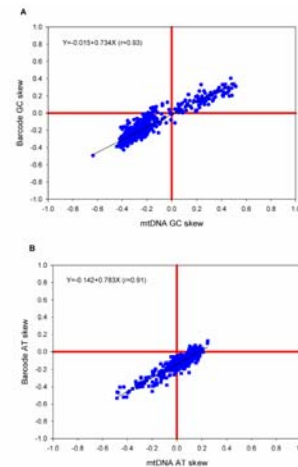


Figure 4. Correlations between nucleotide skews in the entire mitochondrial genome and in the DNA barcode sequences. Panel A shows the results for GC skew [(G-C)/(G+C)]. Panel B shows the results for AT skew [(A-T)/(A+T)].

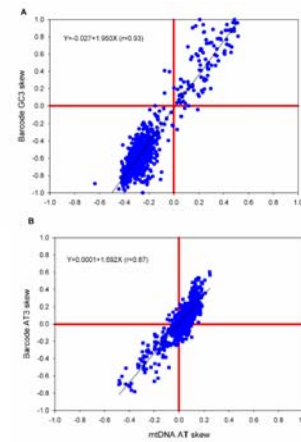


Figure 5. Correlations between nucleotide skews in the entire mitochondrial genome and in the DNA barcode sequences (third codon position only). Panel A shows the results for GC skew [(G-C)/(G+C)]. Panel B shows the results for AT skew [(A-T)/(A+T)].

Summary

- ✓ The nucleotide composition of DNA barcode sequences is highly correlated with nucleotide content of the parent mitochondrial genomes.
- ✓ The nucleotide composition skews (GC- and AT-skews) of DNA barcodes are highly correlated with nucleotide skews of the entire mitochondrial genomes.
- ✓ The nucleotide composition of DNA barcodes is highly correlated with the nucleotide content of total coding DNA sequences. The variation of nucleotide composition at each codon position in barcodes also reflects the variation of nucleotide composition of total coding DNA sequences of the entire genomes.

Acknowledgments

This research was supported through funding to the Canadian Barcode of Life Network from Genome Canada through the Ontario Genomics Institute, NSERC and other sponsors listed at www.BOLNET.ca.

This work is published as a full paper:

Min XJ, Hickey DA (2007) DNA Barcodes Provide a Quick Preview of Mitochondrial Genome Composition. PLoS ONE 2(3): e325. doi:10.1371/journal.pone.0000325