Seminar/Talk

Compression population genomics

Stuart J.E. Baird
Czech Academy of Sciences, Brno

Host: Nick Barton

Population genomics requires us to summarise large quantities of information. Ideally this would be through lossless compression, such that the entire original information could be reconstructed from the summary. Inference could then proceed by co-estimation of parameters and error rates, avoiding the hazards of stepwise estimation. In contrast: a) Site-by-site summaries lose flanking sequence context; inefficient because a site and its flanks may be embedded within a tract of shared history. b) Within-individual summaries lose population context; inefficient because histories can be shared across tracts within individuals. Such tracts of shared history can be compressed without loss of information using run length encoding (RLE) eg: all individuals have the same history for a run of length 1234 sites. This suggests RLE as a potentially efficient compression alternative to site-by-site/within-individual summarisations that in addition retains both sequence and population context. Retaining this context allows better informed thresholding decisions, such as 'calling' sequence state. The robustness of inference to arbitrarily chosen thesholding levels is much more efficiently explored when such decisions occur at the end, rather than the start, of information processing. As a worked example, I explore the construction and properties of RLE-based summaries for population genomics using a sample of 19 mice sampled across Eurasia and the European house mouse hybrid zone.

Friday, October 6, 2017 10:00am - 11:00am
Mondi Seminar Room 2, Central Building

This invitation is valid as a ticket for the ISTA Shuttle from and to Heiligenstadt Station. Please find a schedule of the ISTA Shuttle on our webpage: https://ista.ac.at/en/campus/how-to-get-here/ The ISTA Shuttle bus is marked ISTA Shuttle (#142) and has the Institute Logo printed on the side.