Language Sciences 93 (2022) 101497

Contents lists available at ScienceDirect

Language Sciences

journal homepage: www.elsevier.com/locate/langsci

Linguistic clues suggest that the Indonesian colonizers directly sailed to Madagascar

Maurizio Serva*, Michele Pasquini

Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica, Università degli Studi dell'Aquila, L'Aquila, Italy

ARTICLE INFO

Article history:
Received 6 April 2022
Received in revised form 13 June 2022
Accepted 22 June 2022
Available online 8 July 2022

Keywords:
Settlement of Madagascar
Malagasy dialects
Austronesian languages
Bantu languages

ABSTRACT

The Malagasy language belongs to the Austronesian Family and it is particularly close to some of the languages spoken in Indonesia, a fact that was first noticed at the beginning of the XVII*th* century. The link to a precise Indonesian language is due to Dahl who, in 1951, firmly established a striking kinship with Maanyan, spoken in the South-East of Kalimantan. The introgression of Bantu terms is extremely limited, on the contrary the genetic makeup of the Malagasy people is African and Indonesian with comparable proportions. While genetics and linguistics agree that the colonization of Madagascar by Indonesian sailors took place in the second half of the first millennium, they disagree concerning the role of East-Africa in this event. Here we show that the dichotomy emerges because linguistics uses qualitative arguments where genetics has a consolidated tradition in the use of quantitative methods. After having collected the largest and most complete existing dataset for Malagasy, covering the entire island (207-terms Swadesh lists of 60 different dialects), we adopt new quantitative tools that allow us to confirm the genetics point of view that Indonesian sailors directly colonized Madagascar, without the East-African stopover conjectured in various studies in linguistics. The key point of our approach is the analysis of the geographical distribution of the degree of Bantu languages contamination of Malagasy dialects.

1. Preamble

Madagascar and Comoros are the Western edge of the Austronesian expansion, which is probably the biggest event of maritime exploration and colonization in human history (Blust, 2019).

All dialects spoken in Madagascar, as well two dialects spoken in Mayotte, are varieties of Malagasy, a language that belongs to the Malayo-Polynesian branch of the Austronesian family. It is usually mentioned that the Dutch merchant Frederick de Houtman van Gouda was the first westerner to report around 1600 *C E* that the Malagasy natives speak a language similar to Malay [de Houtman, 1603]. A few years later, the Portuguese Jesuit Luis Mariano arrived to the identical conclusion (Mariano, 1613). However, they probably weren't the first ever, since this similarity was perhaps already remarked, as early as the twelfth century *C E*, by the Arab geographer al-Idrīsī (Tibbetts, 1979).

The clear link with a specific language is due to the Norwegian missionary Otto Christian Dahl (Dahl, 1951) who around the middle of the last century established a striking kinship between Malagasy and Maanyan, a language spoken in the South-East of

* Corresponding author. DISIM, Università degli Studi dell'Aquila, Via Vetoio 1, 67100 L'Aquila, Italy.
E-mail address: serva@univaq.it (M. Serva).

https://doi.org/10.1016/j.langsci.2022.101497
0388-0001/

Borneo, nearby to the Barito river. After 70 years, this special connection, placing Malagasy in the East-Barito group of languages, together with Maanyan and another dozen of varieties spoken in South-East Borneo, is still undisputed (Dahl, 1977).

Words borrowed from other Indonesian languages are also detectable in Malagasy, especially from Malay; Dahl himself addressed the issue (Dahl, 1951; Dahl, 1977; Dahl, 1991), but the systematic study of this topic was performed by Alexander Adelaar (Adelaar, 1989; Adelaar, 1994; Adelaar, 2006). More significant for the subject of this article is that a limited number of Bantu[1] loanwords is also traceable in the language (Dahl, 1951; Dahl, 1954; Adelaar, 2007; Adelaar, 2016; Adelaar, 2017; Adelaar, 2022).

