Stuff casted out of paper

Yu Huang

November 28, 2007

Abstract

Here contains stuff not included in the official paper or trial and failure.

Contents

1 Identity

1.1 2007-10-21 thinning the identity pairs by removing ambiguous NA-rich strains

To enable component to reach transivity (true haplotype), magnus suggested a thinning algorithm.

Sort the individuals so that those with the smallest numbers of NA come on top. Go through the individuals one at a time: if an individual is different (has a mismatch, ignoring NA) from one you have seen, add it to the list of unique haplotypes.

Given the list of unique haplotypes. Now go through the remainder and assign those that can be assigned uniquely (no mismatches). Leave the ones that cannot be assigned uniquely unassigned.

this is a greedy algorithm, which can't gaurantee the transivity. (another algorithm is on the way to settle this...)

the number of identity pairs dropped to 364801 (15% less) but the inter population identities percentage, 244826/364801=0.67112206381 is almost same as before.

Figure 1 is the identity map after thinning. global pattern is still there. and two cross-ocean components are still there (data not shown).

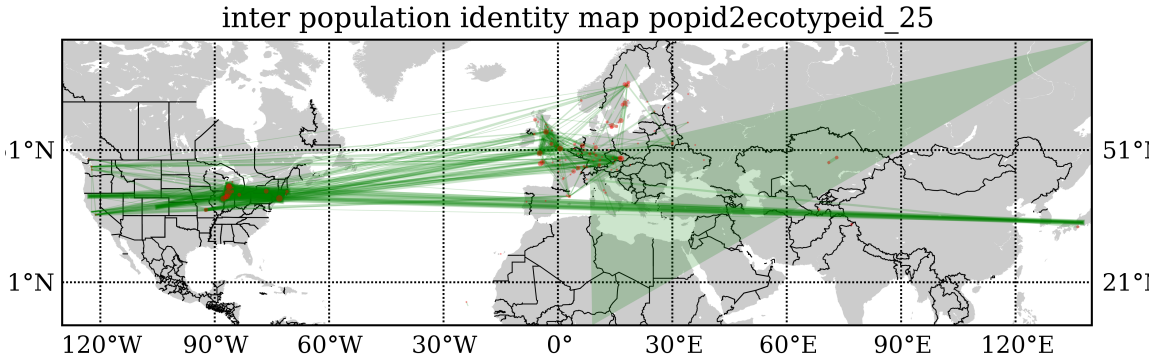

Figure 1: Population Identity Map after thinning