File S2

Supplementary Methods and Results

Measures of differentiation measured on a per SNP basis

δ is the absolute value of the difference in minor allele frequency among populations.

$$\delta = |\text{ Minor Allele Frequency}_{\text{pop1}} - \text{Minor Allele Frequency}_{\text{pop2}}|$$

D_{xy} can be thought of as the number of mismatches between two sets divided by the total number of comparisons between two sets.

$$D_{XY} = \frac{\left(\text{Minor Allele Count}_{\text{pop1}} * \text{Major Allele Count}_{\text{pop2}}\right) + \left(\text{Major Allele Count}_{\text{pop1}} * \text{Minor Allele Count}_{\text{pop2}}\right)}{\text{Number of Alleles}_{\text{pop1}} * \text{Number of Alleles}_{\text{pop2}}}$$

F_{st} is the portion of the variance in the data that lies between two populations.

$$F_{st} = \frac{Pi_{total} - \overline{Pi}_{within}}{Pi_{total}}$$

$$Pi_{total} = \frac{\text{Minor Allele Count}_{\text{total}} * \text{Major Allele Count}_{\text{total}}}{\binom{\text{Total number of Alleles}}{2}}$$

$$Pi_{within\ for\ popk} = \frac{\text{Minor Allele Count}_{\text{popk}} * \text{Major Allele Count}_{\text{popk}}}{\binom{\text{Number of Alleles in popk}}{2}}$$

Runs of fixed differences Another approach to evaluating differentiation across the genome is to consider runs of fixed differences. When sampling is adequate, runs of fixed differences uninterrupted by shared polymorphisms, can also identify fully sorted gene genealogies. For this analysis, we only included genes that contained at least one fixed difference or shared polymorphism from each pairwise comparison. We sampled a single SNP from each gene included in the analysis. Because we were interested in identifying highly differentiated regions, to be conservative, if a gene contained fixed differences and shared polymorphisms, the SNP included in the analysis was selected from among the shared polymorphisms. On average, "pruned" SNPs included in these analyses were ~2.19 Mbs apart. Using publicly available source code, we amended the program SLIDER (McDonald 1996) to generate a distribution of runs of fixed differences based on 10,000 Monte Carlo simulations of coalescence and recombination for each pairwise comparison. In each simulation, the observed number of polymorphisms and fixed differences were distributed randomly among sites such that the number of polymorphisms and fixed differences matched the observed data. These simulations assumed a constant N_e, uniform recombination rates among adjacent sites,

random union of gametes, point mutation, and silent site neutrality. We used data from chromosome two for these simulations as it had, on average, the largest number of topologically informative markers and is the second largest autosome (~182 Mb). We replicated 10,000 simulations over ten recombination parameters ranging from one to ten.

We identified many runs of fixed differences in all pairwise comparisons (Supporting Information Figures 1a, b, c). Consistent with the window analyses, we found that there were more runs of fixed differences in the DM comparison and that those runs were, on average, larger both in terms of number of SNPs and distance covered (Supporting Information Table 5). However, SLIDER analysis failed to reject the null model. Regardless of recombination rate, summary statistics for the distribution of runs did not fall in the extreme tails of results from simulations of coalescence and recombination (Supporting Information Table 6). The X chromosome was characterized by long runs of fixed differences in all three pairwise comparisons (Supporting Information Figures 1a, b, c).

References

McDonald J. H., 1996 Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol Biol Evol 13: 253–260.