While the Malagasy language is Austronesian, with only small traces of Bantu, the genetic makeup of Malagasy people is Asian and African with comparable proportions, possibly with an African preponderance (Hurles et al. 2005; Tofanelli et al., 2009; Razafindrazaka et al., 2010; Cox et al., 2012; Pierron et al., 2014; Pierron et al., 2017; Brucato et al., 2019; Heiske et al., 2021). For what concerns the Asian component, genetics, as well as linguistics, points to Indonesia (Soodyall et al., 1995; Cox et al., 2012; Kusuma et al., 2015), and to be precise, again to South-Eastern Borneo. However, the privileged connection is not with the Maanyan ethnicity but with the Banjar people, who speak a Malay language (Kusuma et al., 2016). This discrepancy can be possibly explained by the presence of an historical Malay trading Post in South-Eastern Borneo, which favored the admixture between the Malay and the autochthonous populations (Brucato et al., 2016).

The picture which emerges from linguistics and genetics is that the Banjar, who spoke a language related to proto-Maanyan, reached Madagascar by a Malay-led expedition bringing the language and becoming the Asian source of the Malagasy gene pool. Later, in Borneo, the Banjar switched to Malay language under the predominant influence of the trading post.

Both genetics and linguistics point to a colonization event occurred in the second half of the first millennium. Concerning linguistics, the argument comes both from careful timing of borrowings, which is derived from the known history of the Indian Ocean and especially of the Srivijaya empire (Adelaar, 1989; Adelaar, 2022) and from the application of new quantitative methodologies inspired by, but nevertheless different from, classical lexicostatistics and glottochronology (Serva *et al.*, 2012, Serva, 2012, Serva and Pasquini, 2020).

However, genetics and linguistics disagree on the role of East-Africa for the colonization of Madagascar. Dahl (Dahl, 1951) initially favored a direct migration from Borneo to Madagascar but soon abandoned this idea and the prevalent picture in linguistics seems to be now a colonization after a stopover in an East African outpost where locals and seafarers interbreed and Bantu words introgressed into Malagasy (Deschamps, 1960; Adelaar, 2017; Adelaar, 2022). On the contrary, genetics is for an Austronesian colonization followed by an arrival of Bantu-speaking groups in Madagascar, where the interbreeding between the new immigrants and the Indonesian colonizers took place. This opinion mostly derives from the observation that the distribution of the Asian and African ancestral components is geographically biased, and also from the observation that there is a marked sex disparity in the Asian/African proportion (Hurles et al. 2005; Tofanelli et al., 2009; Razafindrazaka et al., 2010; Pierron et al., 2014; Pierron et al., 2017; Brucato et al., 2019; Heiske et al., 2021).

We argue in this paper that linguistics also converges to a scenario with an Austronesian colonization followed by a Bantu speakers immigration if a quantitative approach, instead of a qualitative approach, is adopted. The key point is the quantitative representation of the geographical distribution of the rate of Bantu loanwords in the Malagasy language. Ideally, our approach corresponds to the geneticists idea of detecting differences in the genomic composition of the various Malagasy ethnicities, which basically means geographical differences.

2. The theater of events and dataset

To pursue our goal, we have considered 60 different Malagasy varieties scattered all around the Island, 16 Bantu languages in East-Africa and 4 in the Comoros, For each of these 80 varieties we considered a 207-terms Swadesh list, overall about 16,500 terms.

The Malagasy 60 varieties dataset (12,420 terms), which was entirely collected by one of the authors during the years 2018 and 2019, covers the whole Island and it is by far the most complete and extensive for this language. Each list corresponds to a different Malagasy variety, which is not simply identified by the name of the ethnicity, but also by the location where the variety was collected. In turn, the location is identified by the name of a town/village and by latitude and longitude. Each list was furnished and checked at least by three native language speakers which, for each given concept, were asked to furnish the most common word in their dialect as spoken in their town/village. The 60 Swadesh lists, together with the indication of ethnicities, towns/villages and latitudes and longitudes, can be found in the Supporting Information of paper (Serva and Pasquini, 2020).

