Revisiting an old evolutionary question: Did the S mutation of the β-globin gene result from a single or multiple mutations?

Sandra Raquel da Silva Oliveira

Biodiversidade, Genética e Evolução

Departamento de Biologia

2012

Orientador

Jorge Rocha, PhD, Faculdade de Ciências da Universidade do Porto

Co-orientador

António Santos, PhD, Faculdade de Ciências da Universidade do Porto

Todas as correções determinadas
pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Agradecimentos

Ao Professor Doutor Jorge Rocha pela incansável orientação científica e por toda a sabedoria transmitida;

Ao Professor Doutor António Múrias dos Santos pelo valioso apoio ao longo de toda a tese, em especial na parte informática;

Ao Professor Doutor Diogo Meyer;

Ao Rui, à Bárbara, à Catarina, à Filipa, aos meus amigos e colegas de mestrado;

À minha família;

Ao André.

Table of contents

Abstract

The S mutation of the β-globin gene (*HBB*S*) is a well-studied example of an advantageous allele with a protective role against malaria in the heterozygous carrier state. More than 50 years ago, Livingstone proposed that the emergence of tropical agriculture provided ideal habitats for the spread of malaria-transmitting mosquitoes allowing the rapid diffusion of a single, relatively recent *HBB*S* mutation. However, the concept of a single mutation was challenged by the finding that different *HBB*S*-linked haplotypes predominated in various non-overlapping geographical regions of Africa, India and the Arabian Peninsula. Currently, the most favored hypothesis explaining the geographic segregation of *HBB*S*-linked haplotypes is that *HBB*S* variants originated independently by recurrent mutation in each region where a single haplotype predominates. However, little work has been done to examine the effects of the spread of the *HBB*S* allele on linked haplotype variation in spatially explicit evolutionary settings. Here, we explored a computer simulation framework to assess the evolution of *HBB*S*-linked haplotype variation in time and space, using a *wave of advance* model for the dispersal of an advantageous allele. We found that even assuming a single origin for the *HBB*S* variant, it is possible to observe remarkably low levels of haplotype heterogeneity at the edges of variant distribution after as much as 200 generations. Moreover, we demonstrate that the wave of advance of the S allele can mimic the spatial distribution of different S-haplotypes in Africa by creating several patches (or sectors) formed by contiguous populations sharing S-linked modal haplotypes that are different from those observed elsewhere. The comparison of different simulated scenarios with an empirical dataset, consisting of haplotype data from different African populations, additionally showed that the overall levels of haplotype variation that are generated around each center of origin of a *HBB*S* mutation are too high to be compatible with the predictions of the multicentric hypotheses, unless the age of S mutations is unrealistically low (about 60 generations). Although our preliminary study allowed the evaluation of the relative consistency of different evolutionary scenarios, our limited set of simulation conditions were still unable to match all the characteristics of the observed data with sufficient approximation. Thus, the simulated framework that was developed in this work should be used in the future to evaluate a more comprehensive set of demographic alternatives to provide a more robust discrimination of competing evolutionary hypotheses.

Resumo

A mutação S do gene da β-globina (*HBB*S*) é um exemplo bem estudado de como um alelo favorável pode ter uma função protectora contra a malária em portadores heterozigóticos. Há mais de 50 anos, Livingstone sugeriu que o desenvolvimento da agricultura tropical criou as condições ideais para a disseminação dos mosquitos vectores da malária, permitindo a rápida difusão de uma mutação *HBB*S* de origem única relativamente recente. No entanto, o conceito de origem única foi questionado pela descoberta de que diferentes haplótipos ligados à variante *HBB*S* predominam em várias regiões não sobreponíveis de África, da Índia e da Península Arábica. Actualmente, a hipótese mais aceite para explicar a distribuição geográfica dos haplótipos de *HBB*S* sugere que a cada haplótipo predominante numa dada região corresponde uma mutação recorrente. Apesar desta aceitação, não há trabalhos suficientemente detalhados sobre os efeitos da difusão do alelo *HBB*S* na variação haplotípica que lhe está associada, usando cenários evolutivos espacialmente explícitos. Neste trabalho, procurámos explorar um método de simulação que permita estudar a evolução da variação haplotípica no tempo e no espaço, usando um modelo de avanço em onda para a difusão de uma mutação favorável. Com esta abordagem, foi possível mostrar que, mesmo assumindo uma origem única, se podem observar níveis muito baixos de heterogeneidade haplotípica na periferia da distribuição da variante, ao fim de 200 gerações. Adicionalmente, verificou-se que a onda de avanço da variante favorecida pode originar sectores geograficamente delimitados em que predominam diferentes haplótipos, à semelhança dos padrões haplotípicos observados em África. A comparação de diferentes cenários simulados com dados empíricos recolhidos em várias populações africanas, permitiu ainda mostrar que os níveis de variação haplotípica gerados em torno de um determinado centro de origem são demasiado elevados para serem compatíveis com as previsões de modelos multicêntricos, a menos que se assuma que as mutações recorrentes *HBB*S* tiveram origens irrealisticamente recentes (há cerca de 60 gerações). Apesar deste trabalho ter permitido avaliar a consistência relativa de diferentes cenários evolutivos, não foi possível obter, com o número limitado de condições demográficas usadas em várias simulações, uma aproximação suficientemente satisfatória a todas as características dos dados observados. Assim, o enquadramento metodológico agora desenvolvido deverá ser usado no futuro para avaliar um conjunto mais vasto de modelos demográficos a fim de obter uma discriminação mais robusta das hipóteses alternativas.

1. Introduction

Elucidating the factors that shaped the geographical spread of advantageous alleles is fundamental to understand the current distribution of adaptive traits, including resistance to infections (Epperson, 2003). Recent studies have paved the way for a better understanding of these factors by examining different models of spatial dynamics of advantageous alleles, while accounting for allele interactions (Ralph and Coop, 2010) or heterogeneous selection intensity (Novembre et al., 2005).

The S mutation of the β-globin gene (*HBB*S*) is a classic example of a pathogenic genetic variant that can be beneficial in some circumstances (Flint et al., 1993), providing an excellent model for studying natural selection in a spatially explicit approach. Homozygotes for the *HBB*S* allele (SS) suffer from a severe disease, called sickle-cell anemia, which is usually lethal before the age of 5 in the absence of medical care (Piel et al., 2010). However, the *HBB*S* mutation is not pathogenic in heterozygotes (AS), and may confer a protective effect against malaria to its carriers (Allison, 1964). It was the heterozygote advantage over both AA and SS homozygotes (overdominance) that prevented this mutation to be eliminated and allowed its spread to several regions of Africa, Arabian Peninsula and India (Livingstone, 1958; Livingstone, 1964). In West and Central Africa the *HBB*S* gene is especially common (Figure 1), reaching frequencies of 16% (Cavalli-Sforza and Bodmer, 1971; Piel et al., 2010) or even 20% (Livingstone, 1958; Flint et al., 1998). Very high frequencies are also known to occur in the Qatif oases of eastern Saudi Arabia and parts of India (Flint et al. 1998). The allele is also found at lower to moderate frequencies in other parts of the old world, including some Mediterranean areas (Greece and Sicily) (Cavalli-Sforza and Bodmer, 1971; Piel et al., 2010).

Haldane was the first to establish a link between the high frequencies of genetic blood disorders and the selective advantage conferred by protection against *Plasmodium falciparum* infection (Haldane, 1949). His suggestion was based on the correlation between the distribution of the hemoglobinopathies and the historic incidence of malaria (Figure 1). Ever since, much evidence has accumulated in support of the *HBB*S* protective advantage (Beet, 1946; Allison, 1954; Allison 1964; Williams, 2006, Weatherall, 2008) and HBB AS individuals have consistently been shown to enjoy more than 90% protection against severe and lethal malaria and 50% protection against mild clinical attacks (Williams et al., 2005).

According to Livingstone, the *HBB*S* diffusion has only been possible after the adoption of agriculture in the last few thousand years (Livingstone, 1958). Agriculture and forest-clearing created the ideal habitat for malaria-transmitting mosquitoes (*Anopheles*), increasing the risk of infection in agricultural communities and leading to the rapid spread of *HBB*S* in areas infested with malaria (Livingstone, 1958; Carter and Mendis, 2002). The bio-cultural model of Livingstone is based on the assumption that the *HBB*S* mutation has a single origin and became frequent with the emergence of tropical agriculture, about 5000 years ago. This assumption is apparently supported

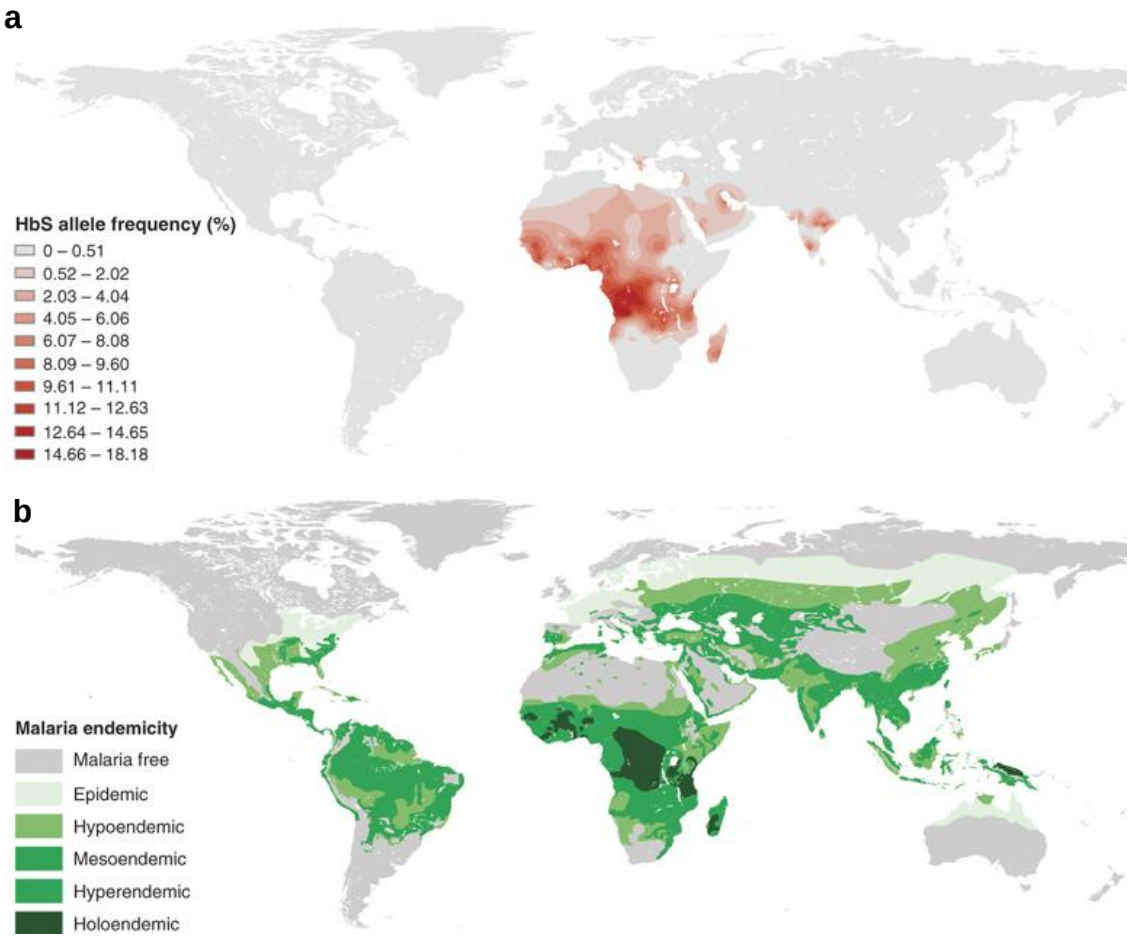

Figure 1 - a) *HBB*S* allele frequency map. b) Historical map of malaria endemicity. Withdrawl from Piel et al., 2010.

by the positive correlation between *HBB*S* frequencies and the oldest agricultural traditions, and also by the demonstration that the *HBB*S* allele could have diffused from a single place of origin throughout its present distribution area in about 5000 years (Livingstone, 1989).

The idea of a single *HBB*S* mutation has been challenged when the first results on the haplotype variation associated with the *HBB*S* variant became known. Using 5 restriction fragment length polymorphisms (RFLPs) it was shown that this allele is associated with five major haplotypes that predominate in non-overlapping geographical regions (Antonarakis, 1984; Pagnier et al., 1984; Nagel et al., 1985; Lapouméroulie et al., 1992). Four major haplotypes are restricted to Africa, and are named Senegal, Benin, Bantu and Cameroon; a fifth haplotype, named Arab-Indian, is found in the Arabian Peninsula and India (Figure 2a). This spatial pattern, together with the high levels of molecular divergence among the different predominant *HBB*S*-linked haplotypes has led to the suggestion that *HBB*S* variants originated independently by recurrent mutation in each region where a single haplotype predominates (Figure 2b) (Pagnier et al., 1984; Flint et al 1998; Nagel and Ranney 1990). This hypothesis can, in

fact, be traced back to Kurnit (1979), who proposed a multicentric hypothesis for the origin of *HBB*S* on the basis of initial insights about haplotype variation obtained with a single linked *Hpa*I restriction site polymorphism.

However, the low likelihood that at least five independent mutations recurred in a short period of time in different lineages was frequently used to argue that the single origin hypothesis was still preferable to a multicentric origin (Fullerton et al., 1994). According to this reasoning, the observed territorial segregation of *HBB*S* haplotypes could be better explained by the appearance of new haplotype associations through interallelic recombination and gene conversion, which could have spread a single mutation across multiple chromosomal backgrounds (Flint et al., 1998; Powers and Smithies, 1986). This interpretation seemed to be favored by the finding of a recombination hotspot immediately 5' to the *HBB*S* gene and 3' to the δ gene (Chakravarti et al., 1984). Moreover, Livingstone (1989) outlined a spatial diffusion framework in which different haplotypes arising from a single *HBB*S* mutation could become geographically segregated. According to this framework, when an allele is introduced into a population with endemic falciparum malaria, its frequency will rapidly increase by selection (Livingstone, 1989). Since the diffusion of a highly advantageous allele usually results from the introduction of a few migrants into new territories, it is reasonable to think that a single *HBB*S* haplotype will be present in the front of the spreading area (Livingstone, 1989; Excoffier and Ray, 2008). As stressed by Livingstone (1989), a major implication of this hypothesis is that the regions where a single major haplotype predominates are peripheral areas where the *HBB*S* allele, originating elsewhere, has been recently introduced by gene flow. However, in spite of describing preliminary simulation analyses, Livingstone (1989) did not provide a thorough quantitative examination of this scenario.

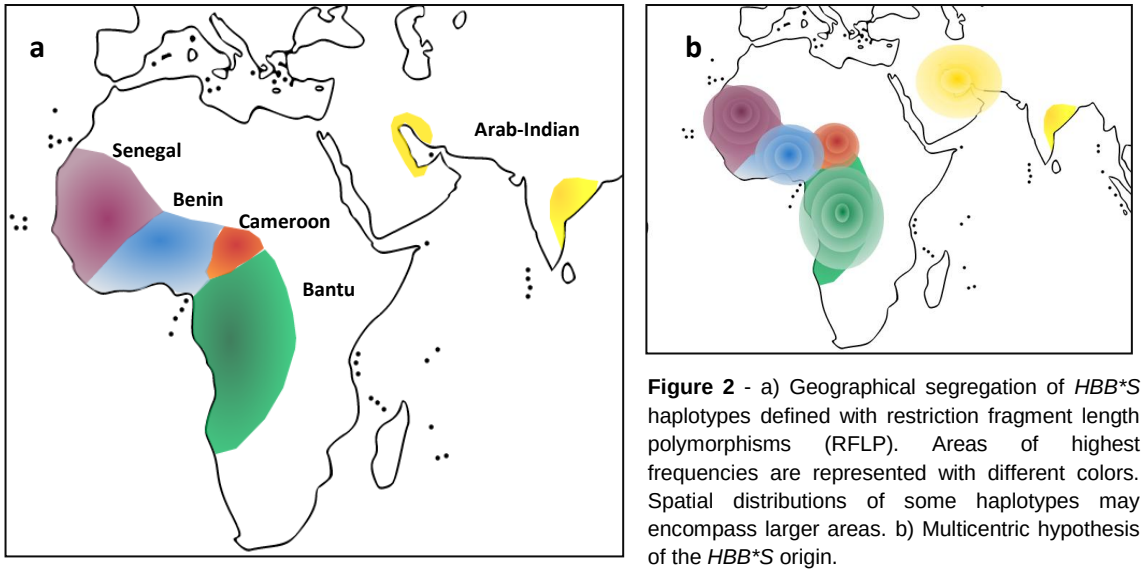

Figure 2 - a) Geographical segregation of *HBB*S* haplotypes defined with restriction fragment length polymorphisms (RFLP). Areas of highest frequencies are represented with different colors. Spatial distributions of some haplotypes may encompass larger areas. b) Multicentric hypothesis of the *HBB*S* origin.

In contrast, the multicentric hypothesis has been strengthened with recent evidence indicating that the ability of very large populations to adapt under strong selection is not mutation-limited. For example, by studying insecticide resistance in *Drosophila melanogaster*, Karasov et al. (2010) provided an elegant demonstration that the same mutations can arise independently several times on different local haplotypes in a short period of time, as long as the population size is large enough to counteract the effects genetic drift (Karasov et al., 2010). More recently, Ralph and Coop (2010), by means of analytical models and simulations, have shown that patterns whereas recurrent mutations predominate in different spatial areas (geographic parallel adaptation) are not unlikely in the presence of spatial structure and homogeneous selective regimes, provided that the selective advantage is high and population sizes are large. In those settings, assuming that each recurrent mutation spreads as an expanding wave with constant speed, the likelihood of geographic parallel adaptation critically depends on the distance reached by a selected allele before it encounters other successful allele arising by recurrent mutation in a different area (also called *characteristic distance*) (Ralph and Coop 2010). Importantly, the authors have shown that, using realistic parameter combinations, the characteristic distance associated with the diffusion of *HBB*S* could be sufficiently low for geographic parallel adaptation to be a likely explanation for the spatial patterns of *HBB*S*-linked haplotype segregation. However, this study focused exclusively on the spatial reach of putatively different *HBB*S* mutations without addressing their linked haplotype variation, which is a critical aspect to assess the likelihood of different models.

