Supplementary Methods for *In silico phylogenomics using complete genomes: a case study on the evolution of hominoids*

Igor Rodrigues Costa, Francisco Prosdocimi, and W. Bryan Jennings

Estimation of the number of Independent Genealogical Units and distance thresholds

In order to estimate an appropriate distance threshold (in bp) separating sampled anonymous loci from annotated genome elements and between sampled loci, we estimated the average physical distances (in bp) between independent loci using available genome and population genetic data. First, we estimated the total number of independent loci in the human genome. Using the calculation by Hudson and Coyne (2002), we estimated the total number of independent genealogical units or "IGUs" with the formula:

$$\text{IGUs} = 4N_e c\ /1000 \quad (1)$$

N_e is the long-term effective population size for humans while c is the per generation recombination rate. For details on how this formula was derived see Hudson and Coyne (2002). To use formula (1) we assumed the genetic map distance for the human genome to be 3,614 cM (Kong et al. 2002) and N_e for humans to be between 7,500 (Tenesa et al. 2007) and 10,000 (Takahata 1993). We next obtained the following two estimates for the number of IGUs in the human genome: [(4)(7,500)(3,614 cM)(0.01 cross-overs per generation)]/1000 = 1,084 and, [(4)(10,000)(3,614 cM)(0.01 cross-overs per generation)]/1000 = 1,446 IGUs. This translates to about one IGU per ~2-3 Mb in the human genome, on average. However, because we are interested in the number of independent loci in common among the human, chimpanzee, gorilla, and orangutan genomes, hereafter referred to as a "hominoid" genome, we must use a long-term N_e that dates back to the hominoid ancestor because we must consider the incidence of recombination throughout the history of the hominoids and not just within the human lineage.

Estimates of the human-chimpanzee ancestral population size (N_{HC}), human-chimpanzee-gorilla ancestor size (N_{HCG}), and human-chimpanzee-gorilla-orangutan ancestor size (N_{HCGO}) vary substantially within and among parameters. Estimates for N_{HC} include: 24,625 (Rannala and Yang 2003), 33,100 (Yang 2002), 27,716-41,263 (Schrago 2014), 47,000 (Hobolth et al. 2011), 99,000 (McVicker et al. 2009), and 99,000-122,000 (Burgess and Yang 2008). Estimates for N_{HCG} include 42,750 (Rannala and Yang 2003), 32,346-51,000 (Schrago 2014), 52,000 (McVicker et al. 2009), and 55,000-66,000 (Burgess and Yang 2008). Estimates of N_{HCGO} include: 24,750 (Rannala and Yang 2003), 84,000 (McVicker et al. 2009),

85,000-110,000 (Burgess and Yang 2008); 187,000 (Hobolth et al. 2011); and 144,483-272,098 (Schrago 2014). Given this range, we chose an intermediate value of 100,000, which is an order of magnitude larger than the presumed N_e for the human lineage, to be our assumed long-term N_e for hominoids. Using Formula (1), and also assuming a similar genome size and linkage map as for humans, we estimate that ~14,000 IGUs exist in the hominoid genome, which translates to one IGU per ~200 kb, on average.

Supplementary References

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al. 2002. A high-resolution recombination map of the human genome. *Nature Genet* **31**: 241-247.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res* **17**: 520-526.

Takahata N. 1993. Allelic genealogy and human evolution. *Mol Biol Evol* **10**: 2-22.