The main core of the East-African Bantu languages dataset concerns 15 Mozambican languages (mostly North of the Zambezi river) and one (standard Swahili) in Zambia. The data were mostly drawn from the series of dictionaries produced by *SIL Moçambique* (Nampula, Moçambique) and integrated and corrected by Professors Aurelio Simango, Narciso Gastene and Davety Mpiuka of the Universidade Eduardo Mondlane in Maputo. Moreover, the lists of Mwani (207 terms), spoken in Capo

[1] The term "Bantu" is becoming increasingly problematic as it conflates ethnic and linguistic connotations. In this work we use this term exclusively with a linguistic meaning. Probably the linguistic synonym "Sintu" would be more appropriate, however we leave the term "Bantu" only because it is still more widespread and known.

Delgado, was integrated by data collected on the field by one of the authors, while the Swahili of Palma (207 terms) was entirely collected on the field by the same author. This dataset can be found in Supplementary Material (Dataset S1).

Finally, the 207 × 4 terms of the four Sabaki Comorian languages were taken from the *ORELC* online dictionary www.swadrii.com/orelc/ShikomoriWords/?i=kmWords, when possible these terms have been cross-checked by consulting other online dictionaries. The corresponding Swadesh lists can be also found in Supplementary Material (Dataset S1).

A complete overview of the geographical locations of the 80 languages, together with some basic information, can be appreciated in Fig. 1. Only the names of the East-African and Comorian towns are reported, while the Malagasy towns are only indicated by numbers. The link between numbers, languages/ethnicities, towns/villages and geodesic coordinates of all the 80 varieties can be found in Supplementary Material (Table S1).

