UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

PhD Defence

Thursday May 5, 2022 at 9am via Zoom

A Novel Statistical Framework for Assessment of Intraspecific Haplotype Sampling Completeness

Jarrett Phillips

Chair: Dr. Joe Sawada
Advisor: Dr. Dan Gillis
Co-Advisor: Dr. Robert Hanner [Integrative Biology]
Non-Advisory Member: Dr. Andrew Hamilton-Wright
External Examiner: Dr. Karen Kopciuk [University of Calgary]

ABSTRACT:

The problem of determining adequate sample sizes necessary for studies of biodiversity conservation and management is a challenging one that has received some attention in recent years. One particular area where the probing of sampling completeness is of utmost priority is DNA barcoding. Species show remarkable genomic marker variation within and among taxa, along with differing evolutionary and life histories. Thus, knowing how many specimens of a given species likely need to be collected to observe the majority of standing COI haplotype diversity present within animal species is a complex question to answer. Estimates of specimen sample sizes for DNA barcoding range from a single individual to hundreds of individuals per species (but typically around 5-10 individuals). However, due to obstacles surrounding project funding and species rarity, often just one or two specimens per species can be reasonably collected. In addition, numerous other factors, especially sequence quality and integrity, hinder the accurate and reliable estimation of specimen sample sizes from existing species-level sequence data found in large DNA repositories.

Here, a deep examination of the genetic specimen sample size problem (GSSSP) is undertaken. Specifically, a novel nonparametric stochastic local search optimization algorithm based on trends in species haplotype accumulation curves, herein called HACSim (Haplotype Accumulation Curve Simulator) is introduced. The method, available as an R package, is tested on a variety of both hypothetical and real animal species mined from the Barcode of Life Data Systems (BOLD). Through a detailed statistical simulation study, the approach is demonstrated to work well across all examined scenarios. As HACSim makes numerous simplifying assumptions that are unlikely to hold well in practice, such as panmixia (random mating), future work in incorporating elements of population structure is imperative.

In addition, it is argued that DNA barcoding currently lacks in statistical rigor needed to robustly estimate the DNA barcode gap, an important quantity expressing the difference between intraspecific and interspecific genetic variation. A number of accessible statistical solutions revolving around sample sizes needed for gap assessment, as well as visualization and inference are offered in this regard.