FREGENE datasets: 'Population C'

SUMMARY OF MODELLING ASSUMPTIONS USED TO SIMULATE WORLDWIDE HUMAN POPULATION.

Population C mimics the principle features of genetic variation in major worldwide human populations. This simulation required the following steps:

- **Founding population in Africa:** an homogeneous population (N=25K sequences) evolves for 125K generations.
- **Expansion in Africa:** the population expands from 25K to 48K sequences and evolves during 17K further generations.
- **Out of Africa (OoA) split and bottleneck:** among the 48K sequences, 8.5% (= 4,080) leave Africa. Simultaneously, the population in Africa encounters a bottleneck of size ratio 0.8%, leaving 380 sequences in that subpopulation.
- **African and OoA expansion:** African population expands back to N=48K. Similarly the OoA population expands to N=15.4K and evolves for 3.5K generations.
- **Asian and European split:** the OoA population encounters a bottleneck of size N=1,360, and splits with N=320 moving to Europe and N=1,040 to Asia.
- **Asian and European expansion:** Asian and European populations both expand to N=15.4K, and evolve for 2K generations. During this stage, migration occurs symmetrically, first between Asia and Africa (with rate 0.8×10^{-5} per chromosome), and between Europe and Africa (with rate 3.2×10^{-5} per chromosome).
- **Independent evolution of the three populations:** African, Asian and European populations evolve, without migration, during 200, 400 and 350 generations respectively, while each population expands to reach a final population size of N=50K sequences.

Each sequence is 10 *Mb* long. Two simulations are available: one neutral (see results in `Output_neutral`) and one with selection (see results in `Output_selected`). Scripts used to generate these simulations are also provided (`fregene_POPNC_neutral.sh` and `fregene_POPNC_sel.sh` respectively). The population at the end of each step is recorded in the `Input` directory (see `rin_*.xml`). Modelling assumptions are recorded in the input files (in `Input` directory) and summarized below. Finally one `R` script that generate the evolution of diversity along generations can be found in `R_Scripts` together with the corresponding figure (in `Figures`).

General parameters	
Chromosome Length	$10\ Mb$
Per-site mutation rate	1.5×10^{-8}
Recombination model	
Per site crossover rate:	1.1×10^{-8}
Per site Gene conversion rate:	4.5×10^{-9}
Proportion of recombination events occurring in hotspots	80%
Hotspot length:	$2.0kb$
Gene conversion length:	$0.5kb$
Mean distance between hotspots:	$8.5kb$
Selection parameters (if applicable)	
Prop. of sites under selection:	5×10^{-4}
Proportion of selected sites locally under selection:	0.5
Mean # generations before selected sites are switched off	$50,000$
selection coefficient:	$s \sim 0.1 \times \mathcal{N}(0.005, 0.05^2)$ $+0.9 \times \mathcal{N}(-0.01, 0.005^2)$
dominance coefficient:	$h \sim 0.8 \times \mathcal{N}(0.5, 0.2)$ $+0.3 \times \mathcal{N}(1.2, 0.2^2)$