The purpose of this thesis is to explore a computer simulation framework that can be used to study the effects of the spread of the *HBB*S* allele on linked haplotype variation, in order to ultimately assess to what extent the current patterns of geographic segregation of *HBB*S* haplotypes are consistent with a single mutational origin, or are better explained by parallel adaptation. Our approach consisted in first describing the basic properties of simulated haplotype variation, and then comparing the outputs of different simulated scenarios with observed data on extended haplotype variation linked to *HBB*S* alleles from various regions of Africa. Although, due to computation time restrictions, the data presented here are still preliminary and not sufficiently detailed to discriminate between alternative hypotheses, we are confident that our attempt provides useful insights into the geographical spread of *HBB*S* and other adaptive traits, as well as into the interdependence between spatial structure and evolutionary processes.

2. Material and Methods

2.1 The empirical dataset

The empirical dataset to which simulated data were compared consists of 330 high resolution *HBB*S*-linked haplotypes, available at the *Human Evolutionary Genetics* group at CIBIO (unpublished), which were sampled from 35 populations in eight different African countries (Table 1). The *HBB*S* bearing chromosomes were initially genotyped at four polymorphic restriction sites (*HincII 3', HindIIIγG, HincIIΨβ, HindIIIγA*) in order to identify the four classical RFLP–defined haplotypes associated with each *HBB*S* allele in Africa (Bantu, Benin, Senegal and Cameroon). After this initial classification, the haplotypes were additionally characterized with 11 microsatellites distributed across a 525 kb region encompassing the HBB gene, which is 10-fold larger than the one including the RFLP markers (Figure 3). Haplotypes were inferred from genotype data using the Bayesian approach implemented in the Phase v2.1 software (Stephens et al., 2001).

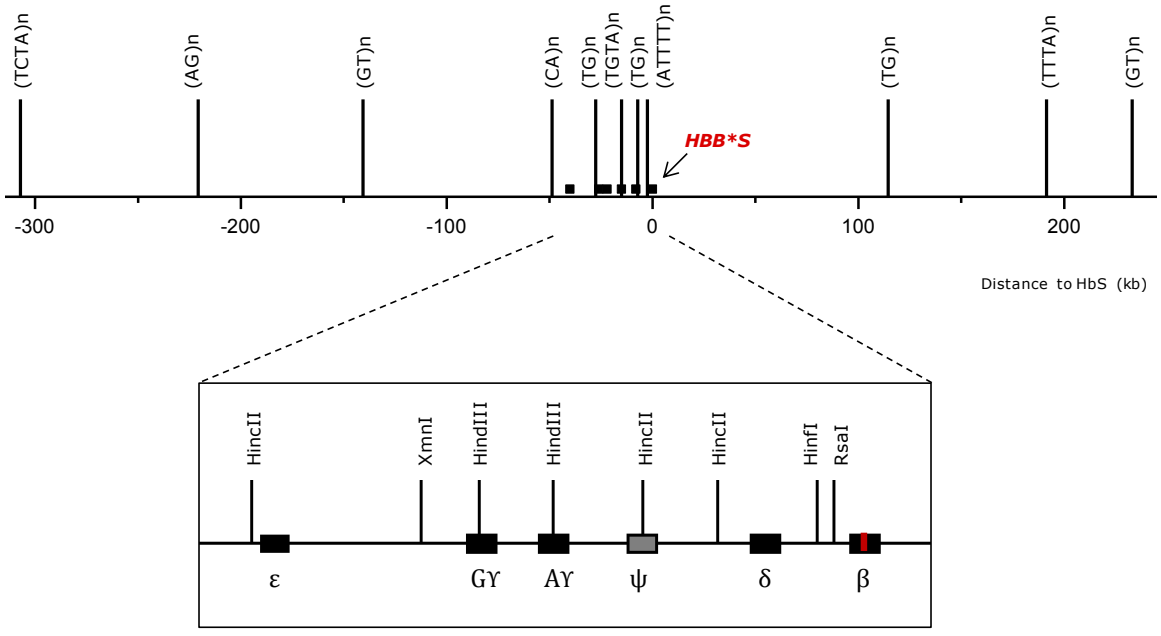

Figure 3 – Approximate location of microsatellites used in the characterization of haplotype diversity linked to the *HBB*S* mutation. Genes of the β-globin family are represented by black squares. The position of the *HBB*S* mutation is indicated with an arrow.

Table 1 – Provenance of samples used in the study of the *HBB*S* variant.

Region	Country	Ethnic group
West Africa	Cape-verde	a)
Central Africa	Cameroon	Baka Pygmies
		Mbenzele Pygmies
		Beti/Ewondo
		Bamileke
		Bassa
		Fali
		Foulbe
		Ngoumba
		Sanga
		Podokwo
		Uldeme
	Central African Republic	Sanga
	São Tomé	a)
East Africa	Sudan	b)
	Kenya	Rendille
		Luo
Southeast Africa	Mozambique	Chewa
		Chibarué
		Chuabo
		Coti
		Lomwe
		Macua
		Marenje
		Sena
		Yao
Southwest	Angola	Kimbundu
		Herero/Kuvale
		Kinkongo
		Nyaneka-Humbe
		Kioko
		Umbumdu
		Cuanhama
		Muhumbi
		Mumuíla

a) The island of Cape Verde and São Tomé were peopled by slaves captured in adjacent areas of Africa and may be regarded as typical sink populations. Although the original ethnicities and spatial context have been blurred during the peopling of Cape Verde and São Tomé, the islands provide the opportunity to collect very divergent haplotypes in relatively limited geographic areas.
b) Samples obtained in hospital environment without specification of ethic group.

In cases where a given *HBB*S* haplotype was collected in populations formed by relatively recent migration, we treated that haplotype as belonging to the African mainland area from where it likely originated. For example, Bantu and Benin haplotypes sampled in São Tomé were assigned to Congo/Angola and Benin/Nigeria regions, respectively, because they are likely to have been carried into the island by slaves recruited in those areas (Tomás et al., 2002). Likewise, Senegal haplotypes sampled in Cape Verde were considered to have originated from the Senegambia region, where most slaves that colonized the archipelago were captured (Curtin, 1969).

In the extended haplotype dataset, each classical, RFLP-defined haplotype constitutes a population of chromosomes with a modal haplotype that is distributed across a considerably large area, and a series of less frequent haplotypes having more restricted distributions, which probably arose from recombination and/or mutation at the microsatellite markers. For example, as illustrated in Figure 4, most classical Bantu haplotypes sampled in Central-West Africa (São Tomé, Angola, Cameroon) share the same extended haplotype consisting of 11 microsatellite loci. Interestingly, the predominant Bantu haplotype in samples from Mozambique, differs from the modal Central-West African haplotype at the 5' and 3' ends of the studied chromosomal region, revealing a previously undetected local heterogeneity within Bantu range (Figure 4). Thus, we decided to name the two subtypes as Bantu-West and Bantu-East, and treat them separately in data analyses.

In addition to the Bantu, Cameroon, Benin and Senegal haplotypes, we found that different Afro-Asiatic speaking populations from northern Cameroon shared an atypical RFLP profile corresponding to a different modal extended haplotype that is not found in other populations. We named this haplotype Afro-Asiatic, although the low number of *HBB*S*-bearing chromosomes sampled in this region (n=14) is not enough to know the full range of its distribution.

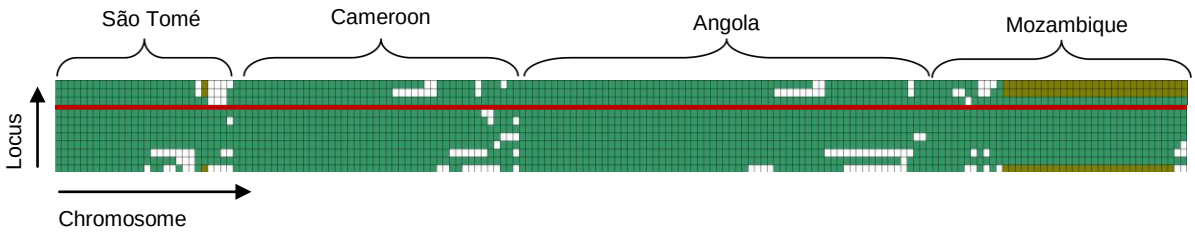

Figure 4 – Schematic representation of haplotype variation within the Bantu haplotype in four different populations. Each column represents one chromosome and each line one microsatellite. The modal haplotype is represented in green. Microsatellites that have alleles other than the modal haplotype are represented in white or brown. The position of the *HBB*S* mutation is indicated in red.

All together, we divided the extended haplotype data into six major groups (Cameroon, Benin, Senegal, Bantu-West, Bantu-East and Afro-Asiatic), whose frequencies in different sampled countries are shown in Figure 5. Figure 6 depicts the frequencies of the different haplotypes belonging to each major group. Note that, with the exception of the Cameroon type, which has a flattened distribution, most groups have a single predominant haplotype that is much more frequent than the other haplotypes.

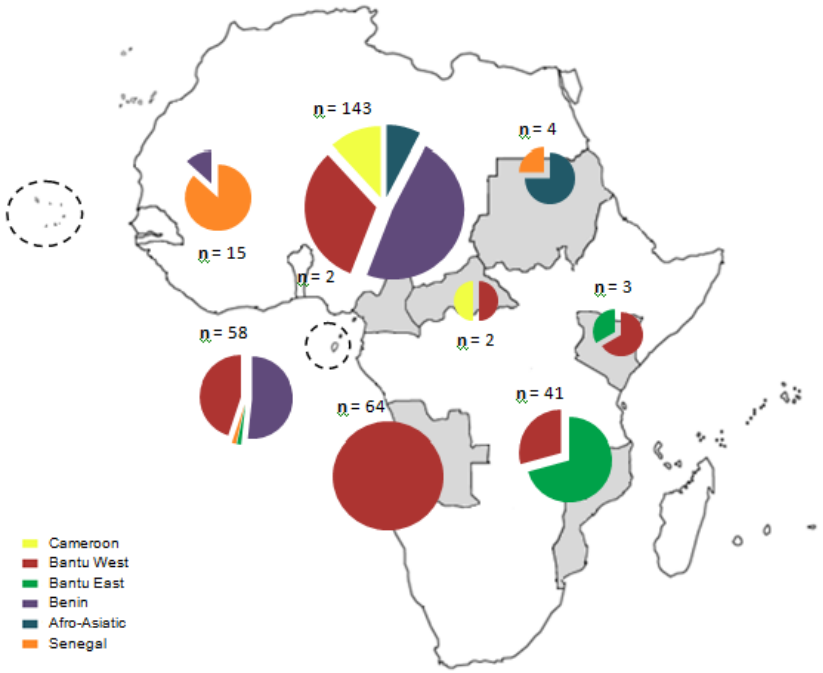

Figure 5 – Haplotype frequencies of the six major groups from the empirical dataset. Sampled countries are depicted in grey, with the exception of Cape Verde and São Tomé that are highlighted with a dashed circle. The number of samples (n) obtained in each country is shown together with the pie charts of HBB*S haplotype frequencies.

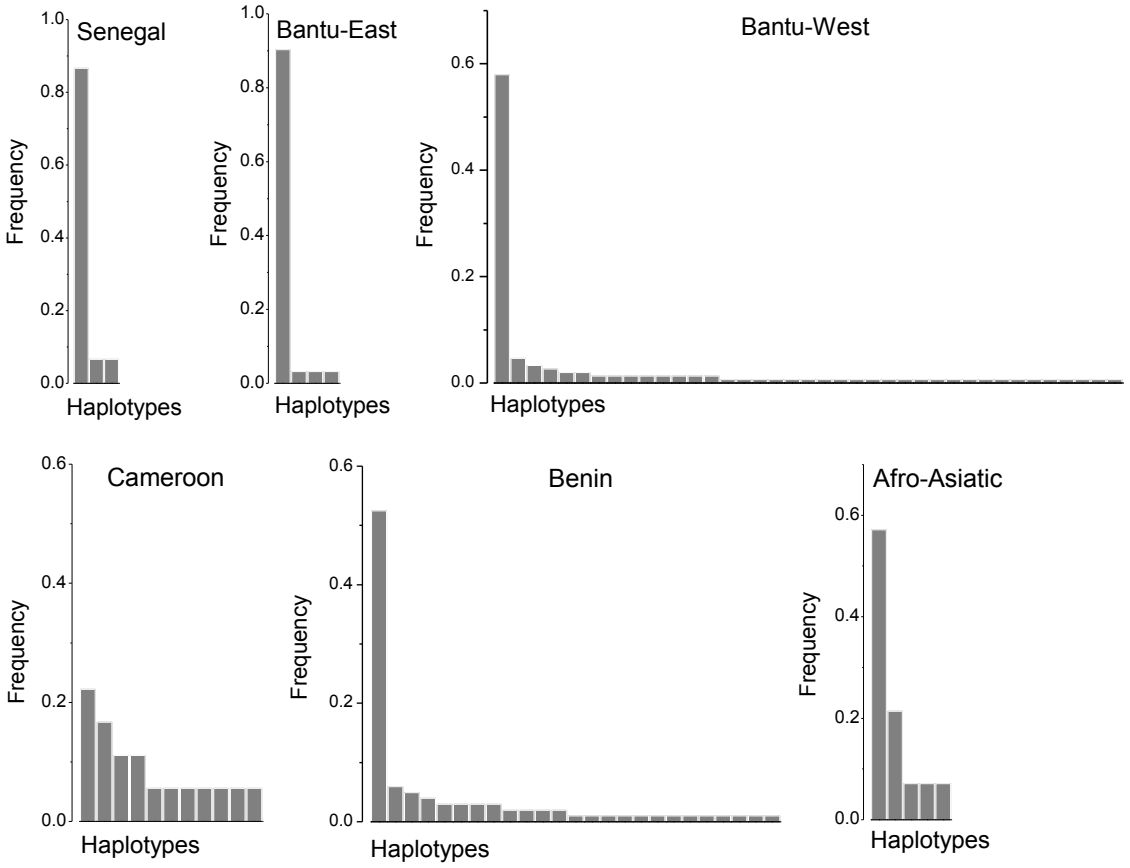

Figure 6 – Frequencies of the different haplotypes found in each major group (Senegal, Bantu-East, Bantu-West, Cameroon, Benin, and Afro-Asiatic).

2.2 Simulation approach

Using a previously developed program (Santos AM, unpublished; appendix 1), we explored a forward-in-time strategy to simulate the diffusion of haplotypes linked to a favorable mutation. The simulation has two main components, a diffusion component representing the increase in frequency of a favorable variant and its dispersal across space, and a diversification component representing the generation of haplotype diversity.

2.2.1 Diffusion of a favorable variant

The classical Wright-Fisher model assumes that populations behave as panmitic units (each individual of the population is equally likely to interact with any other), with finite and constant size and discrete generations (Ewens, 1979). However, natural populations are not panmitic and, for species occupying large areas, individuals are more likely to mate with their geographic neighbors. To account for this lack of randomness in mating across a species range, Wright explored what he called a "isolation by distance" model, in which geographic distance limits the exchange of migrants throughout space (Wright, 1943). Following Wright's work, Kimura and Weiss (1964) introduced the "stepping stone model" of population structure, in which individuals are distributed into discontinuous populations (Kimura and Weiss, 1964). In this model populations are arranged on a grid and migration takes place between neighboring populations; i.e. in each generation an individual can migrate at most one step between populations.

Here, we simulated the diffusion of haploid individuals across one-dimensional or two-dimensional grids using a stepping stone model. Each cell of the grid is hereafter referred to as "population" or "deme", interchangeably. Isotropic migration occurs between nearest-neighbors in each generation and the size of each deme, *N,* remains constant. In this haploid model, haplotypes (hereafter also referred to as "individuals") that will form the next generation remain in the parental population with probability 1-*m* and migrate to any of the *n* adjacent demes with probability *m/n*. The number of adjacent demes is 1 or 2 for one-dimensional simulations and 2, 3 or 4 for two-dimensional simulations, depending on the position of the cell in the grid (Figure 7). Grid boundaries are reflective, which means that potential migrants in marginal populations are not lost by migration outside the grid.

Since selection at β-globin gene is overdominant, the relative fitness of AS heterozygotes was set to 100% (W_{AS}=1) and the fitness of AA (W_{AA}) and SS (W_{SS}) types were parameterized as 1-s_1 and 1-s_2, with s_1 and s_2 representing the selection coefficients against the AA and SS genotypes, respectively.

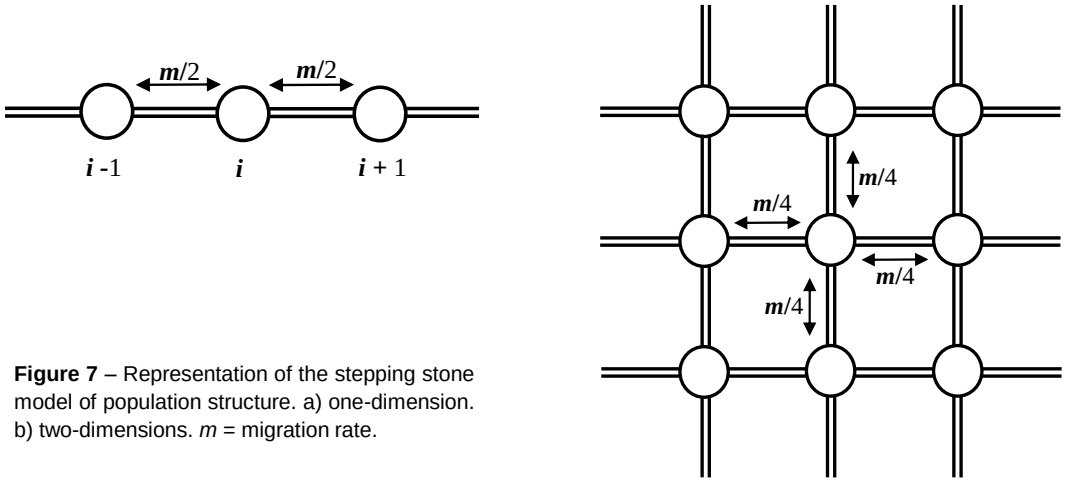

Figure 7 – Representation of the stepping stone model of population structure. a) one-dimension. b) two-dimensions. *m* = migration rate.

