UNIVERSITY of GUELPH

COLLEGE *of* ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

PhD Seminar 2

Friday February 28, 2020 at 1:30PM in Reynolds, Room 2224

A Novel Statistical Framework for Assessment of Intraspecific Haplotype Sampling Completeness

Jarrett Phillips

Advisor: Dr. Dan Gillis

Co-Advisor: Dr. Bob Hanner [Integrative Bio and Biodiversity]

Advisory Committee: Dr. Deb Stacey

Advisory Committee: Dr. Graham Taylor [Engineering]

ABSTRACT:

Biodiversity manifests itself in many ways. One way is through the lens of species' genetic diversity. DNA barcoding offers a systematic means of documenting such diversity in the face of global species extinction. One question of interest among DNA "barcoders" is: How many specimens of a given species do we need to sample before we can stop? That is, what sample size is needed to capture a given level of a species' genetic diversity? This is not an easy question to answer! At a practical level, specimen sample sizes typically range from 1-10 individuals per species. However, a number of studies have demonstrated through both empirical investigations and statistical simulations that such small, largely arbitrary sample sizes are far from enough. In fact, some studies even suggest that hundreds to thousands of individuals may be necessary to be sampled before sampling can safely be stopped.

A common tool of the trade to greatly aid biodiversity scientists in assessing the overall completeness of specimen sampling is to generate haplotype accumulation curves. These curves depict the degree of haplotype saturation as a function of the number of individuals sampled at random. If the curves level off to an asymptote, this is a good indication that we have probably found all the haplotype diversity existing for a given species. If, on the other hand, curves show no evidence of reaching a plateau, then greater sampling effort is needed.

HACSim, short for **H**aplotype **A**ccumulation **C**urve **Sim**ulator is a new R package developed to greatly facilitate the process of estimating likely required specimen sample sizes for recovery of species' genetic variation. The approach underlying HACSim is fundamentally different than what has been attempted before. What separates HACSim from the rest of the pack is that it is a nonparametric approach that combines two clever statistical ideas to propose plausible specimen sample sizes necessary to adequately recover a given level of haplotype diversity for a species of interest. These are: 1) stochasticity (randomness) and 2) iteration. These two characteristics together represent a step forward in thinking critically about how genetic variation manifests itself in contemporary patterns of biological diversification and how well current sampling efforts contribute to our understanding of these patterns and the underlying evolutionary processes that generate them.