**SUMMARY OF THE PROJECT**

Historians, archaeologists and geneticists have been trying for a long time to answer how the spread of farming has occurred. Currently, there are two main competing hypothesis that can potentially explain it: (1) farming was transmitted by the diffusion of ideas and technology between societies (Cultural Diffusion); or, alternatively, (2) by population expansion of the first farmers that subsequent displaced foraging societies (Demic Diffusion). So far, the scarcity of data and the limited application of quantitative analysis have restricted the conclusions to local and regional scales. Here, we propose that the two process (Cultural and Demic diffusion) would lead to different patterns of language evolution and that such differences could be captured by modern phylogenetic methods applied to phylogenies generated from comparative linguistics. We expect that the spread of farming through Demic Diffusion would lead to more speciation of farming societies (new societies driven by expansion and conquering), more societies extinctions of foragers (caused by conflicts) and an aggregation of farming technology among societies that descend from the first farmers. By the other side, the cultural diffusion of farming would represent an adaptation of already established societies, only decreasing the general extinction rates. Assuming that the selection of farming should be at the society level, we aim to generate a society based modelling simulation to first confirm if there could be differences among phylogenetic trees exposed to different cultural spread modes (Cultural and Demic diffusion), as we suspect, and what would be such differences. After, we plan to compare the modelling results with the real world linguistic phylogeny to indicate how the farming technology was spread among human societies. The results of this work has the potential to answer an important question regarding human history, that after all could have been always there, hidden in the relationship of our languages altogether.

**THE SIMULATION**

***Rules***

* In the simulation, a society can only occupy one location. The locations are determined by the occurrence points in the real dataset (*N*) [Note that any set of occurrence points is accepted to make the model flexible]. Hence, the maximum number of societies is restricted by the number of locations.
* The locations are connected by the 7 closest neighbors that keep all the connections symmetric. Previous research shows that his number represent better the overall connection among societies in this specific dataset.
* All the societies have one trait associated, either foraging or farming.
* Each location has one binary value indicating whether a society can farm or not (environmental conditions), determined by previous studies using real data.
* The model starts with one foraging society. The location of the starting society can be pre-determined by the user or randomly assigned by the model.
* The total number of time steps (*T*) in which a simulation will run is defined *a priori* by the user.
* Every society has a chance to expand into un-colonized neighbor locations at each time step. Because societies can only occupy one location, a speciation occurs when a society expand into a new location, and a new society is created (Speciation module). The decedent society will keep the same trait of the parent society.
* Every society has a chance to go extinct at each time step (Extinction module).
* Every society has a chance to change their trait from foraging to farming at every time step (Arisal module).
* Every society has a chance to diffuse their trait to other neighbor society at every time step (Diffusion module).
* Every society has a chance to takeover a neighbor society (with speciation of the source society and extinction of the target society) at every time step (Takeover module). Takeover only happens when all neighbor cells are occupied [Note that because the Speciation module requires empty neighbors, while Takeover only occurs in the absence of them, only one of these two modules will happen at each time step per society. This avoids an artificial inflation of the speciation rate with Takeover]. Farming societies will always prefer to invade locations suitable for farming if they are available.
* All the parameters involving the probabilities related to each action module are decided at the beginning of the simulation by the user.
* Any module can be turned off at the simulation by setting its probabilities to zero.
* In locations that are not suitable for farming, it is *X* times harder to diffuse farming. It is also *X* times harder to diffuse foraging to farming societies that are in suitable locations for farming. Farming societies have *X* times higher probability of takeover and resistance from takeover when in suitable locations for farming. Being *X* defined *a priori* by the user.
* All the modules will happen at every step. The modules order of action is always shuffled at the beginning of every time step.
* Inside the modules, the actions will involve all existing societies at a specific time step (*Ni*). However, for the Speciation, Takeover and Diffusion modules, the actions cannot happen for all the societies at the same time, because what happen to one society may affect the behavior of other societies in the same module. So, the simulations inside these modules happen sequentially among societies with sub time steps (*Tsub = 1 / Ni*), such sequence is decided by shuffling the order of societies at the begging of each module. Societies that got extinct in the Takeover module are excluded from the sequence.
* During the simulation, all the evolutionary relationships are tracked using a phylogenetic tree.
* The branch lengths (*BL*) correspond to the sum of time steps from the origin of a node (*k*) or tip to the current time of the simulation (*BLk = Ti - Tk*).
* During a speciation process (through the Speciation or Takeover module) one society will bifurcate into two new societies in the phylogenetic tree (so that the tree is always binary).
* All societies extinctions (through the Extinction or Takeover module) will be pruned from the phylogenetic tree.
* At the end of each time step, all the tip lengths are extended by adding the current time step (so that the tree is always ultrametric).

***Output***

* The simulation will output two files, a table including the spatial data and a phylogenetic tree.
* The spatial table, will include for each location their ID, latitude, longitude and environmental conditions. If a society is occupying the location at the time *T*, the table will also include the society trait and tip label (corresponding to the phylogenetic tree).
* The phylogenetic tree will include all existing societies at the time *T* (no extinct society will be included). The final phylogenetic tree shall be binary and ultrametric.

***Experiment design***

* The simulations will be running over 4 different scenarios that include different modules combination: 1) Speciation, Extinction and Arisal; 2) Speciation, Extinction, Arisal and Diffusion; 3) Speciation, Extinction, Arisal and Takeover; 4) Speciation, Extinction, Arisal, Diffusion and Takeover.
* For each scenario, a series of phylogenetic and spatial metrics will be calculated.
* We will choose the parameters of each of the scenarios using an Approximate Computational Bayesian (ABC) approach that maximize the likelihood of the simulation phylogenetic and spatial metrics to the metrics calculated for the real dataset.
* Each of the four calibrated scenarios will be repeat 1000 times. This will allow us to identify the key metrics that change among the scenarios and what scenario best fits the real dataset.

***Voilá!***