Beyond DNA barcoding: The unrealised potential of genome skim data in sample identification

Kristine Bohmann[1], Siavash Mirarab[2], Vineet Bafna[3], and Tom Gilbert[4]

[1]University of Copenhagen
[2]UC San Diego
[3]UCSD
[4]University of Copenhagen Globe Institute

May 5, 2020

Posted on Authorea 26 Feb 2020 — CC BY 4.0 — https://doi.org/10.22541/au.158274776.61753752 — This a preprint and has not been peer reviewed. Data may be preliminary.

Abstract

Genetic tools are increasingly used to identify and discriminate between species. One key transition in this process was the recognition of the potential of the ca 658bp fragment of the organelle cytochrome c oxidase I (COI) as a barcode region, which revolutionised animal bioidentification and lead, among others, to the instigation of the Barcode of Life database (BOLD), containing currently barcodes from >7.9 million specimens. Following this discovery, suggestions for other organellar regions and markers, and the primers with which to amplify them, have been continuously proposed. Most recently, the field has taken the leap from PCR based generation of DNA references into shotgun sequencing-based 'genome skimming' alternatives, which the ultimate goal of assembling organellar reference genomes. Unfortunately, in genome skimming approaches, much of the nuclear genome (as much as 99% of the sequence data) is discarded, which is not only wasteful but can also limit the power of discrimination at or below the species level. Here, we advocate that the full shotgun sequence data can be used to assign an identity (that we term for convenience its 'DNA-mark') for both voucher and query samples, without requiring any computationally intensive pretreatment (e.g., assembly) of reads. We argue that if reference databases are populated with such 'DNA-marks', it will enable future DNA-based taxonomic identification to complement, or even replace PCR of barcodes with genome skimming, and we discuss how such methodology ultimately could enable identification to population, or even individual, level.

Hosted file

`Bohmann_et_al.pdf` available at https://authorea.com/users/300456/articles/430111-beyond-dna-barcoding-the-unrealised-potential-of-genome-skim-data-in-sample-identification