



Virginia Commonwealth University  
VCU Scholars Compass

---


Biology and Medicine Through Mathematics  
Conference

---

## Statistical inference of adaptation at multiple genomic scales using supervised classification and a hidden Markov model

Lauren A. Sugden  
*Duquesne University*, [sugdenl@duq.edu](mailto:sugdenl@duq.edu)

Follow this and additional works at: <https://scholarscompass.vcu.edu/bamm>

 Part of the [Biostatistics Commons](#), [Computational Biology Commons](#), [Evolution Commons](#), [Genomics Commons](#), [Population Biology Commons](#), and the [Statistical Models Commons](#)

---

<https://scholarscompass.vcu.edu/bamm/2020/talk/17>

This Event is brought to you for free and open access by the Dept. of Mathematics and Applied Mathematics at VCU Scholars Compass. It has been accepted for inclusion in Biology and Medicine Through Mathematics Conference by an authorized administrator of VCU Scholars Compass. For more information, please contact [libcompass@vcu.edu](mailto:libcompass@vcu.edu).

## **Statistical inference of adaptation at multiple genomic scales using supervised classification and a hidden Markov model**

Lauren A Sugden, Department of Mathematics, Statistics, and Computer Science  
Duquesne University, Pittsburgh PA

Identifying evidence for adaptation in the genome is a longstanding challenge in evolutionary biology. While a handful of highly adaptive mutations in humans carry strong genomic signatures that are easily identifiable, a deeper understanding of the full scope of human adaptation requires the development of robust statistical tools that can register subtler patterns, and integrate information across genomic regions. I will present here two such approaches: the first approach focuses on gaining power to detect single mutations under positive selection, while the second uses a hidden Markov model to integrate evidence for selection across a gene region.

To detect single mutations, we use a supervised classification method called Averaged One-Dependence Estimation to produce calibrated probabilities of adaptation at every genomic site, using simulated training data generated using multiple evolutionary scenarios. This published approach, which we call SWIFr, identifies many putative targets of adaptive evolution in human population data from the 1000 Genomes Project, including well-studied adaptive mutations from the literature. We applied this approach to data from the from the ‡Khomani San hunter-gatherer population of southern Africa, finding an enrichment of adaptive mutations in genes associated with metabolism, suggesting that efficiency of energy storage has been a source of selective pressure in this population.

To address the integration of selection signatures across gene regions, I will describe ongoing work in which we use a hidden Markov model (HMM) to leverage the physical linkage of mutations in a linear strand of DNA. Our hidden states correspond to mutations that are neutral, adaptive, and under the influence of nearby adaptive mutations, and our transition matrix reduces the noise of site-by-site predictions by ensuring that only a single mutation be chosen as adaptive in a given region. By drawing thousands of representative samples using stochastic backtrace through our HMM, we can generate measures of uncertainty for each site-by-site prediction, as well as region-level probabilities of selection, enabling downstream analysis of adaptive signatures in gene pathways and networks.