2.2.2 Evolution of haplotype diversity

The simulated haplotypes were designed to match the characteristics of the haplotype system used to analyze the β-globin haplotype variation in the empirical dataset (Figure 3). Haplotypes are encoded in a vector of 12 bytes, with positions 1-8 representing eight microsatellites 5' of the β-globin gene, position 9 representing the β-globin gene itself, and position 10-12 representing three microsatellites 3' of the β-globin gene. Except for position 9 (which is used to identify mutant and non-mutant haplotypes by holding a character value of "A" or "S"), all other positions (bytes), representing microsatellite loci, can have a maximum of 256 different allele states. To establish the initial assemblage of haplotypes in each deme, the 9th position of all haplotypes is set to "A". For the A-bearing haplotypes, allele states at each locus are drawn from a multinomial distribution computed from the observed data. At the microsatellite loci, the maximum number of alleles was set to be equal to that observed in the data.

At the beginning of the simulation the population of haplotypes carrying the ancestral "A" allele may have a facultative burn-in period of mixing, drifting and purging. Then, a single mutant allele is inserted at just one deme, by randomly selecting a haplotype and changing its 9th position to "S". If this mutant allele goes extinct the simulation restarts, and this process is repeated until a predefined number of generations is reached without loss of the original mutant allele or any of its derivatives. Hence, all results are conditioned on the non-extinction of the mutant allele.

As the mutant *S* allele increases its frequency due to selection, and disperses due to migration, the haplotype to which the S allele became originally associated can be eroded by mutation at the microsatellite loci, recombination between microsatellites and/or between microsatellites and the β-globin locus, as well as gene conversion at the β-globin locus. In each generation, mutation, recombination and gene conversion are followed by a resampling step where the following generation is formed under the influence of selection. Random haplotypes are sampled with replacement (simulating

an infinite pool of gametes) until *N* is reached. In each population, the number of haplotypes that mutate, recombine or is converted is determined by a Poisson distribution. Note that the events being modeled (mutation, recombination, and gene conversion) are actually the outcomes of discrete trials, and would be more precisely modeled by using the Binomial distribution. However, when the population size is large and the occurrence of a given event is rare (law of rare events), the Binomial distribution can be approximated by a Poisson distribution with significant computational advantages.

Microsatellite mutations were modeled using the symmetric stepwise model, where copy number changes only by one unit, and increases or decreases in copy number are equally likely. The minimum and maximum number of alleles per locus defines a lower and an upper boundary. In such bordering conditions, if the allele is hit by a mutation event, it undergoes a change in the opposite direction.

Gene conversion was modeled as double-recombination.

2.2.3 Simulation parameters

The list of simulation parameters is displayed in table 2. In order to reduce the parameter space used in each simulation, some parameters were directly estimated from the data or from the literature (tables 2 and 3). Due to computation time constrictions, variable parameters were chosen from a limited grid of values that were considered sufficiently disparate to assess their impact on the outcome of simulations.

Table 2 –Parameters of the simulation.

Parameters	*Fixed*[a]	*Variable*[b]
Number of demes		X
Deme size		X
Generations		X
Number of mutants	X	
Original deme (deme where mutation arises)	X	
AA/ AS/ SS fitness values	X	
Microsatellite mutation rates	X	
Recombination rates	X	
Gene conversion rate	X	
Migration rate		X
Generation at which migration starts/stops		X

[a] Parameters that were fixed in simulations
[b] Parameters that were variable in simulations

Microsatellite mutation rate estimates were based on the variance of repeat number in the observed data on the background of the ancestral A allele, which is not affected by selection (table 3). Specifically we assumed that the most variable microsatellite (that with highest repeat number variance) had a mutation rate μ=0.001 (Weber & Wong, 1993), and calculated the mutation rate of the remaining microsatellites according to their variance relatively to the most variable microsatellite. Recombination rates were interpolated from the HapMap recombination map (http://hapmap.ncbi.nlm.nih.gov/) (table 3). Based on Livingstone estimates (Livingstone, 1989), we assumed that AA and SS homozygotes have selective disadvantages of s_1=0.20 and s_2=1 relative to the AS heterozygote, respectively. This values are not very different from those assumed in Currat et al. (2002), s_1= 0.152 and s_2 = 1, or in Cavalli-Sforza and Bodmer (1971, p.150), s_1= 0.15 and s_2= 1.

Table 4 displays the basic set of parameter combinations, based on different deme sizes, *N*, and migration rates, *m*, in variable one- and two-dimensional grids. Each set of parameters (simulations) was replicated 1000 times. In all simulations a single A→S mutation is seeded at the central deme of the geographic grid, hereafter also referred to as the "original deme". One-dimensional simulations were run for maximums of 200 or 300 generations. Two-dimensional simulations were run for a maximum 200 generations. Considering a generation time of 25 years, the maximum number of simulated generations corresponds to a period of 5000 or 7500 years, in agreement with the hypothesis that the malarial selective pressures favoring the spread of the *HBB*S* mutation only became relevant after the expansion of tropical agriculture across Africa (Livingstone, 1989). While older ages for malaria and for the *HBB*S* mutation are entirely possible, and should be evaluated, an increase in the number of generations would have required unreasonable computational times.

Table 3 – Features of microsatellite markers used in simulations of haplotype variation linked to the *HBB*S* mutation.

	STR	Physical position (bp)[a]	Maximum number of alleles[b]	Mutation rate	Recombination rate[c]
5'	(TCTA)n	5,501,925	11	0.000232904	0.002006359
5'	(AG)n	5,416,891	5	0.0000342	0.000512
5'	(GT)n	5,341,113	23	0.001	0.004003
5'	(CA)n	5,254,000	7	0.0000815	0.0000753
5'	(TG)n	5,231,503	17	0.000838	0.0000112
5'	(TGTA)n	5,220,382	4	0.0000253	0.0000944
5'	(TG)n	5,207,520	16	0.000865	0.00000189
5'	(ATTTT)n	5,206,333	5	0.0000389	0.001134
	HBB	5,204,807	-	-	0.002387
3'	(TG)n	5,094,990	17	0.000694	0.000645
3'	(TTTA)n	5,017,841	6	0.000101	0.000203
3'	(GT)n	4,976,041	10	0.000213	-

[a] Data from UCSC Genome web browser (http://genome.ucsc.edu).

[b] Total number of alleles observed in the sampled populations.

[c] Recombination rate between adjacent markers, measured to the right-side of the first marker.

Table 4 – Parameter combinations and grid size

	Simulation	*N*	*M*	*Nm*	Grid
1D	1	1000	0.05	50	101
	2	1000	0.1	100	101
	3	2000	0.05	100	101
2D	4	200	0.05	10	45 x 45
	5	200	0.1	20	70 x 70
	6	500	0.01	5	25 x 25
	7	500	0.05	25	45 x45
	8	500	0.1	50	70 x 70
	9	1000	0.01	10	35 x 35
	10	1000	0.05	50	55 x 55
	11	1000	0.1	100	80 x 80

In addition to the basic set of scenarios displayed in table 4, we explored additional parameter combinations that were designed *ad hoc* to assess the influence of particular features in the outcome of the simulations (table 5). For example in simulation 13 we tested the outcome of beginning to spread the S mutation only 50 generations after its appearance in the original population (table 5). We have also analyzed the effect of the gene conversion rate between the A and S alleles, by assuming that gene conversion was 7.3 times higher than the recombination rate for a mean tract length of 500 bp, as estimated by Frisse et al. (2001) (simulations 12 and 14, in table 5). To implement the Frisse et al (2001) estimate, we divided the gene conversion rate estimated for the 111343 bp track contained between the two microsatellites immediately flanking the β-globin gene and then multiplied it by the number of possible conversions with tract length of 500 bp involving the A and the S alleles.

Table 5 - Parameter combinations, grid size and particular features

	Simulation	*N*	*M*	*Nm*	Grid	GC[a]	Start migration[b]
2D	12	200	0.05	10	45 x 45	0.000115	-
	13	500	0.1	50	70 x 70	-	gen 50
	14	1000	0.1	100	80 x 80	0.000115	-

[a] – Gene conversion rate
[b] – Generation in which migration starts

3. Results and Discussion

3.1 Basic properties of the simulation system

3.1.1 β-globin S diffusion

The spatial distribution of the S allele frequency can be described as concentric growing rings with the highest frequency at the diffusion center. As shown in figure 8, the S frequency increases with time until its equilibrium frequency is reached in each population. The wavelike pattern is caused by the delay in the arrival time of the favored mutation at populations that are located at increasing distances from the center. The simulated equilibrium frequency is in agreement with the predicted equilibrium given by $s_1/(s_1+s_2)=0.167$, a value that seems realistic considering the observed frequencies in Africa (Livingstone, 1989).

Figures 9 and 10 illustrate the speed of the S variant dispersal in one- and two-dimensional population grids, respectively, using different combinations of deme sizes, *N*, and migration rates, *m,* for 25 independent runs (cf. Table 4). As predicted by the theoretical models of Fisher (1937) and Kolmogorov, Petrovskii, and Piscunov (KPP) (Kolmogorov et al. 1937), the spread of the favorable mutation forms a travelling wave of constant speed. However, as noticed in a previous simulation study (Ralph & Coop, 2010), the constant state is preceded by a brief period of variable speed. Comparisons

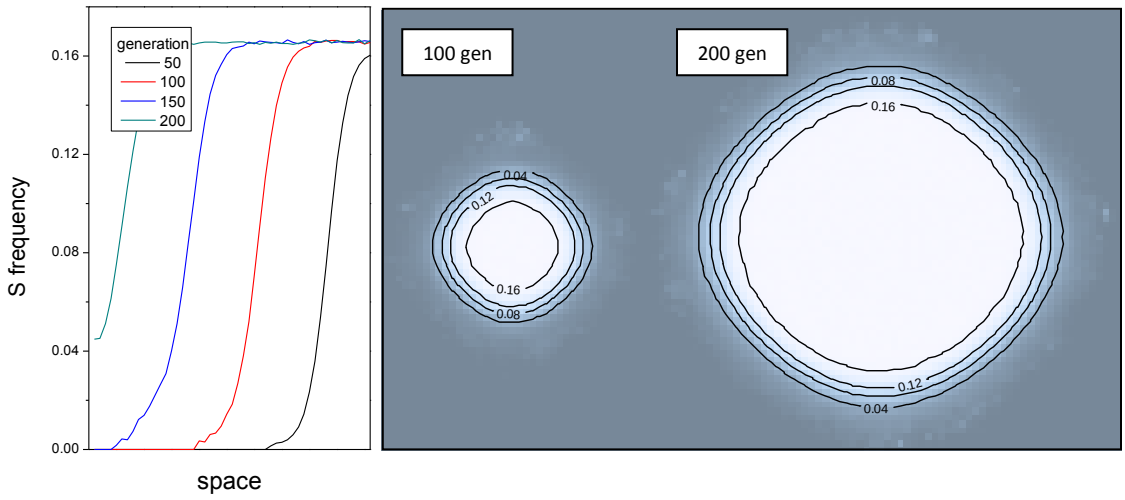

Figure 8 – Rate of advance of the S allele (average of 1000 runs). a) Four time slices from one-dimensional diffusions using the parameters of simulation 2 (cf. Table 4); b) Two time slices from two-dimensional diffusions using the parameters of simulation 11 (c.f. Table 4). In each case a single S mutation was seeded in the central population. The numbers in concentric rings refer to S allele frequencies. Gen - generations.

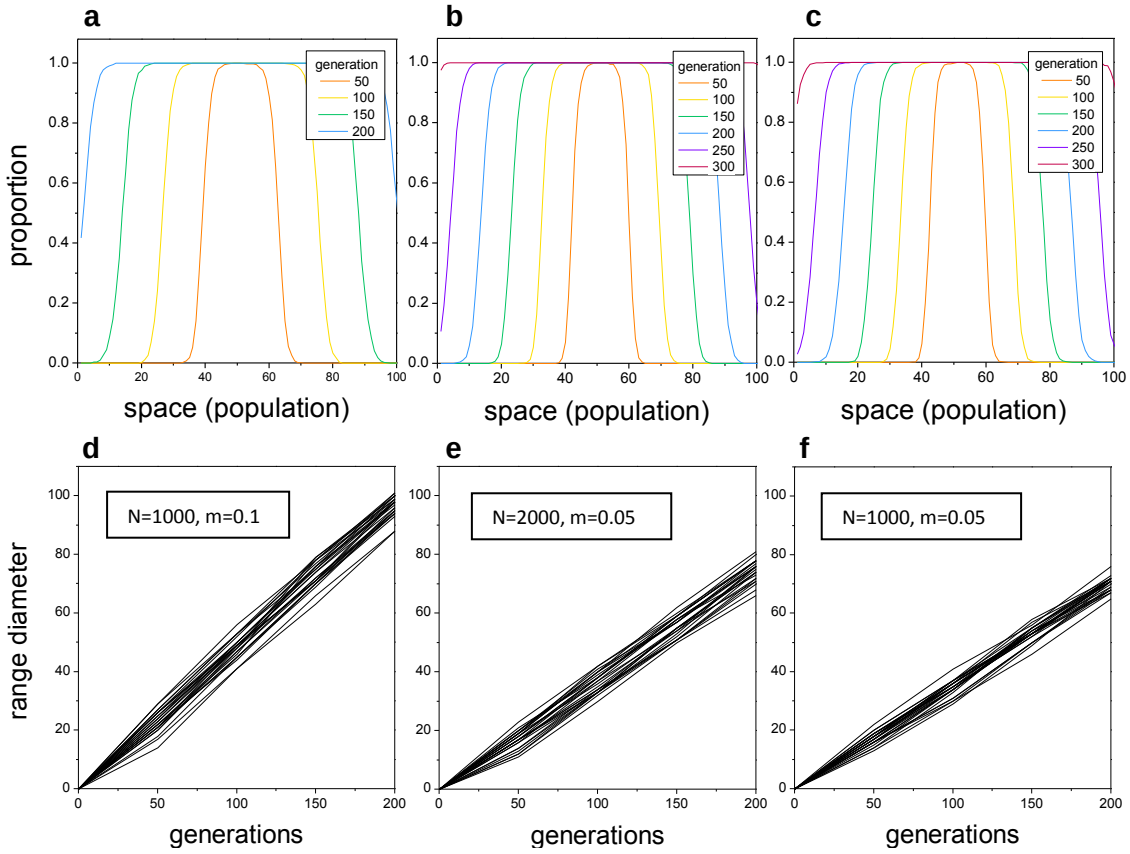

Figure 9 - Spatial reach and wave speed in one-dimensional simulations. (a-c) Proportion of simulation runs in which the S variant reached a given population at different time slices, using parameter combinations shown in panels d-f. (d-f) Relationship between the range diameter (maximum distance, in demes, reached by the variant) and the number of generations in 25 individual runs, using different parameter combinations. In all simulations a single S mutation was seeded in the central population.

of the average wave speed of 1000 runs between different parameter combinations are depicted in Figure 11. The wave speed increases with migration (Figure 12a) but is less sensitive to variations in population size (Figure 12b). In fact, whereas genetic drift associated with low population size seems to retard the spread of the mutant, the effect of increasing population size on the speed is characterized by diminishing returns (Figure 12b). These observations are in agreement with the theoretical expectations of the continuous isolation-by-distance model, which predict that, for sufficiently large populations, the wave speed depends exclusively on the migration rate and the selection coefficient, according to the relationship (Fisher, 1937; Ralph & Coop, 2010):

$$v = \sqrt{m 2^{-(d-1)}\, 2s} \quad (1)$$

where v is the speed (in demes/generation), m is the migration rate, d is the number of dimensions and s is the selection coefficient (for details see appendix 2). In the discrete

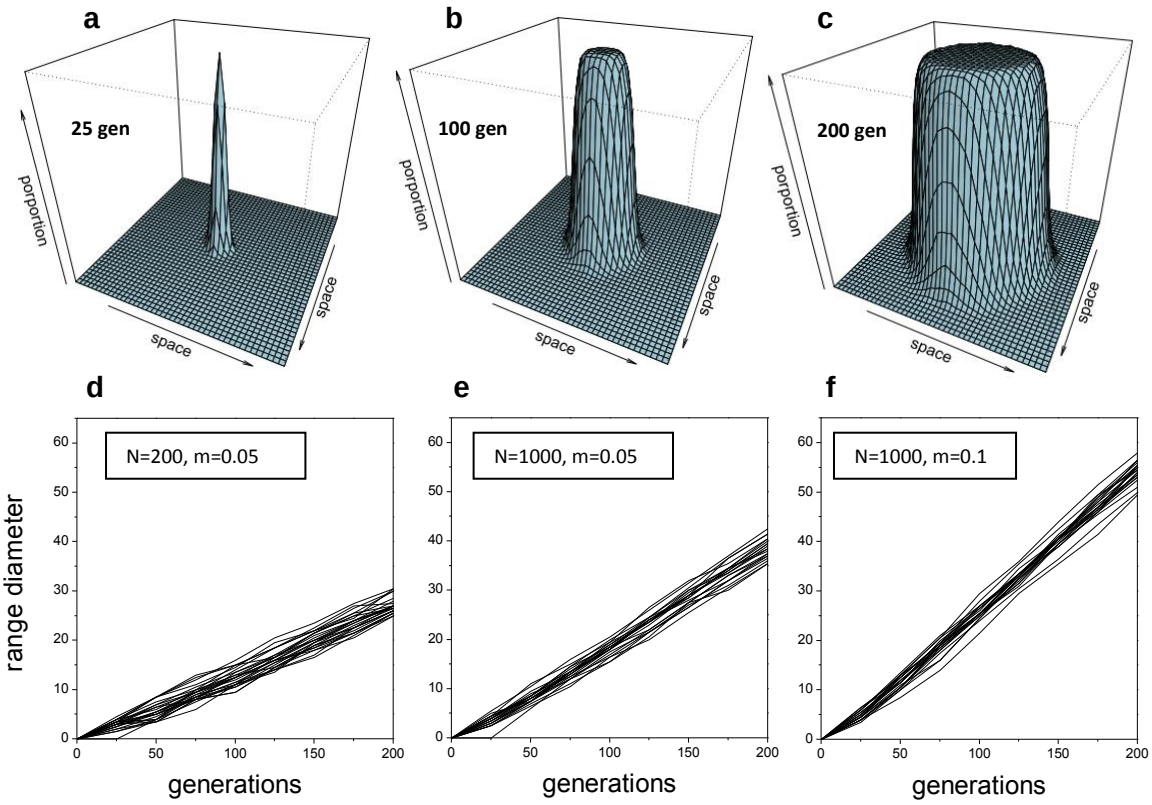

Figure 10- Spatial reach and wave speed in two-dimensional simulations. (a-c) Proportion of runs in which the S variant reached a given population at different time slices, with *N* = 200 and *m* = 0.05. (d-f) Relationship between the range diameter (maximum distance, in demes, reached by the variant) and the number of generations in 25 individual runs for different parameter combinations. In all simulations a single mutant allele was seeded in the central population.

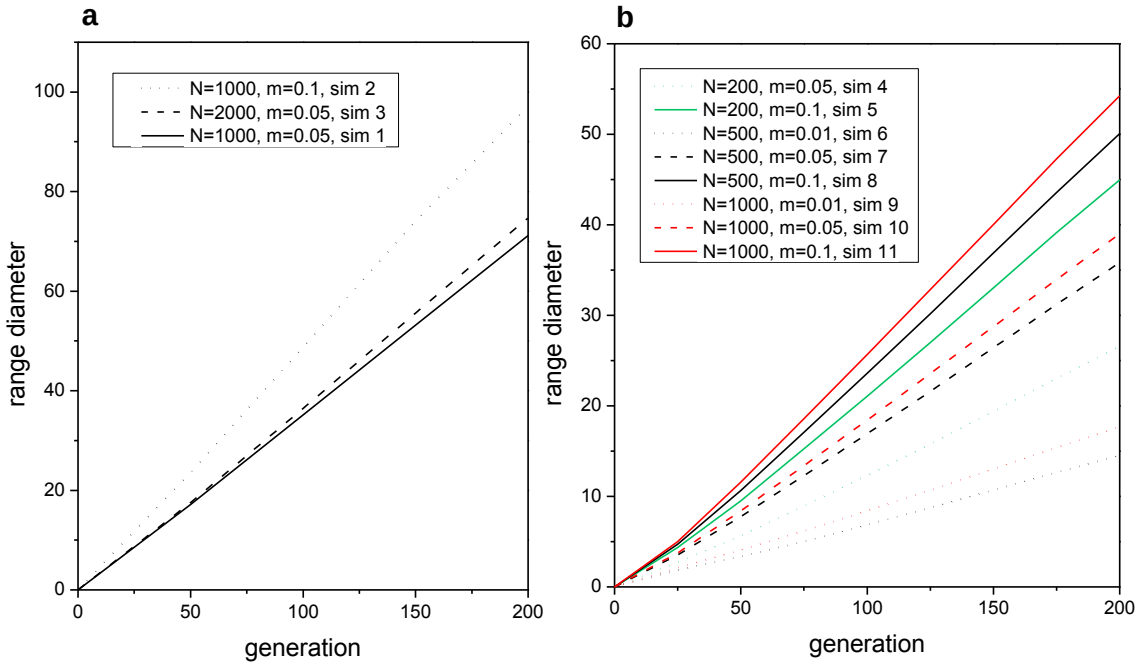

Figure 11 - Relationships between range diameter (in demes reached by the mutation) and number of generations. The lines were obtained by averaging 1000 independent simulation for each of different parameter combinations. a) One-dimensional simulations (simulations 1 – 3, cf. table 4); b) two-dimensional simulations (simulations 4 – 11, cf. table 4).

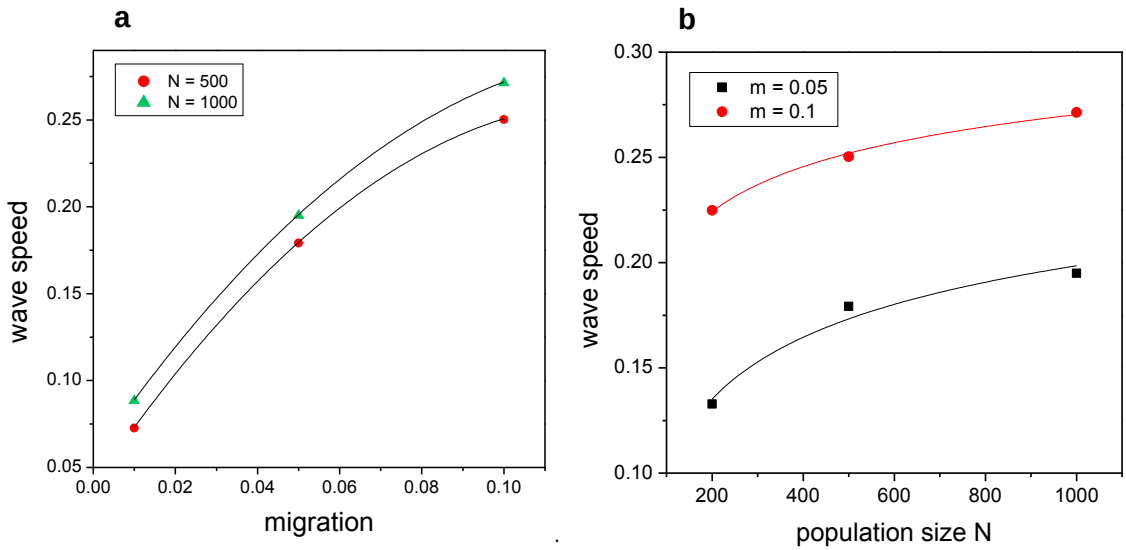

Figure 12 – a) Influence of migration (m) on the wave speed (in demes/generation) for two-dimensional simulations (simulations 6 – 11, cf. table 4). b) Influence of population size (N) on the wave speed (in demes/generation) for two-dimensional simulations (simulations 4, 5, 7, 8, 10 and 11, cf. table 4).

stepping stone model, the relationship also predicts, for the same parameter values, a wave of advance faster in one-dimensional simulations than in two-dimensional ones.

This can be observed by comparing Figures 11a and 11b. Despite the general trends being in agreement with the theoretical model, the computed speed values were one to two times faster than predicted, similarly to previous results by Ralph & Coop (2010). As discussed by the authors, the discrepancy may be explained by the discrete nature of generations and demes in the stepping stone model, which contrasts with the use of continuous time and space of the Fisher-KPP equation, from which equation 1 ultimately derives.

3.1.2 Haplotype diversity in time and space

3.1.2.1 Intrapopulation diversity

As discussed above, S variants arriving in each population by migration (or mutation in the case of the original population) are expected to increase in frequency due to selection. During this process, S-bearing haplotypes will also become increasingly diverse because of the cumulative effects of microsatellite mutation, recombination and gene conversion. However, recurrent sampling associated with migration counteracts the accumulation of haplotype diversity during the spread of the variant, generating consecutive bottlenecks causing substantial reduction of genetic diversity at the wave front. The genetic composition of these edge populations may be regarded as a subset

of the diversity generated by mutation, recombination and gene conversion that is found in populations that are closer to the diffusion center.

We have assessed the space-time evolution of S-linked haplotype diversity using a number of different summary statistics to assess consistency between different aspects of the data. Figure 13 shows how the mean number of different S-linked haplotypes within each population (n_h) increases with time but decreases in space towards the edge of the S variant distribution. As expected, n_h increases linearly with the population size, *N,* (Figure 14a), since genetic drift (inversely related with *N*) tends to eliminate the diversity accumulated in S-bearing chromosomes. Migration also increases n_h because new haplotype combinations arising in a specific population will be exchanged among populations (Figure 14b).

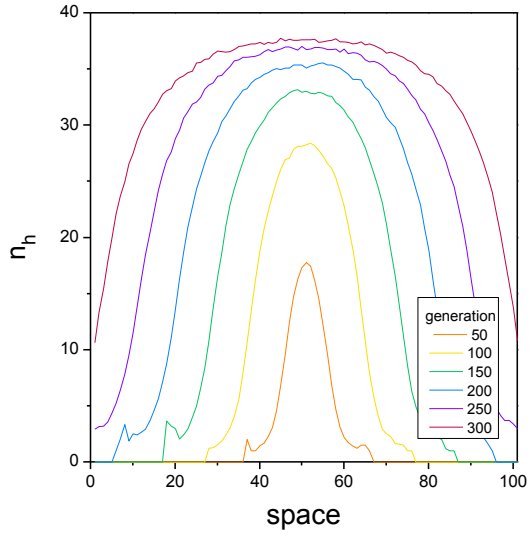

Figure 13 – Spatial-temporal variation in the number of different S haplotypes (n_h). The lines are time slices of simulation 1 (cf. table 4), and were obtained by averaging 1000 individual runs. A single mutant allele was seeded in the central population.

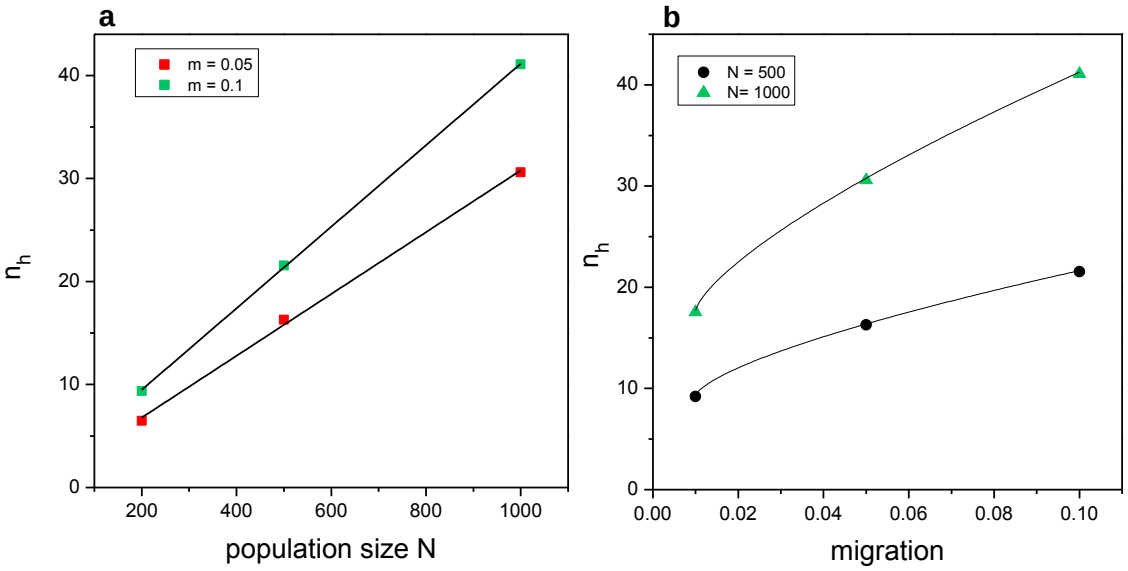

Figure 14 – a) Influence of population size in the number of different S haplotypes (n_h) in two-dimensional simulations (simulations 4, 5, 7, 8, 10 and 11, cf. table 4). b) Influence of migration in the number of different S haplotypes (n_h) in two-dimensional simulations (simulations 6 – 11, cf. table 4). The number of different S haplotypes was computed by averaging the top 1% n_h values for the average of 1000 runs, after 175 generations.

The frequency of the most common haplotype within S-bearing haplotypes, *P* (here also called predominant haplotype), which is inversely related to haplotype variation, can also be used to summarize the levels of S haplotype variation within each population. As shown in Figure 15, at any given generation, *P* tends to be highest at the edge of the variant distribution (Figure 15a) and this tendency is counteracted by the number of migrating gametes (*Nm*) that are exchanged between populations, which reflect the joint effects of *N* and *m* in increasing S-linked haplotype variation (Figure 15b).

Besides analyzing the frequency of the most common haplotype we have also evaluated the behavior of the original haplotype where the S mutation initially arose (Figure 16). Characteristically, the average frequency of the original haplotype (P_o) remained approximately constant through space, except for the most peripheral demes, where the number of simulations that were reached by the S mutation and could be used to calculate the average is very low (Figure 16a). This behavior is predictable since, theoretically, the expectation of P_o depends only on the age of the S mutation (Stephen et al., 1998). However, due to genetic drift, P_o values are expected to vary substantially across the different replications of a given simulation, which are equivalent to different realizations of the evolutionary process. This property is illustrated in Figure 16b, which shows that variance of P_o across replications sharply increases at the spatial edges of the S variant distribution where, due to serial bottlenecks, the effects of genetic drift are more intense. Another implication of the wave-like spread of the S mutation is that the proportion of replications where the original haplotype remains the most frequent haplotype (i.e. P_o=*P*) tends to be lowest at the edges of the distribution where genetic drift tends to increase the frequency of newly generated haplotypes (Figure 16c).

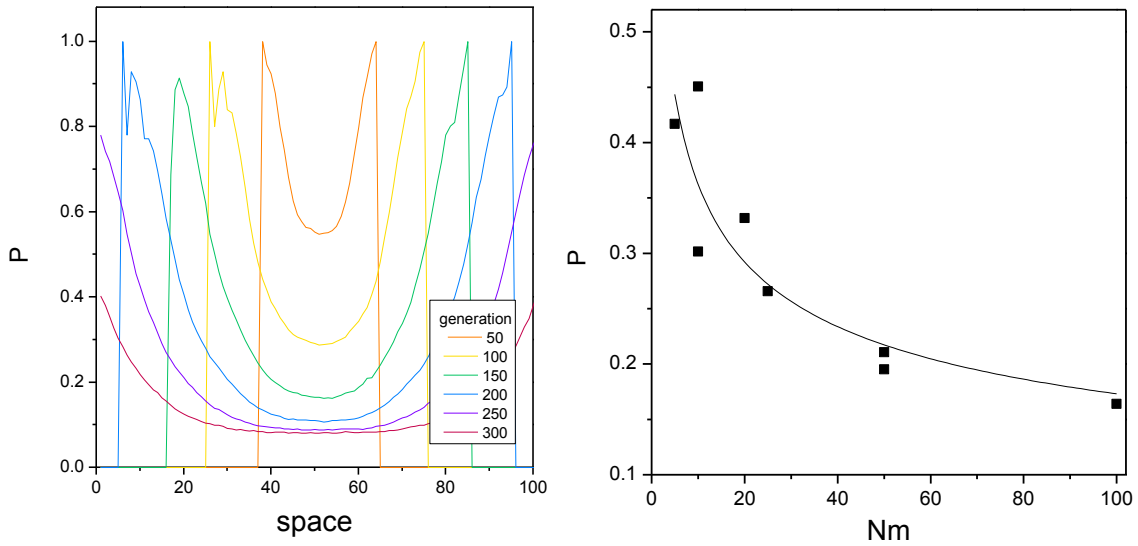

Figure 15 – a) Spatial-temporal distribution of the average frequency of the predominant haplotype (*P*) in simulation 3 (cf. table 4). b) Influence of the parameter combination *Nm* in the frequency (*P*) of the predominant haplotype for two-dimension simulations (simulation 4 – 11, cf. table 4). The predominant haplotype (*P*) was computed by averaging the lowest 1% *P* values for the average of 1000 runs, after 175 generations.

The behavior of n_h, *P and* P_o (Figures 6, 8 and 9) is consistent in showing a progressive loss of haplotype diversity towards the wave front. This observation is in pace with Livingstone's hypothesis that the areas where different homogeneous *HBB*S* haplotypes predominate should not be regarded as centers of origin, but as edges where the variant has arrived only recently (Livingstone, 1989). However, descriptive statistics like n_h, *P and* P_o do not provide information on the levels of molecular divergence across different haplotypes, which is a critical aspect of the observed S-linked haplotype distribution that must be addressed by simulations.

To overcome this caveat, we defined a more sophisticated summary statistic, which takes into account the molecular similarity of haplotypes by quantifying the average relative length of haplotype tracts encompassing the HBB locus that are shared between two *HBB*S*-bearing chromosomes randomly sampled from a given population (*Hii*) (Figure 17). Moreover, in contrast to n_h, *P and* P_o, this statistic may additionally be used to measure levels of interpopulation divergence in haplotype composition (*Hij*), when *HBB*S*-bearing chromosomes are sampled from different populations.

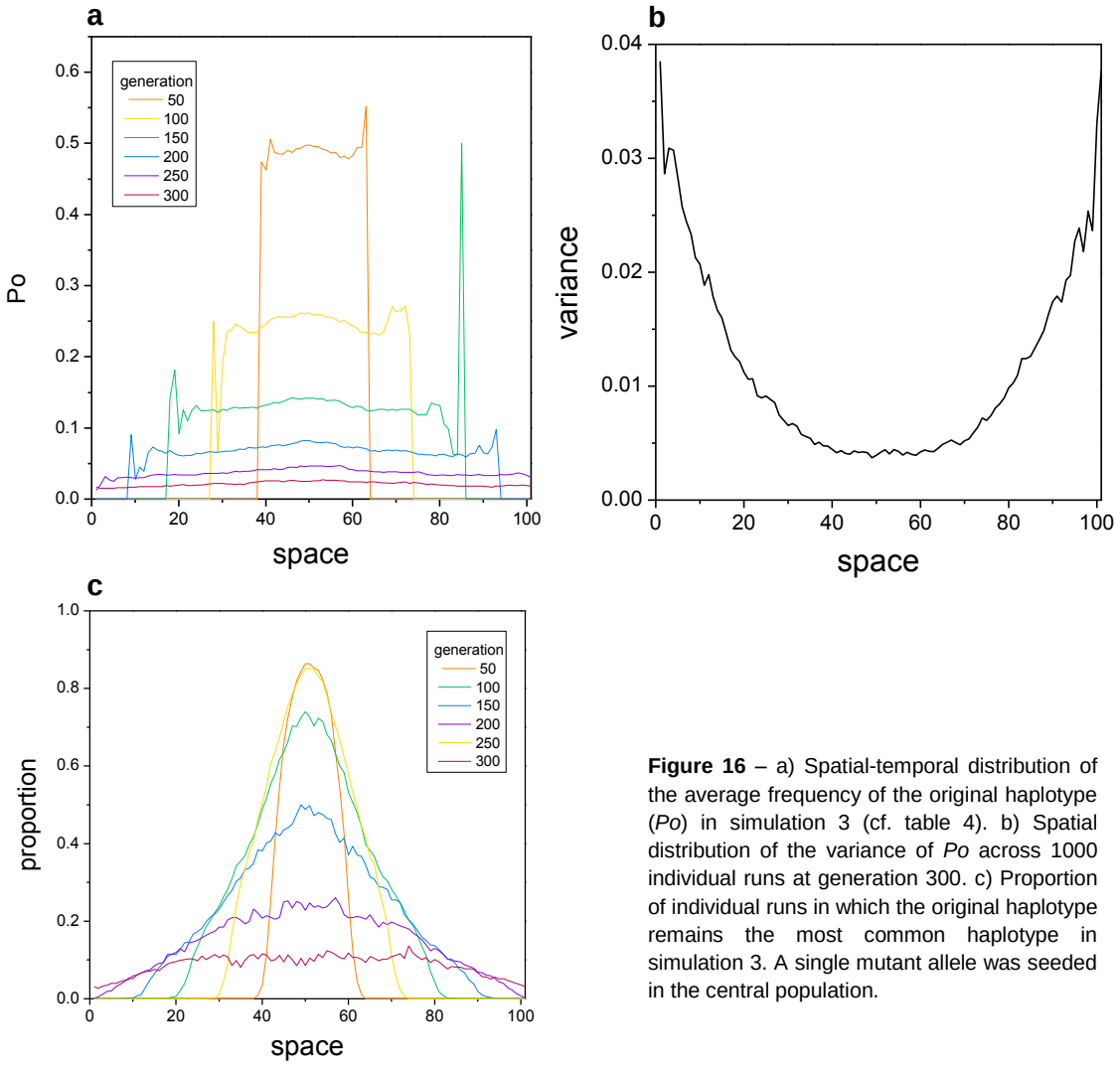

Figure 16 – a) Spatial-temporal distribution of the average frequency of the original haplotype (*Po*) in simulation 3 (cf. table 4). b) Spatial distribution of the variance of *Po* across 1000 individual runs at generation 300. c) Proportion of individual runs in which the original haplotype remains the most common haplotype in simulation 3. A single mutant allele was seeded in the central population.

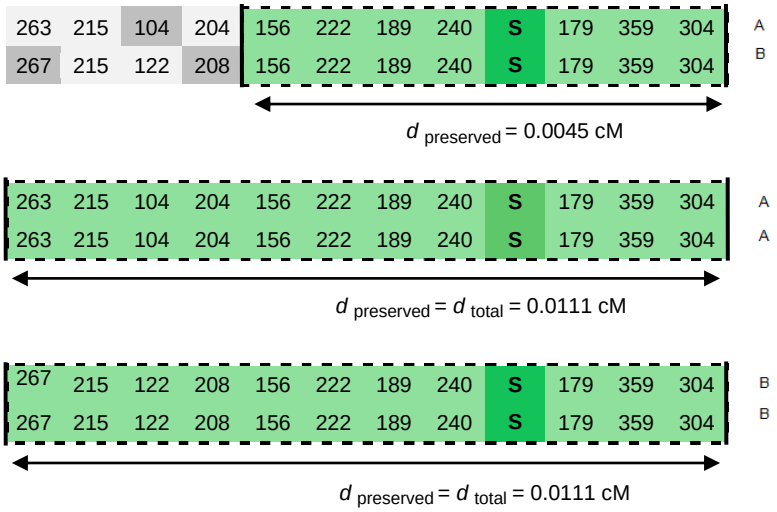

Figure 17 – Schematic representation of *Hii* calculation for a population with 2 different haplotypes (A and B). The *HBB*S*-linked tract shared by each pair of haplotypes is represented in green. The homogeneity between two haplotypes is computed by multiplying the relative length of the tracts by the probability of sampling a given pair of haplotypes. The *Hii* of a population is the sum of the homogeneities computed for each pair.

$$H_{AB} = \frac{d_{preserved}\,(cM)}{d_{total}\,(cM)} \times 2 \cdot frequency_A \cdot frequency_B$$

$$H_{AA} = \frac{d_{preserved}\,(cM)}{d_{total}\,(cM)} \times frequency_A^{\,2}$$

$$H_{BB} = \frac{d_{preserved}\,(cM)}{d_{total}\,(cM)} \times frequency_B^{\,2}$$

Figure 18 illustrates the spatial-temporal patterns of *Hii* in a one-dimensional scenario. Consistent with the previous results, haplotype homogeneity exhibits a sharp increase in populations located at the edges of the distribution. For example, the *Hii* values at the edges of the S variant distribution at generation 200 are higher than those observed in the population of origin at generation 50 (Figure 18). Two-dimensional simulations generated similar patterns (not shown).

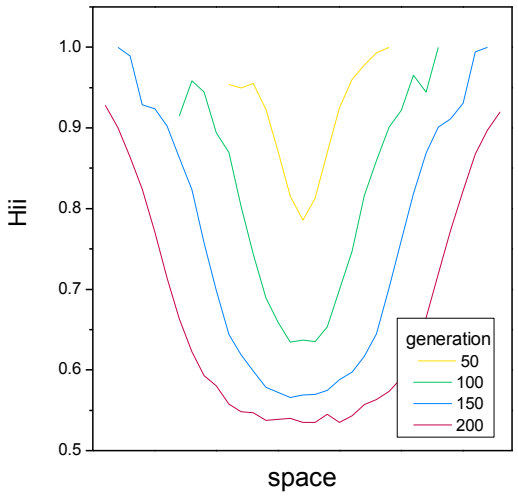

Figure 18 – Spatial-temporal distribution of the intrapopulation haplotype homogeneity (*Hii*) obtained by averaging 1000 runs of simulation 1 (cf. table 4). A single mutant allele was seeded in the central population.

3.1.2.2 Interpopulation diversity

To assess the levels of interpopulation differentiation in the haplotype composition of S-bearing chromosomes, we analyzed how *Hij* values varied with distance of separation between pairs of peripheral populations located in opposite directions from the central population, as well as between pairs of central-peripheral populations. Figure 19a exemplifies the patterns of *Hij* variation as functions of the distance of separation between populations in a one-dimensional system. The pairwise S-linked haplotype similarities decreases as the distance separating two populations increases, and this decrease is more marked when the populations are both located at the periphery of the distribution than when central and peripheral populations are compared (Figure 19a).

Intuitively, this can be explained by the directional flow of haplotypes to each side of the wave of advance. While selection increases the number of S haplotypes in central populations, the low number of haplotypes that reach the wave front will more easily spread in the forward direction than backwards, since the frequency of the S mutation is higher in the center of diffusion, where these haplotypes are easily diluted. This behavior is again consistent with Livingstone's suggestion that the differences between *HBB*S* haplotypes predominating in different regions may result from these regions being located at the periphery of the S-value distribution (Livingstone, 1989).

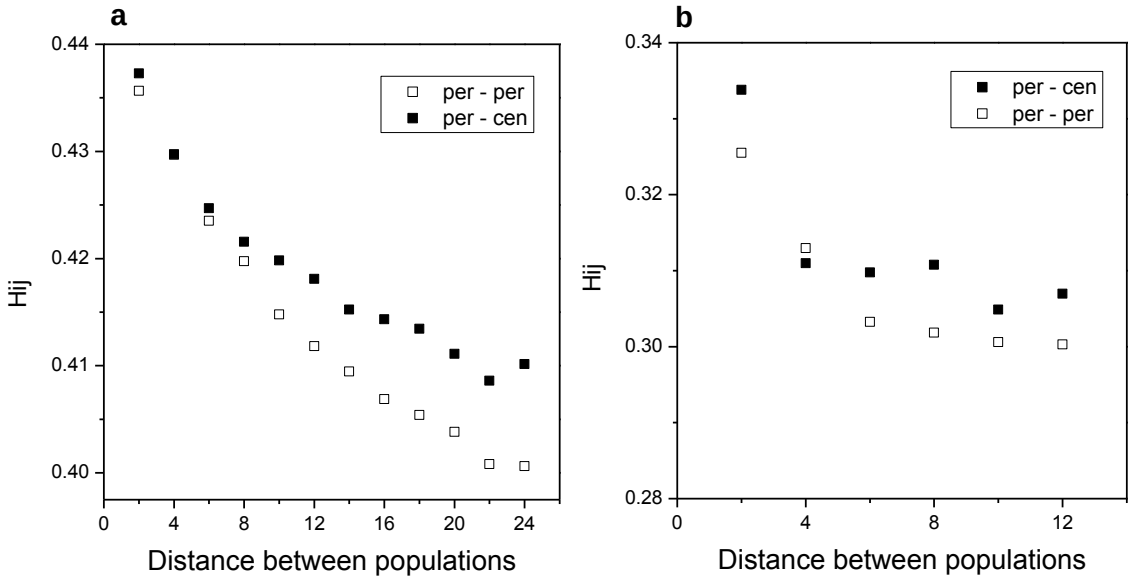

Figure 19 - Variation of the interpopulation haplotype homogeneity (*Hii*) as function of the distance of separation between pairs of populations (in number of populations) . a) One dimensional simulation (simulation 1, cf. table4). b) Two-dimensional simulation (simulation 4, cf. table 4). Per – peripheral population, Cen – central population. The points are averages of 1000 independent runs.

Figure 19b shows the relationships between *Hij* and distance of separation in two spatial dimensions, which are reminiscent of Moran's *I* spatial correlograms obtained with the isolation-by-distance framework (Moran, 1950; Epperson, 2003). Interestingly, the form of the decrease of *Hij* with distance in two-dimensional models is quite different from that observed in Figure 19a for one-dimension, highlighting the influence of the number of spatial dimensions in the patterns of haplotype variation.

Figure 20 shows that both *Hii* and *Hij* are largest when the number of migrating gametes is low (*Nm*), indicating that migrant exchange across populations increases the within population diversity and blurs the inter-population differentiation of S-linked haplotypes.

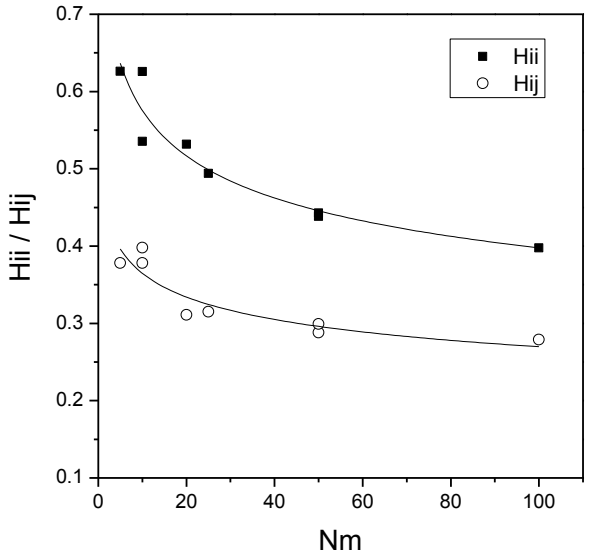

Figure 20 – Influence of the parameter combination *Nm* in the intra- (*Hii*) and interpopulation (*Hij*) homogeneity for two-dimensional systems, using simulations 4-11 (cf. table 4). Displayed values are averages of 1000 individual runs after 200 generations

3.2 Identifying sectors

The analyses performed in the previous sections using different summaries of the S-linked haplotype diversity were all consistent in showing that consecutive sampling of S-bearing chromosomes results in a noticeable increase of haplotype homogeneity at the edges of the variant distribution. This homogeneity is caused by the dramatic reduction of the number of S variants reaching populations that are far from the location where the S mutation originated. Eventually, the rare "edge haplotypes" will increase in frequency due to natural selection and will give rise to S-linked haplotype profiles that are different from those observed at the center of the spatial distribution.

The observed increase in the levels of S-haplotype homogeneity as one moves further away from the origin is strikingly analogous to the increase in frequency that can be experienced by mutations arising in the wave front of populations expanding into new territories (Edmonds, 2004; Hallatschek and Nelson, 2008). This occurrence, termed "allele surfing" (Klopfstein; 2006), has been extensively studied by means of simulation and analytical methods, in the context of range expansions, in which new mutants are neutral or even deleterious (Edmonds et al., 2004; Klopfstein et al., 2006; Travis et al., 2007; Hallatschek and Nelson, 2008; Excoffier and Ray, 2008) Moreover, Hallatschek et al. (2007), using microbial populations, have experimentally shown that allele surfing during range expansions is expected to create temporarily stable sectors where alternative alleles become completely fixed in different areas of the geographic range. This outcome has been subsequently replicated by simulation (Excoffier & Ray, 2008; François et al., 2010).

In contrast with these studies, individuals in our simulation framework are not expanding towards uninhabited territory, and the S variant, although being deleterious in homozygotes, behaves like a favored allele. However, the wave of advance of the S mutation can still be considered analogous to range expansions, if we focus on the dynamics of the populations of S-chromosomes. In this setting, spreading S-chromosomes may be viewed as haploid individuals entering newly colonized territory; overdominant selection is equivalent to logistic population growth, and mutant alleles correspond to new S haplotypes resulting from recombination or mutation at linked microsatellites. This analogy, which is supported by our observations on the loss of S-linked haplotype variation at the wave front, prompted us to assess if large patches of the S-spreading area could become occupied by different predominant S-linked haplotypes.

To this end, we looked for groups of populations sharing the same predominant haplotype in each replication of the simulations. Figure 21 shows the end results of individual realizations of simulation 8, 10 and 11 (cf. table 4) clearly demonstrating that the wave of advance of the S allele can create several patches (or sectors) formed by contiguous populations sharing S-linked modal haplotypes that are different from those observed elsewhere. These sectors are evocative of the currently observed geographic segregation of the *HBB*S* haplotypes in Africa.

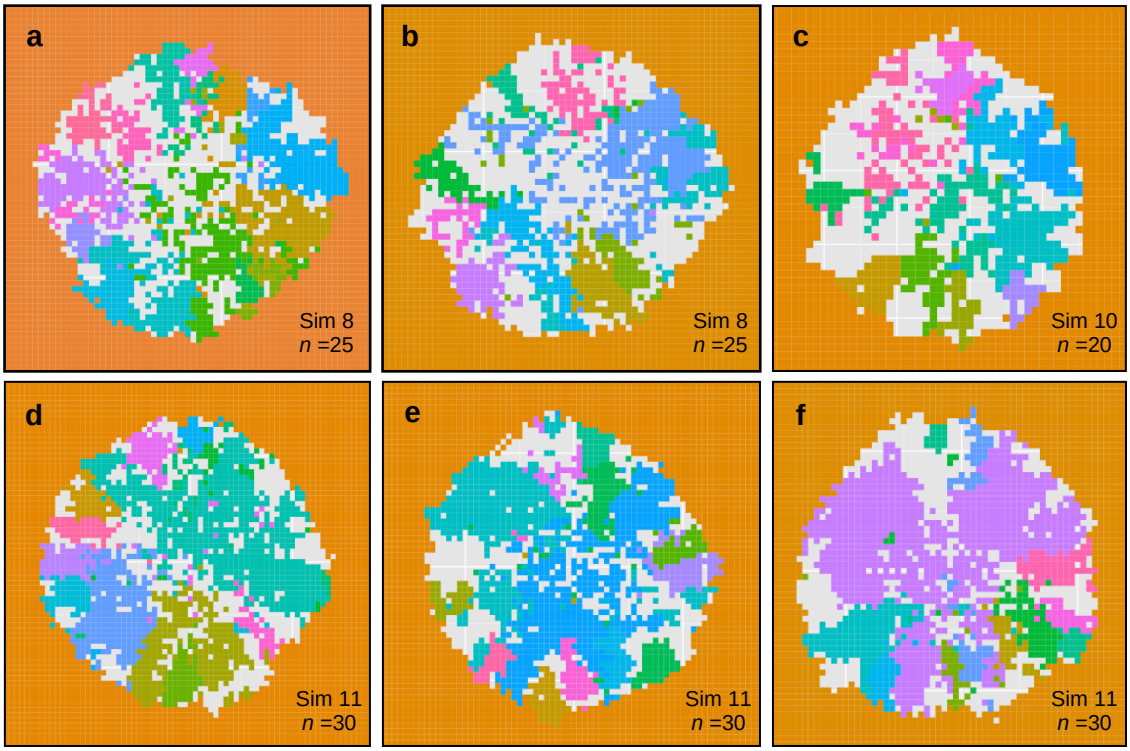

Figure 21 – Outputs of six independent runs of 200 generations showing the spatial distribution of sectors encompassing populations with the same predominant S-linked haplotype. Each color represents a different haplotype. Orange stands for areas not reached by the S mutation. Haplotypes that are not shared by a minimum of *n* populations are not displayed (grey areas). Running conditions were according to simulations 8, 10 and 11 (cf. table 4).

Importantly, the area close to the diffusion center is devoid of sectors in the majority of simulations (eg. Figure 21b). When sectors are formed in this area, they usually involve a small number of populations or are characterized by low frequencies of predominant haplotypes. This is not unexpected, since haplotype diversity is highest and S-linked haplotype frequencies are flatter at the place of origin.

Figure 22 illustrates how the frequency of the predominant haplotype in each sector varies across the whole geographic range of the S mutation, using one realization of simulation 11 (cf. table 4). The surfing phenomenon is clearly shown by the gradual increase of haplotype frequencies towards the edge of haplotype distributions, even within the limits of each sector (eg. Figures 22a6 and 22a10). Interestingly, sectors seem to be quite sharp, as S-haplotypes predominating in a sector seldom reach populations that are far beyond the limits of that sector (eg. Figure 22b). Finally, we observed that the haplotype with the widest distribution (corresponding to the largest sector) is frequently the original haplotype to which the S mutation became associated (eg. Figure 22a6). However, this original haplotype may not be the most common one, and may even be extinct in the central population after 200 generations (cf. Figure 16c).

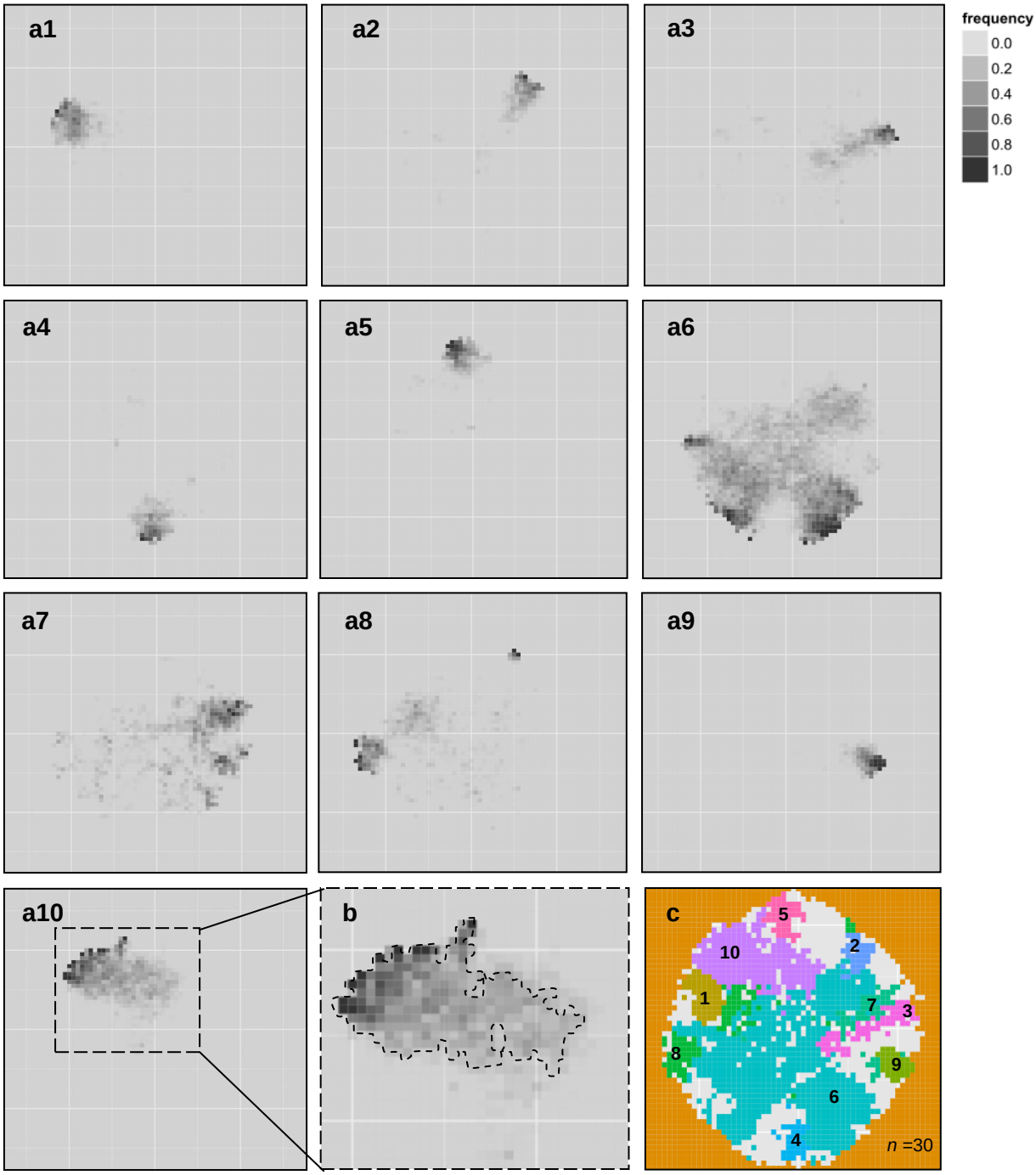

Figure 22 – Spatial distribution of S-haplotype frequencies in one replicate of simulation 11 (cf. table 4) after 200 generations. a1-a10) Spatial variation of the frequencies of S-haplotypes predominating in each of the sectors identified in panel c. b) Frequency distribution of the haplotype predominating in a10 c) Spatial variation of the frequency of the S-haplotype predominating in sector 10 (also shown in panel a10) with delimitation of the sector's limits (dashed contour). *n* – minimum number of populations defining a sector.

3.3 Evaluating fitting

In the previous section we have shown that geographic patches dominated by unique S-linked haplotypes may arise during the dispersal of a single advantageous *HBB*S* mutation, without being necessary to invoke recurrent mutation. However, a patchy distribution of haplotypes is not the only property of the observed *HBB*S* haplotype distribution. In fact, besides displaying geographic segregation, the six different haplotype classes discriminated in the observed dataset (cf. section 2.1) are also characterized by high levels of haplotype homogeneity (high *P* and *Hii*), while modal haplotypes from different classes are highly divergent (low *Hij*). Tables 6, 7 and 8 quantify the most important characteristics of the haplotype groups found in the empirical dataset using the *P*, *Hii* and *Hij* summary statistics (cf. section 3.1), assuming that each of haplotype group corresponds to a geographic area (or sector) encompassing different populations (cf. section 3.2).

3.3.1 Pairwise population comparisons in one-dimensional simulations

To further evaluate the extent to which single-mutation simulations could approximate the empirical data, we started by qualitatively comparing the observed *Hii* and *Hij* statistics (Table 6) with the values obtained within and between pairs of populations located at different positions of simulated one-dimensional waves of advance (Figure 23).

Figure 24 illustrates the influence of the number of simulated generations and the relative position of compared populations in the levels of resemblance between simulated and observed data. After 50 generations (Figure 24a), when the comparisons involve peripheric and central populations, or peripheric populations that

Table 6 – Frequency of the intra (*H*ii) and interpopulation (*H*ij) homogeneity observed in African populations. *H*ii values are represented in grey.

	Ba-W	Ba-E	Ca	Be	AA	Se
Ba-W	0.681					
Ba-E	0.599	0.893				
Ca	0.103	0.102	0.513			
Be	0.113	0.111	0.113	0.708		
AA	0.083	0.080	0.084	0.269	0.511	
Se	0.022	0.014	0	0.185	0.144	0.867

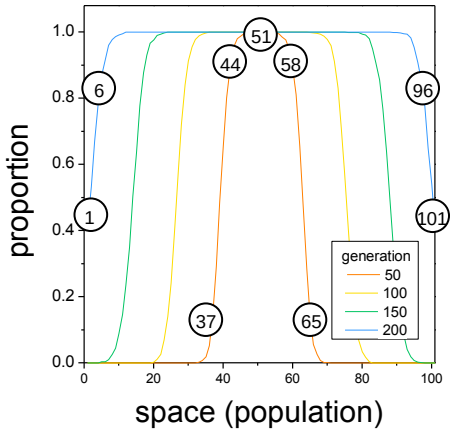

Figure 23 – Schematic representation of the geographic location of populations compared in figure 24 (simulation 2, cf. table 4).

are close to the center, the simulation data displays a strong correlation between the levels of haplotype diversity calculated within (*Hii*) and between (*Hij*) the populations involved in each pairwise comparison, (Figure 24a1-a3 and Figure 23); i.e., the decrease of *Hii* is usually followed by a decrease in *Hij*. In these cases, there is a noticeable separation between the simulated and observed data, as the latter are characterized by combinations of high intrapopulation homogeneity (high *Hii*) and high interpopulation divergence (low *Hij*) values. In contrast, when comparisons involve populations situated at opposite edges of the wave of advance (Figure 24a4), the correlation between *Hii* and *Hij* becomes less pronounced in simulated data and a better approximation with observed data is achieved. However, the number of simulated data points in these comparisons is low, reflecting the low number of simulations in which the *HBB*S* allele reaches the populations located at the edge of the spatial grid.

After 200 generations (Figure 24b) the simulated and observed data become much closer when comparisons involved peripheral populations (Figure 24b3 and 24b4), even if these populations are not strictly located at the edges of the grid (Figure 24b3).

These preliminary assessments show that 50 generations are probably not enough to create the level of divergence observed between different haplotypes and that populations with high levels of intra-population homogeneity (high *Hii*) can hardly represent a center of diffusion of mutation. For example, if we consider that the area where the Bantu haplotype predominates is located in the periphery of a wave of advance of a single mutation, it would be very unlikely that this mutation originated in any of the regions where other S-haplotype categories predominate. Instead, under the hypothesis of single mutation, the most likely interpretation of the observed data is that, as proposed by Livingstone (1989), current patches occupied by different haplotypes are located at the edges of a spreading area whose center is to be found elsewhere.

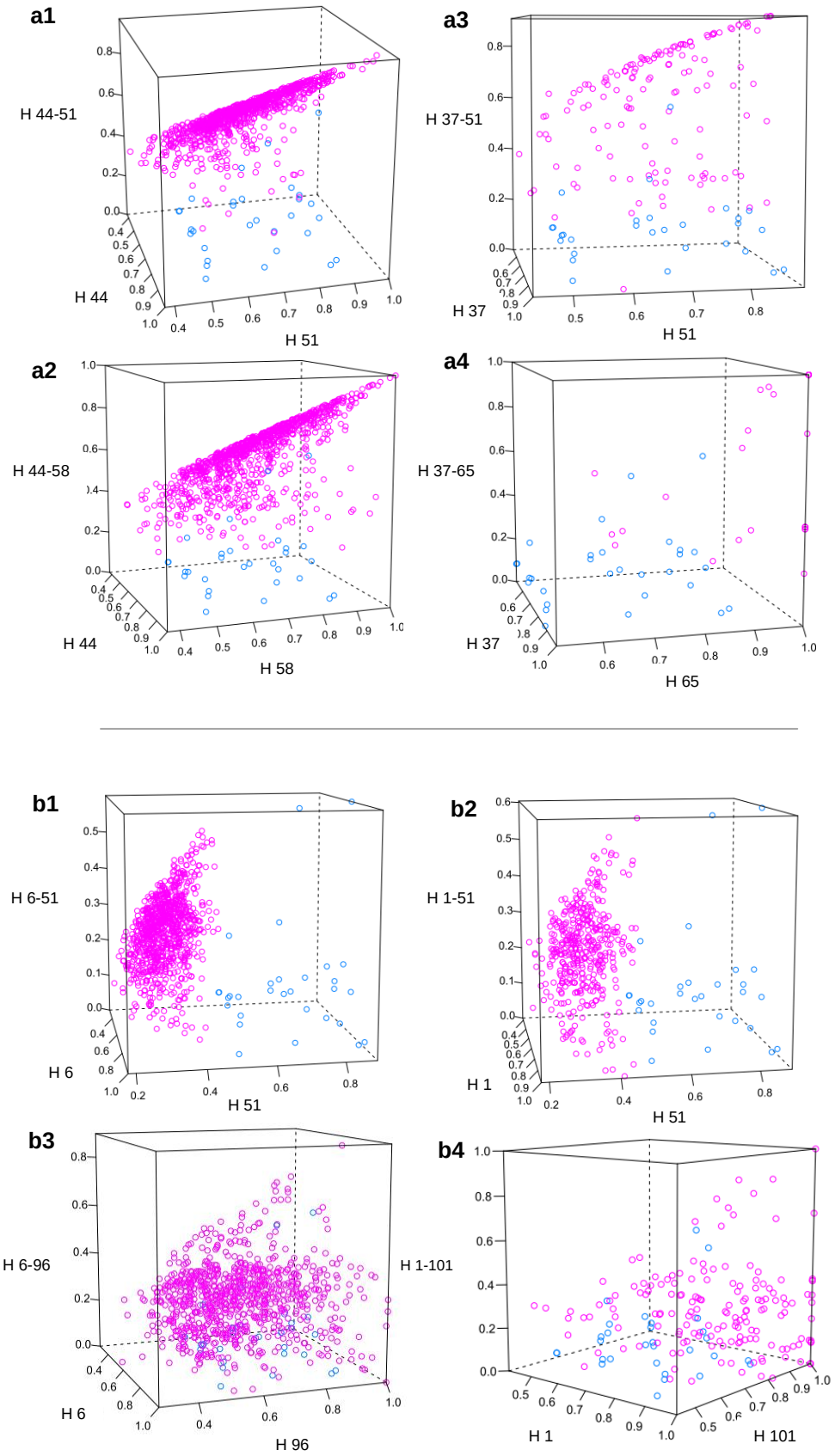

Figure 24 – Comparison between simulated and observed *Hii* and *Hij* values at 50 (a) and 200 (b) generations, using simulated populations whose location is shown in Figure 23. Simulated values are indicated in pink and observed values are indicated in blue.

Table 7 – Frequency of the predominant haplotype (*P*) observed in African populations.

	Ba-W	Ba-E	Ca	Be	AA	Se
P	0.58	0.903	0.222	0.525	0.571	0.867

Table 8 – Average of the frequency of the predominant haplotype (*P*), intra (*H*ii) and interpopulation (*H*ij) homogeneity.

$\overline{Hii}$	$\overline{Hij}$	$\overline{P}$
0.695	0.135	0.611

3.3.2 Pairwise comparisons of sectors in two-dimensional simulations

The approach developed in the previous section, although informative about some of the factors influencing the distribution of haplotype variation, only provides a poor analogy between simulated and real data, because the spread of *HBB*S* was likely to have been two-dimensional, and because well-defined simulated populations are not equivalent to large haplotype spatial patches encompassing multiple populations.

To overcome this limitation, we further attempted to evaluate the fit between simulated and empirical data by characterizing the levels of haplotype diversity within and across sectors that were defined as discussed in section 3.2. To this end, we first chose the minimum number of populations defining a sector, and kept it fixed in each analysis. Although this choice is arbitrary, the use of different minimum sizes for a sector has a small influence in the results (not shown). Then, for each of 1000 simulation runs, the frequency of the predominant haplotype and the intra-population homogeneity were computed by averaging *P* and *Hii* values (cf. section 3.1) across all sectors. The interpopulation homogeneity was measured by averaging pairwise *Hij* values computed between all possible pairs of sectors formed in each run. A set of observed *P*, *Hii* and *Hij* were also obtained by applying the same rationale to the six-haplotype groups from the empirical data, which are equivalent to six different geographic sectors (Table 8). Finally the three summary statistics values obtained for each run of each simulation were compared with the averaged observed values by means of Euclidean distances computed according to:

$$d = \sqrt{(P_{sim} - P_{obs})^2 + (Hii_{sim} - Hii_{obs})^2 + (Hij_{sim} - Hij_{obs})^2} \quad (2)$$

In all, a maximum total of 1000 distances between simulated runs and observed data were obtained for each simulation.

To assess the relative consistency between different simulation scenarios and the empirical data, we ordered the simulated data sets by increasing distance to the observed data and estimated the fraction of datasets obtained with each simulated scenario in the nδ smallest distances. Figure 25 displays the fractions of simulated data closest to the empirical data for simulation scenarios 4 to 11 (cf. Table 4) for different nδ values. Simulation 4 (*N*=200, *m*=0.05) is the one that better fits the observed data. Simulation 6 (*N*=500, *m*=0.01) has also a relatively good fit, especially for low nδ values. Differently, simulations 5, 8, 10 and 11, all with high *Nm* are much more discordant from the observed dataset and are not even represented in the range of nδ smallest distances. Taken together, the results indicate that conditions where genetic drift is high (low *N*) and migration is low (*m*), tend to favor spatial patterns with the closest resemblance to the observed data.

In Figure 26 we compare the fit to the observed data of simulations 12 and 14, which unlike simulations 4 to 11 also take gene conversion into account (cf. Tables 4 and Table 5). The importance of gene conversion has been stressed by several proponents of the single-origin hypothesis of *HBB*S*, since direct A->S gene conversion provides a simple mechanisms for the S allele to become associated with haplotypes that are very different from the chromosome to which the S mutation was originally associated (Flint, 1998; Livingstone, 1989). Consistent with Figure 25, simulation 12 with *Nm=10* shows a much better fit to the observed data than simulation 14 with *Nm*=100.

We have additionally compared simulation scenarios 4 and 12 (cf Tables 4 and 5) to assess the implications of including or not direct A->S gene conversion in the parameter set. As shown in Figure 27 there is no clear discrimination between the best gene conversion and non-gene conversion scenarios, suggesting that gene conversion does not improve the fitting of simulated to observed data.

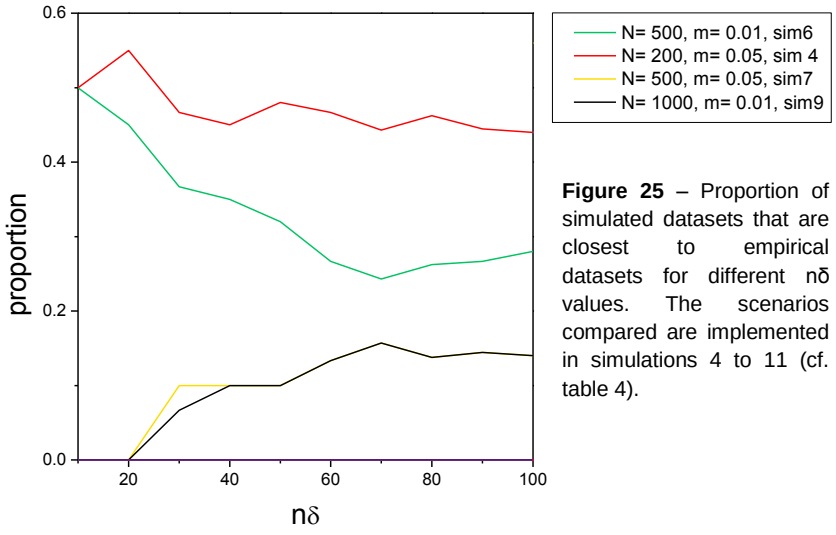

Figure 25 – Proportion of simulated datasets that are closest to empirical datasets for different nδ values. The scenarios compared are implemented in simulations 4 to 11 (cf. table 4).

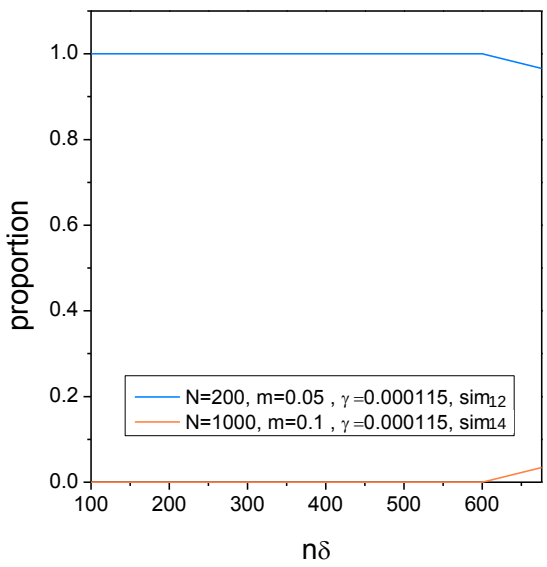

Figure 26 – Proportion of simulated datasets that are closest to empirical datasets for different nδ values. The two compared scenarios are implemented in simulations 12 and 14 (cf. table 5). γ - gene conversion rate.

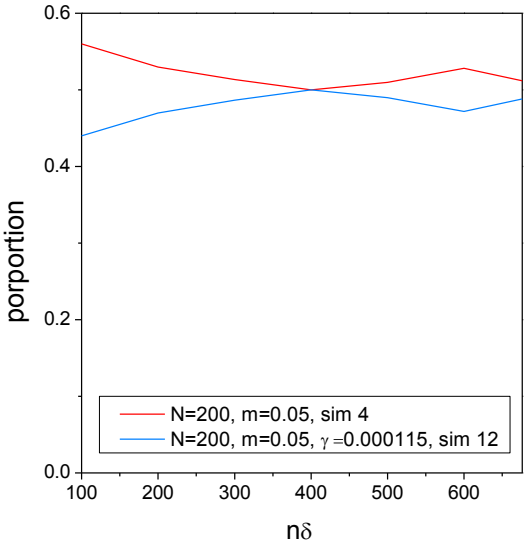

Figure 27 – Proportion of simulated datasets that are closest to empirical datasets for different nδ values. The two compared scenarios are implemented in simulations 4 and 12 (cf. table 4 and 5, respectively).

3.3.3 Preliminary assessment of multicentric origins

In the previous analyzes, we focused on evaluating the outcomes of different simulation conditions, assuming a single origin for *HBB*S*. In this setting, the spatial grid was viewed as a miniaturized (and simplified) model of the African continent, where *HBB*S* haplotype patches were considered to be equivalent to current areas of Africa where different *HBB*S* predominate. However, our simulated grids can be also used to assess, at least preliminarily, the implications of multicentric origins. In fact, if the *HBB*S* is considered to have arisen by multiple recurrent mutations, the simulated spatial grid can be viewed as the dispersal area of one of the several independent mutations. In this setting, to compare the fit between observed and simulated data, we merged all populations in the spatial grid into a single metapopulation and computed the Euclidean distance between the frequency of the predominant haplotype (*P*) in the meta-population and the observed dataset.

Figure 28a displays the relative fit to the data of simulation scenarios 4 to 11 (cf. table 4) using this new approach. In this case simulation 6 is the one that better fits the observations, but no marked differences were observed between most simulated scenarios. To assure comparability between the muticentric and the single-origin scenarios we reassessed the fit of simulations under the single origin scenarios, using *P* as the only summary statistic (but calculating it as the average of *P*s from different sectors and not from a single metapopulation). As shown in Figure 28b, the most suitable simulation was found again to be simulation 4 (cf Table 4 and Figure 25).

Finally, as show in Figure 28c, we compared the two best simulations of the single origin and multicentric frameworks and found that the single origin model clearly outperforms the muticentric model. The low performance of the multicentric setting is most probably due the fact that in each center of diffusion, at least in a wave of advance scenario, haplotypes linked to the spreading mutation will tend to become very diverse with time, and regions of high homogeneity, like those currently observed in Africa, will be difficult to find unless the mutation is very recent. Figure 29 illustrates this reasoning by showing the relationship between *P* at a metapopulation representing a center of diffusion, under the multicentric hypotheis and the number of simulated generations. Note that to fit some observed values of *P* which are on the order of 0.6 (cf. Table 7), each recurrent mutation had to appear at most about 60 generations ago,

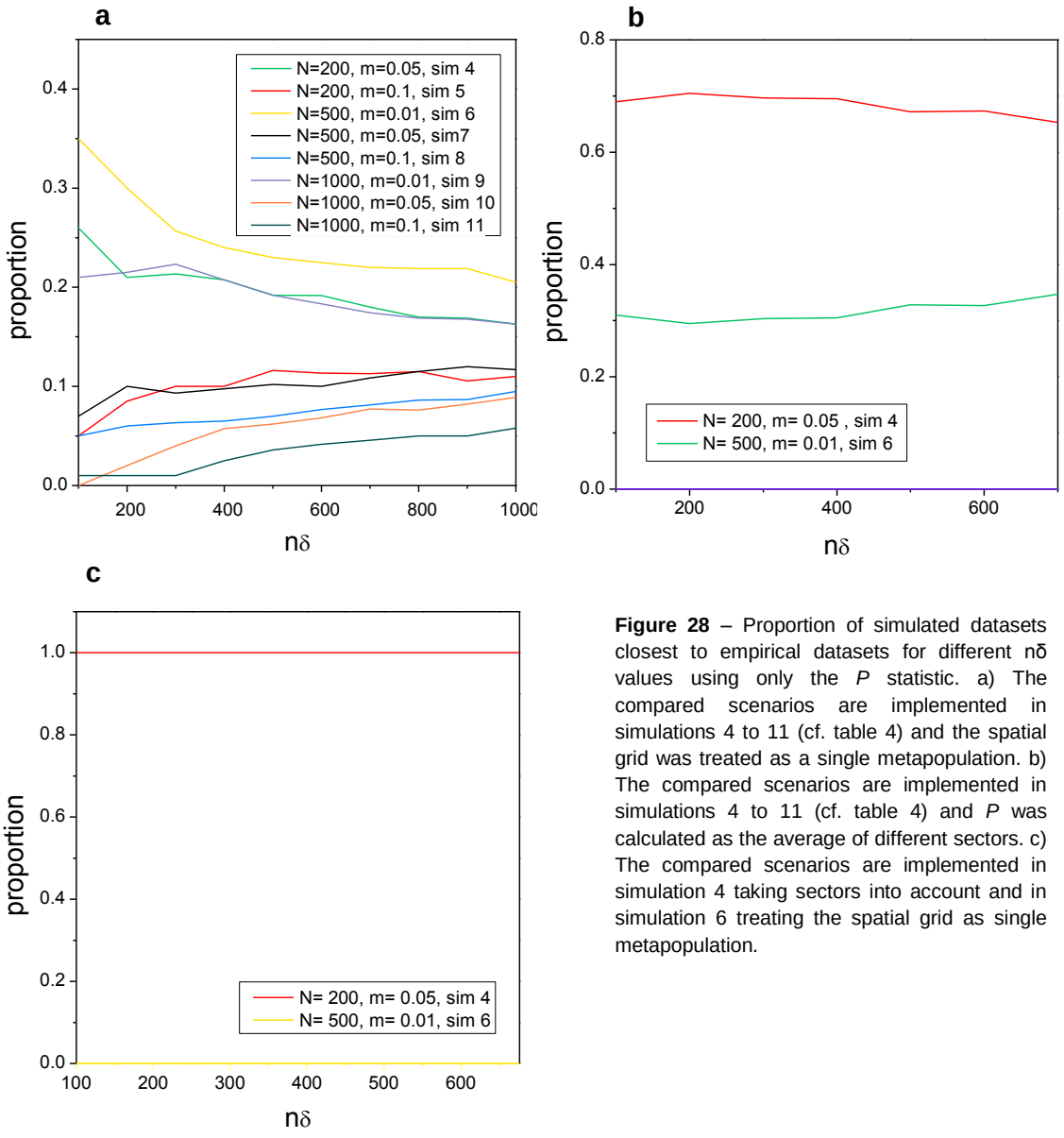

Figure 28 – Proportion of simulated datasets closest to empirical datasets for different nδ values using only the *P* statistic. a) The compared scenarios are implemented in simulations 4 to 11 (cf. table 4) and the spatial grid was treated as a single metapopulation. b) The compared scenarios are implemented in simulations 4 to 11 (cf. table 4) and *P* was calculated as the average of different sectors. c) The compared scenarios are implemented in simulation 4 taking sectors into account and in simulation 6 treating the spatial grid as single metapopulation.

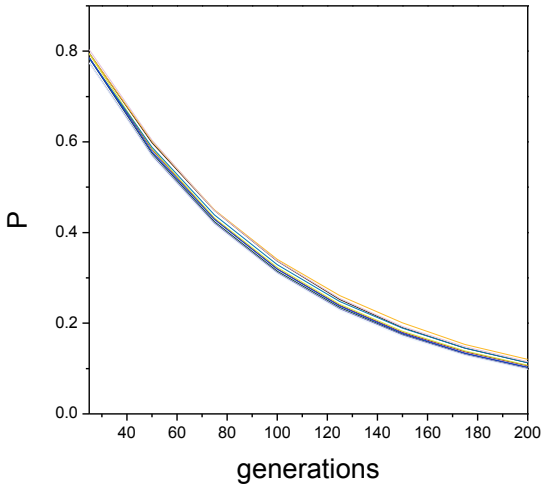

Figure 29 – Variation of the frequency of the predominant haplotype (*P*) with time (generations). *P* was obtained by averaging 1000 independent simulation runs of each simulation, treating the spatial grid as a single metapopulation. The lines correspond to simulation 4-11 (cf. Table 4) and simulation 12, 13, 14 (cf. Table 5).

which may be an unreasonably short time interval for each *HBB*S* independent mutations to spread across their current geographic areas. In any case more simulations including wider ranges of demographic models (e.g. range expansions) will be needed to draw more robust conclusions about this issue.

3.3.4 Hypothesis testing

In the previous sections we have compared the relative approximation between observed and simulated data using a limited set of demographic scenarios. This approach was sufficiently informative to discriminate the relative fit of different scenarios, but does not assess the extent to which these scenarios could be accepted or rejected using more standard hypothesis testing approaches.

To implement such approaches, we took into account that the observed dataset is characterized by high frequency of the predominant haplotype (*P*), high intra-population homogeneity (*Hii*) and low inter-population homogeneity (*Hij*), and estimated the probability that a single *HBB*S* mutation generates simulated sectors with the following properties: i) *P* values that are equal or higher than those obtained in the observed dataset; ii) *Hii* values that are equal or higher than observed; and iii) *Hij* values are equal or lower than observed. The models were rejected when the observed summary statistics lied on the upper 95% percentile (for *P* and *Hii*) or lower 5% percentile (*Hij*) of null distributions obtained by different simulation scenarios (equivalent to $p<0.05$ in one-tailed tests).

Figure 30 shows how this approach was used to test the single-mutation scenario implemented in simulation 4. Note that, even though simulation 4 was considered the best-fitting of all compared scenarios, the probability of generating *Hii* and *Hij* values that resemble the observed data under this simulation (Figure 30b and 30c) is lower than 0.05.

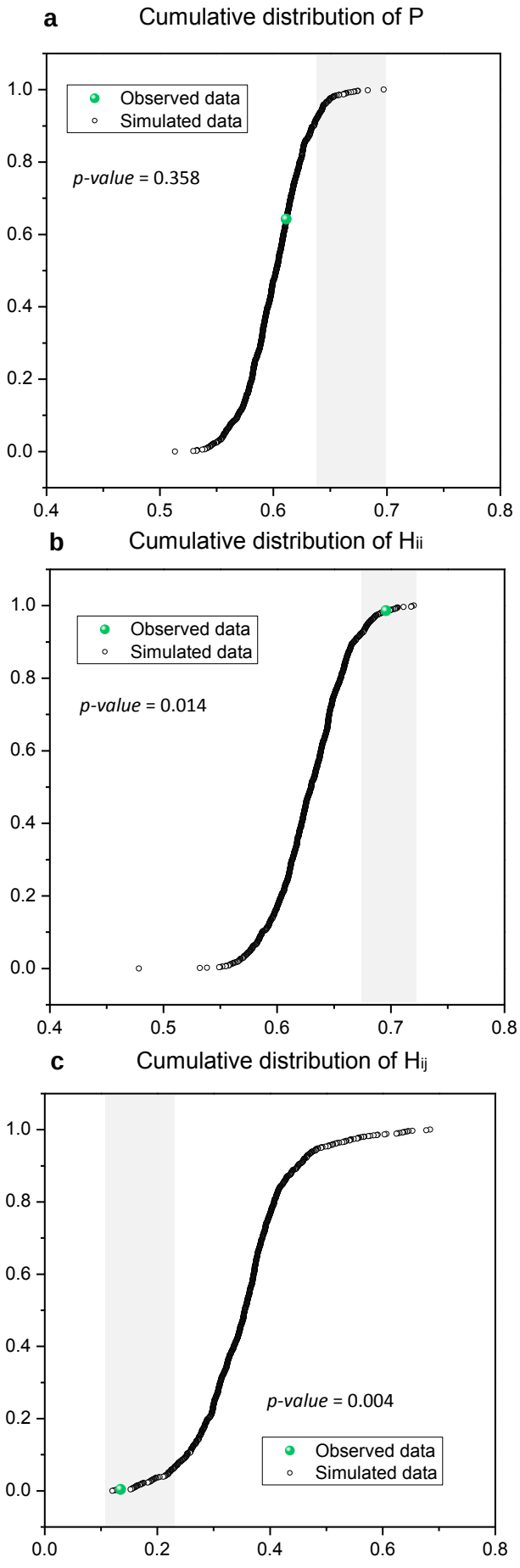

Figure 30 – Cumulative frequency distribution of *P (a)*, *Hii* (b) and Hij (c) obtained by using simulation 4. The shaded area represents the significance level α =0.05. Green point represents observed values.

Table 9 shows the results of the tests applied to all simulated two-dimensional scenarios listed in tables 4 and 5. Simulations 4, 12 and 6 appear to generate *P* values that are close to real data with high probability. Simulation 6 was the only one being consistent with the observed data for *Hii*. However, none of the simulations produced *Hij* values that were sufficiently close to the observed data, indicating that the diffusion conditions that were used in the present work, although being illustrative of the effects of mutation spreading on linked haplotype variation, can still not explain the observed patterns of haplotype diversity.

In the future, the simulation framework that was developed in this work should be used to explore a more comprehensive set of demographic scenarios and provide a more robust framework for assessing alternative evolutionary hypotheses.

Table 9 - Probabilities of obtaining a summary statistic at least as extreme as the one that was observed, assuming that the null hypothesis is true. *P* – Frequency of the predominant haplotype, *Hii* – intra-population homogeneity, *Hij* – inter-population homogeneity.

	P		*Hii*		*Hij*	
	P α[a]	*p*[b]	*Hii* α	*p*[b]	*Hij* α	*p*[c]
sim 4	**0.644**	**0.358**	0.680	0.014	0.220	0.004
sim 5	0.508	0	0.565	0	0.221	0
sim 6	**0.662**	**0.256**	**0.701**	**0.057**	0.175	0.041
sim 7	0.453	0	0.534	0	0.379	0
sim 8	0.371	0	0.476	0	0.211	0
sim 9	0.538	0.004	0.600	0.002	0.218	0.019
sim 10	0.360	0	0.477	0	0.217	0
sim 11	0.296	0	0.436	0	0.210	0
sim 12	**0.652**	**0.343**	0.680	0.023	0.234	0.009
sim 13	0.293	0	0.433	0	0.203	0
sim 14	0.365	0	0.465	0	0.198	0
Observed	0.611		0.696		0.135	

[a] Simulated value corresponding to an alpha of 0.05.

[b] Estimated probability of obtaining simulated values that are greater than or equal to the observed value.

[c] Estimated probability of obtaining simulated values that are lower than or equal to the observed value.

4. Conclusions

In this work, we set up a simulation framework that can be used to study the effects of the spatial diffusion of adaptive mutations on linked haplotype variation. Our major motivation was to understand to what extent the current geographical segregation of different haplotypes linked to the HBB*S variant could be explained by a single mutational origin, instead of resulting from multiple recurrent mutations as is currently assumed.

By exploring a limited set of different simulation scenarios under the *wave of advance* model for the dispersal of an advantageous allele, we were able to show that different predominant S-linked haplotypes may arise at the edges of spatial distribution of a single mutation after as much as 200 generations. The major outcome of this phenomenon is the formation of several, relatively homogenous patches (or sectors) encompassing contiguous populations sharing S-linked modal haplotypes that are different from those observed in other regions. These patterns, which clearly mimic the spatial distribution of different S-haplotypes in Africa, show that, as originally proposed by Livingstone (1989), when a single mutational origin is assumed, the geographical areas where different homogeneous haplotypes predominate must be regarded as edges where the variant has recently arrived from elsewhere.

By comparing the simulated data with an empirical dataset, consisting of haplotype data from several African populations, we were able to evaluate the relative approximation between observed and simulated data, and discriminate the relative fit of different scenarios. However, although this approach was sufficiently informative to indicate that the overall levels of haplotype variation are more likely assuming a single-mutation than multicentric origins, we were still unable to generate simulated single-mutation conditions matching all the characteristics of the observed data with sufficient approximation.

In the future, we plan to take advantage of the simulation framework developed in this thesis to explore additional demographic scenarios, including range expansions of populations, in order to provide a more robust discrimination between competing hypotheses.

5. References

Allison, AC (1954) The distribution of the sickle-cell trait in East Africa and elsewhere, and its apparent relationship to the incidence of subtertian malaria. *Trans. R. Soc. Trop. Med*. *Hyg*. 48, 312-318.

Allison, AC (1964) Polymorphism and natural selection in human populations. *Cold Spring Harb. Symp. Quant. Biol.* 29, 137-149.

Antonarakis SE (1984). Origin of the βS-globin Gene in Blacks: The Contribution of Recurrent Mutation or Gene Conversion or Both*. Proceedings of the National Academy of Sciences*. 81, 853–856.

Beet, EA (1946) Sickle cell disease in the Balovale District of Northern Rhodesia. *East African medical journal* 23, 75.

Carter R, Mendis KN (2002). Evolutionary and historical aspects of the burden of malaria. *Clinical Microbiology Reviews.* 15, 564-594.

Cavalli-Sforza, LL and Bodmer, WF (1971) *The Genetics of Human Populations*. W. H. Freeman, San Francisco, p.148.

Chakravarti A, Buetow IKH, Antonarakis SE, Waber PG, Boehm CD, Kazazian HH (1984) Gene Cluster β-Globin. *American journal of Human Genetics*. 36, 1239–1258.

Currat M, Trabuchet G, Rees D, Perrin P, Harding RM, Clegg JB, Excoffier L (2002) Molecular Analysis of the β-Globin Gene Cluster in the Niokholo Mandenka Population Reveals a Recent Origin of the b S Senegal Mutation. *American journal of human genetics*. 70, 207–223.

Curtin PD (1969) *The Atlantic slave trade: a census.* Milwaukee. University of Wisconsin Press.

Edmonds CA, Lillie AS, Cavalli-Sforza LL (2004) Mutations arising in the wave front of an expanding population. Proceedings of the National Academy of Sciences of the USA. 101, 975–9

Epperson, BK (2003) *Geographical genetics.* (Princeton Univ Press).

Ewens WJ (1972) The sampling theory of selectively neutral alleles. *Theoretical Population Biology.* 3, 87–112.

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in ecology & evolution*. 23, 347–51.

Fisher RA (1937) The wave of advance of advantageous genes. Annals of Eugenics. 7, 355–369.

Flint J, Harding RM, Clegg JB, Boyce AJ (1993) Why are some genetic diseases common? Distinguishing selection from other processes by molecular analysis of globin gene variants. *Human Genetics*. 91, 91–117.

Flint J, Roalind M. Harding RM, Boyce AJ, Clegg JB (1998) The population genetics of the haemoglobinopathies. *Baillière's Clinical Haematology*. 11, 1-51.

François O, Currat M, Ray N, Han E, Excoffier L, Novembre J (2010) Principal component analysis under population genetic models of range expansion and admixture. *Molecular biology and evolution*. 27, 1257–68.

Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A (2001) Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *American journal of human genetics*. 69, 831–43.

Fullerton SM, Harding RM, Boyce AJ, Clegg JB (1994) Molecular and population genetic analysis of allelic sequence diversity at the human β-globin locus. *Proceedings of the National Academy of Sciences USA*. 91, 1805-1809.

Haldane, J. B. S. (1949). The rate of mutation of human genes*. Hereditas*. 35, 267-273.

Hallatschek O, Hersen P, Ramanathan S, Nelson DR (2007) Genetic drift at expanding frontiers promotes gene segregation. Proceedings of the National Academy of Sciences of the United States of America. 104, 19926–30.

Hallatschek O, Nelson DR (2008) Gene surfing in expanding populations. *Theoretical population biology*. 73, 158–70.

Karasov T, Messer PW, Petrov D (2010) Evidence that adaptation in Drosophila is not limited by mutation at single sites. *PLoS genetics*. 6, e1000924.

Kimura M and Weiss GH (1964) The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*. 480, 561–576.

Klopfstein S, Currat M, Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion*. Molecular biology and evolution*. 23, 482–90.

Kolmogorov A, Petrovskii I, Piscunov N (1937) A study of the equation of diffusion with increase in the quantity of matter, and its application to a biological problem. *Byul. Moskovskogo Gos. Univ.* 1 , 1-25.

Kurnit DM (1979) Evolution of sickle variant gene. *The Lancet*. 313, 104.

Lapouméroulie C, Dunda O, Ducrocq R, Trabuchet G, Mony-Lobé M, Bodo JM, Carnevale P, Labie D, Elion J, Krishnamoorthy (1992) A novel sickle cell mutation of yet another origin in Africa: the Cameroon type. *Human genetics*, 89, 333-337.

Livingstone F (1958) Anthropological Implications of Sickle Cell Gene Distribution in West Africa. *American Anthropologist*. 60, 533–562.

Livingstone F (1964) The Distributions of the Abnormal Hemoglobin Genes and Their Significance for Human Evolution. *American Anthropologist*. 18, 685–699.

Livingstone FB (1989) Who gave whom hemoglobin S: The use of restriction site haplotype variation for the interpretation of the evolution of the βS-globin gene. *American Journal of Human Biology*. 1, 289-302.

Moran PAP (1950) Notes on continuous stochastic phenomena. *Biometrika*. 37, 17-23.

Nagel RL, Fabry ME, Pagnier J, Zohoun I, Wacjman H, Baudin V, Labie D (1985) Hematologically and Genetically Distinct Forms of Sickle Cell Anemia in Africa — The Senegal Type and the Benin Type. *The New England Journal of Medicine*. 312, 880–884.

Nagel RL, Ranney HM (1990) Genetic epidemiology of structural mutations of the beta-globin gene. *Seminars in Hematology*. 27, 342-59.

Novembre J, Galvani AP, Slatkin M (2005) The geographic spread of the CCR5 Delta32 HIV-resistance allele. *PLoS biology*. 3, e339.

Pagnier J, Mears JG, Dunda-Belkhodja O, Schaefer-Rego KE, Beldjord C, Nagel RL, Labie D (1984) Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. *Proceedings of the National Academy of Sciences USA.* 81, 1771–1773.

Piel FB, Patil AP, Howes RE, Nyangiri OA, Gething PW, Williams TN, Weatherall DJ, Hay SI (2010) Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis*. Nature communications*. 1, 104.

Powers PA, Smithies O (1986) Short gene conversions in the human fetal globin gene region: a by-product of chromosome pairing during meiosis? *Genetics*, 112, 343-58.

Ralph P, Coop G (2010) Parallel adaptation: One or many waves of advance of an advantageous allele? *Genetics*. 186, 647-668.

Stephens JC, Reich DE, Goldstein DB, et al. (1998) Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *Am J Hum Genet* 62,1507-1515.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American journal of human genetics.* 68, 978–89.

Tomás G, Seco L, Seixas S, Faustino P, Lavinha J, Rocha J (2002) The peopling of São Tomé (Gulf of Guinea): origins of slave settlers and admixture with the Portuguese. *Human biology*. 74, 397–411.

Travis JMJ, Münkemüller T, Burton OJ, Best A, Dytham C, Johst K (2007) Deleterious mutations can surf to high densities on the wave front of an expanding population*. Molecular biology and evolution.* 24, 2334–43.

Weatherall, DJ and Clegg, JB (2001) *The Thalassaemia Syndromes* (Blackwell Science).

Weber JL, Wong C (1993) Mutation of human short tandem repeats. *Human molecular Genetics.* 2, 1123–1128.

Williams TN (2006) Human red blood cell polymorphisms and malaria. *Current opinion in microbiology*. 9, 388-394.

Williams TN, Mwangi TW, Wambua S, Alexander ND, Kortok M, Snow RW, Marsh K (2005) Sickle cell trait and the risk of Plasmodium falciparum malaria and other childhood diseases. *Journal of Infectious Diseases*. 192, 178-186.

Wright S (1943) Isolation by distance. *Genetics*. 28, 114-138.

6. Appendix

6.1 Program manual

The order of variables in a configuration file is more or less arbitrary. However, note that some variables depend on others. In these cases the dependent variables should be defined after those on which they depend. In this document dependencies are given within parenthesis using the word DEPS. Default values are between square brackets ([]). Variables are described in a 'natural' order that can be broken, provided that all dependent variables are defined after those on which they depend.

The only mandatory variable is POPULATIONS, which also describes the model (one- or two-dimensional model). All other variables will assume the default values. If no other variable is defined, no right and left STR markers will be used, but the simulation still proceeds. This is useful to test the spread of the mutation (S) under several migration scenarios and selection models, by manipulating only MIGRATE, AAFITNESS, ASFITNESS and SSFITNESS (additionally, POPSIZE and SINIT may also influence these).

POPULATIONS <num> <num> Number of populations to use. SHOULD BE THE FIRST THING DEFINED in the configuration file. Note that it is possible to build one- or two-dimensional models. For one-dimensional models the second number is optional. Therefore 'POPULATIONS 10' or 'POPULATIONS 10 1' builds a one-dimensional model with 10 populations. 'POPULATIONS 10 4' builds a two-dimensional model with 40 populations, defining an area (matrix) of 10 columns and 4 rows. Population 1 is in the left upper corner of the matrix, and population 40 is in the right lower corner. Hence, in two-dimensional models, populations are read from left to right and from top to bottom of a matrix.

In one-dimensional models the number of populations effectively participating is determined by POPULATIONS. For two-dimensional models, the number of populations may, or may not be equal to POPULATIONS. In two dimensional models POPULATIONS is equal to the product of columns (the first number after keyword POPULATIONS) and rows (the second number after keyword POPULATIONS) defining a rectangular area (matrix). The variable POPMASK allows one to define which cells in this 'area' are effective populations and which are not. Hence, in two-dimensional models the effective number of populations can be less than or equal to POPULATIONS. NOTE that many variables should be initialized for *ALL* putative POPULATIONS defined by the the 'area' and not only for those populations that will actually participate in the 2D model.

SIMULATIONS <num> [1] Define number of simulations.

POPSIZE <array> [1000] (DEPS: POPULATIONS) Array of POPULATIONS sizes for each deme. If no growth rate is specified this is the constant size of each deme. If a

growth rate is specified, this is the initial deme size. For multi population simulations values are given after keyword, e.g., for a three population simulation, this should be 'POPSIZE 1000 1000 1000'. Additionaly, a wildcard (*) can follow the first value, meaning that all subsequent values are equal. For any case where POPULATIONS > 1, 'POPSIZE 1000 *' means that all populations have a size of 1000.

GENERATIONS <num> [100] Number of generations to simulate for each simulation.

MUTANTS <num> <array> [1,'S'] Defines the number of mutant haplotypes and their names. The first parameter is the number of mutant haplotypes (default is one) followed by a list of characters (as many as defined by the number of mutants). Default mutant haplotype is named 'S'. Use only single characters to name mutant haplotypes. Do not use 'A' since it will get confused with background haplotypes. Note that the software is case sensitive, so 'MUTANTS 2 S s' is perfectly valid (two mutants, one called 'S' and the other called 's').

ORIGINALPOP <array> [1] (DEPS: POPULATIONS, MUTANTS) Defines the population where the mutation arises. Should fall between 1 and POPULATIONS. If MUTANTS is larger than 1, then there should be as many numbers as MUTANTS. One may insert two (or more) different mutants in the same population, each with SINIT/POPSIZE.

SINIT <num> [1] Number of chromosomes carrying the mutant allele at the original population. Default is 1 (with frequency 1/POPSIZE) but can be anything from 1 to POPSIZE. Note that in multiple mutant injections (MUTANTS > 1), SINIT is the same for each mutant.

LEFTLOCI <num> [0] and RIGHTLOCI <num> [0] Number of marker loci to the left (LEFTLOCI) or to the right (RIGHTLOCI) of the mutation. There are no limits for both variables, but the fewer the better!

MAXALLELES <array> [10] (DEPS: LEFTLOCI, RIGTHLOCI) Maximum number of alleles per locus. Wildcard notation (*) applies. The maximum number allowed is 220 (alleles are coded as bytes in a string, which would technically allow 256 different alleles; however, the first 32 bytes (0-31) in ASCII represent special codes, with the 0 (zero or null) representing the end of a string! these are not used, so there remain 256-32 = 224 ~ 220 alleles. 'MAXALLELES 10 10 10 10 10' or 'MAXALLELES 10 *' is equivalent for a simulation with 5 loci, each with 10 alleles. Note that the first number corresponds to the leftmost left locus, followed by all other left loci up to the marker, and then from the leftmost right locus up to the rightmost right locus. Hence 'MAXALLELES 3 6 7 3 4' translates into 3-6-7-A-3-4 for a case of LEFTLOCI=3 and RIGHTLOCI=2.

ALLELEFREQS <matrix> (DEPS: LMAXALLELES) Initial frequencies for each allele per locus. If not given, alleles per locus will be equi-frequent [1/(N alleles)]. No wildcard notation can be used! In each separate line one should give allele frequencies for each locus. For a MAXALLELES of '3 4 6' (2 loci left and one right). ALLELE_FREQS should read:

ALLELE_FREQS

0.2 0.3 0.5 <- for locus 1 (leftmost left one)
0.1 0.1 0.3. 0.5 <- for locus 2 (rightmost left one)
0.6 0.4 <- for locus 3 (right one)

AAFITNESS <array> [0.8], ASFITNESS <array> [1.0] and SSFITNESS <array> [0.0] (DEPS: POPULATIONS) Arrays with fitness values of AA, AS, and SS genotypes for each population. For all three cases, the wildcard notation (*) is valid. Thus for three populations the instructions 'AAFITNESS 0.8 0.8 0.8' or 'AAFITNESS 0.8 *' are valid ways to tell that AAFITNESS is the same (0.8) in all three populations.

RIGHTMUTRATE <matrix> (DEPS: POPULATIONS, RIGHTLOCI) A matrix of right loci mutation rates (POPULATIONS*RIGHTLOCI). Populations go in each line, and loci in each column. The wildcard notation (*) can be used but in a different way. Mutation rates should be set for all loci at the right side. After the first line, a wildcard means that all populations have the same mutation rates at each locus. For a four population example, with four loci at the right side,

RIGHTMUTRATE
0.01 0.02 0.003 0.002
0.01 0.02 0.003 0.002
0.01 0.02 0.003 0.002
0.01 0.02 0.003 0.002
and
RIGHTMUTRATE
0.01 0.02 0.003 0.002
*

mean exactly the same. All population will have mutrates of 0.01, 0.02, 0.003, 0.002 for each locus.

LEFTMUTRATE <matrix> (DEPS: POPULATIONS, LEFTTLOCI) A matrix of left loci mutation rates (POPULATIONS*LEFTLOCI). Populations go in each line, and loci in each column. The wildcard notation (*) can be used (see RIGHTMUTRATE).

RIGHTRECRATE <matrix> (DEPS: POPULATIONS, RIGHTLOCI) A matrix of right loci recombination rates (POPULATIONS*RIGHTLOCI). Populations go in each line, and loci in each column. The wildcard notation (*) can be used (see RIGHTMUTRATE).

LEFTRECRATE <matrix> (DEPS: POPULATIONS, LEFTLOCI) A matrix of left loci recombination rates (POPULATIONS*RIGHTLOCI). Populations go in each line, and loci in each column. The wildcard notation (*) can be used (see RIGHTMUTRATE).

CONVERTRATE <array> [0] (DEPS: POPULATIONS) An array with conversion rates for each deme. These are the rates at which A alleles are converted to S. Wildcard notation (*) applies. For a three population simulation (POPULATIONS=3) 'CONVERTRATE 0.01 0.01 0.01' or 'CONVERTRATE 0.01 *' mean the same.

MIGRATE <num> [0.05] Define migration rate between adjacent cells. For a linear model (rows=1) each cell has a total migration rate of 2*MIGRATE, except the first and last cells (for which MIGRATE applies). In a two-dimensional model (rows>1) each cell has a total migration rate of 4*MIGRATE, except for 'corners' (where it is 2*MIGRATE) and 'marginal cells' (where it is 3*MIGRATE). MIGRATE should be between 0 and 1.

STOPMIGATGEN <num> [2000000000] Define a generation where migration stops. Should be less than GENERATIONS. If it is larger than GENERATIONS it will be ignored!

STARTMIGATGEN <num> [1] Define generation when to start migration. Should be less than GENERATIONS. If it is larger, no migration will occur.

GROWTHRATE <array> [0] (DEPS: POPULATIONS) An array with growth rates for each deme (or population). Wildcard notation(*) applies. For a three population simulation (POPULATIONS=3) 'GROWTHRATE 0.5 0.5 0.5' or 'GROWTHRATE 0.5 *' mean the same.

MAXPOPSIZE <array> [1000] (DEPS: POPULATIONS) An array of maximum population size for each deme. When a growth rate is specified, this defines the carrying capacity in the logistic growth equation. The wildcard (*) applies. For a three population simulation (POPULATIONS=3), 'MAXPOPSIZE 1000 1000 1000' or 'MAXPOPSIZE 1000 *' mean the same.

BURNIN <num> [0] Before each simulation (insertion of a mutant) there may be a period of mixing, drifting and purging of background variation of normal haplotypes. This is done during a number of generations defined by BURNIN, after which the real simulation starts.

POPMASK <matrix> (DEPS: POPULATIONS) POPMASK allows one to specify in a two dimensional model which populations will participate in the model. This is only useful if POPULATIONS was given in the form 'POPULATIONS <cols> <rows>' and 'rows' > 1. For example, in a model with 16 populations where POPULATIONS was defined as 'POPULATIONS 4 4' variable POPMASK would be defined as,

```
POPMASK
0 1 1 0
1 1 1 1
1 1 1 0
1 1 0 0
```

to set a scenario where the first, fourth, 12th, 15th, and 16th cells would not be used (no populations created there).

VERBOSE [FALSE] Produce very verbose output (to stderr) during the initialization of variables.

OUPUTALL [FALSE] If this keyword is provided all haplotypes (and not only the mutant haplotypes) will be output. Note that this option is essentially for debugging purposes, and will produce very large data files.

SIMOFFSET <num> [0] Simulations are numbered from 0 (zero) to SIMULATIONS-1. There are cases where it is necessary to start output numbering at higher values. If the program is run separately in two different machines, each with 500 simulations, using SIMOFFSET 500 in one of them will allow subsequent merging of the two files without duplicating simulation numbers. The second run will start at simulation 500 up to 999.

TIME_SLICES <num> [0] [DEPS: GENERATIONS] Normally, the program only outputs results at the end of a full run (after g GENERATIONS). To produce output at different times before reaching the desired number of generations one can provide this variable. There is, however, a *CATCH* (see below). TIME_SLICES is limited internally to 10 slices maximum (since this is the number of files that will be opened for writing during the run of the simulation). This can be altered in 'config.h' (MAX_TIME_SLICES) but implies recompilation of the software. The algorithm is as follows:

N = GENERATIONS / (TIME_SLICES + 1) Eq. $1
M = N Eq. $2

The simulation starts and runs until the current number of generations (g) is greater than M (g > M). It outputs the results to a individual file, increments M by N (M=M+N) and continues until g > M. The procedure is repeated until g<(GENERATIONS-1) avoiding the duplication of the output of the last iteration. Note that because of the way N is computed (Eq. $1), providing a TIME_SLICES of 10 when GENERATIONS is 100 will produce 10 time slices plus the final output, meaning that each time slice will span ~9 generations and not 10 generations.

With GENEARATIONS 100 and TIME_SLICES 10 the program outputs data at generations 9, 18, 27, 36, 45, 54, 63, 72, 81 and 90, and finally at generation 99 (10 time slices plus the final output). Since generation numbers start at zero (1st generation is numbered 0, 100th generation is numbered 99) the first slice spans 10 and not 9 generations [note that the first output is at generation 9 (10th) and not 8 (9th)]. To produce slices spanning 10 generations for a simulation with GENERATIONS = 100, one should use TIME_SLICES = 9. The output will be at generations 10, 20, 30, 40, 50, 60, 70, 80 and 90, plus the final output at generation 99 [Note again that the first output spans 11 generations and not 10!].

CATCH (IMPORTANT!): The files to output time sliced data are opened at the beginning of the run, and remain opened until all simulations are done. This avoids excessive opening/closing of files which, in turn, will slow down the run. The problem is that for each individual simulation there are instances where it has to be reset due to loss of mutant haplotypes. However, if this loss happens after some time slices, the state results for those time slices will already be written in the corresponding output files. Always check for duplicated results of simulation/population in earlier time slices!

This is more likely for smaller than larger time slices (i.e., time slices spanning less generations). To put it in other words, when TIME_SLICES is large (combined with a small value for GENERATIONS) each time slice will span less generations and the chances of earlier sliced data having duplicated results is higher.

6.2 The wave speed of an advantageous allele

The spread of favorable mutations with the formation of travelling waves was first modeled by fisher (1937) and by Kolmogorov, Petrovskii, and Piscunov (Kolmogorov et al. 1937), independently. Assuming that the dispersal distance is Gaussian, the advantage of the mutation is additive and the population density is large enough so that stochastic effects are small, they showed that the wave of advance resulting from the evolution of the mutant frequency in time and space moved asymptotically with speed,

$$v = \sigma\sqrt{2s}$$

where σ is the standard deviation of the dispersal distance and s is the selective advantage. Under the above assumptions the speed of the advancing wave is constant and identical for the one-dimensional case and for radially symmetric waves (two dimensions).

Contrarily to this model of continuous time and space, the simulations presented here use discrete generations and discontinuous space. Nevertheless, results obtained with the discrete stepping stone model can be adapted to the former model if we consider the limiting case in which both generation time and the actual distance between adjacent subgroups approach zero (Kimura & Weiss, 1964; Ralph & Coop, 2010). Thereby, σ must be replaced by $\sqrt{m2^{-(d-1)}}$ (Ralph & Coop, 2010), where *d* is the number of dimensions, to satisfy the Fisher-KPP equation.