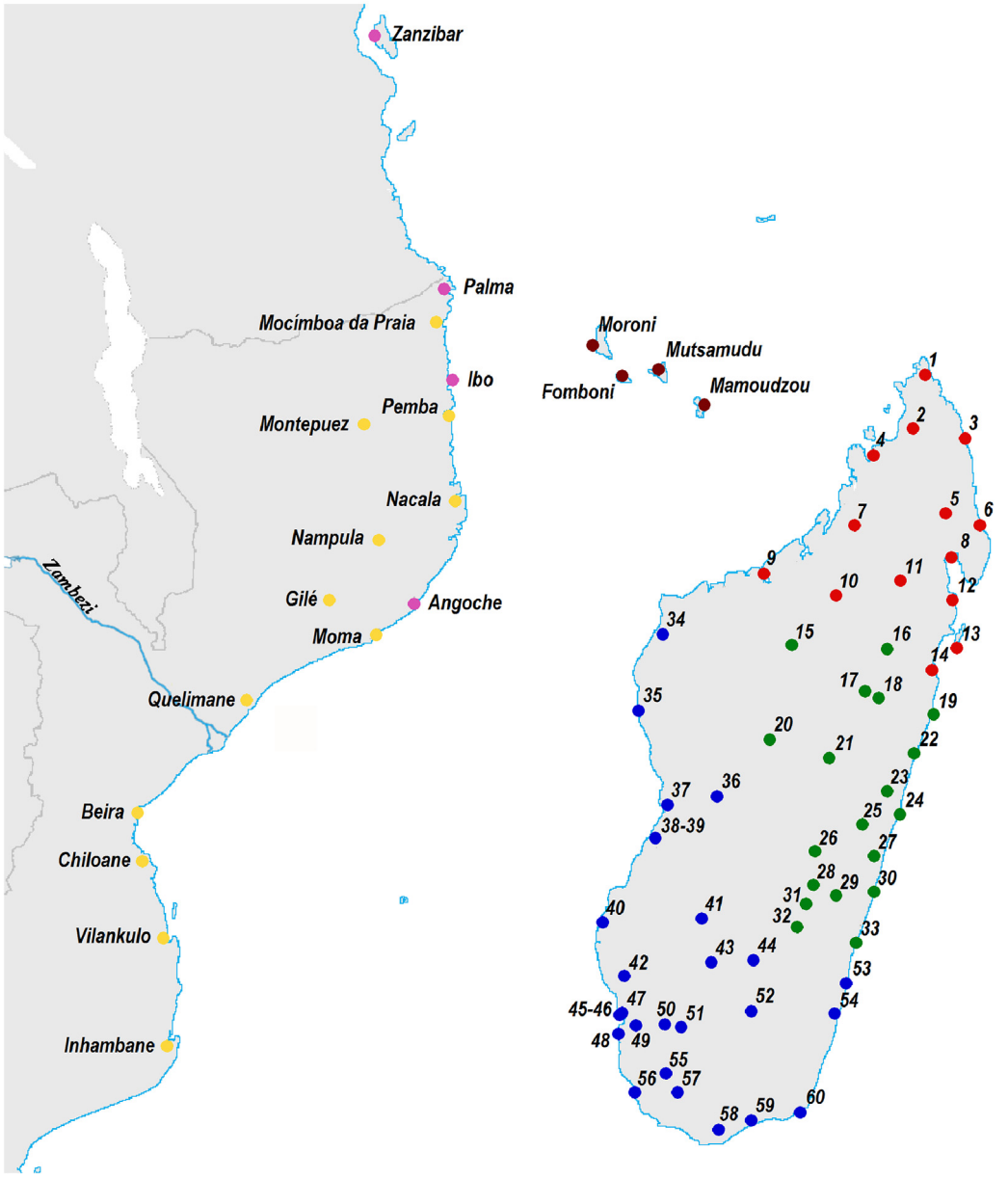

Fig. 1. A complete overview of the geographical locations of the 80 languages. Only the names of the East-African and Comorian towns are reported while the Malagasy towns are only indicated by numbers. The link between numbers, languages/ethnicities, towns/villages and geodesic coordinates of all the 80 varieties can be found in Supplementary Material (Table S1). The purple (East Africa) and brown spots (Comoros) refer to Sabaki languages, while the yellow spots refer to Makhuwa and other non-Sabaki Bantu languages. The three different colors red, green and blue for the Malagasy varieties refer to the three main branches of their phylogenetic tree (Serva and Pasquini, 2020). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Although Bantu loanwords in the Malagasy languages indisputably indicate a contact between the Austronesian settlers and the Bantu-speaking populations of East-Africa and Comoros, the exact geographical origin of the African loans remains largely unknown. Contact with populations North of the Zambezi River is typically favored (see chart of Fig. 1). Nevertheless, given the incertitude concerning the source of loans, we have considered a much larger geographical area which also includes languages and populations South of the Zambezi river. The purple (East Africa) and brown spots (Comoros) refer to Sabaki languages, while the yellow spots refer to Makhuwa and other non-Sabaki Bantu languages (see (Serva and Pasquini, 2021b) for a cladistics of these languages).

As it can be also appreciated in Fig. 1, the 60 Malagasy varieties uniformly cover the Island, allowing a high resolution geographical analysis of Bantu loanwords all over Madagascar. The different colors refer to the three main branches of the phylogenetic tree of the Malagasy constellation of varieties (Serva and Pasquini, 2020).

3. The introgression of Bantu words into the constellation of Malagasy varieties

This section contains the main piece of evidence that Madagascar was independently reached by African and Asian settlers, rather than colonized by an already admixed population.

Our argument goes as follows:

- As a first step, we determine the number $n(\alpha)$ of Bantu terms that can be detected in the Malagasy Swadesh list corresponding to the dialect $\alpha(\alpha = 1, \ldots., 60)$. To obtain this result, we compare each of the 207 terms which corresponds to the same 207 concepts of each Malagasy variety with the corresponding terms of the 20 Bantu languages. Our procedure is automatic and is described in the Methods section;
- The second step is to assign to any Malagasy geographical location x, identified by latitude and longitude, an intensity $B(x)$ of Bantu words introgression which is obtained by an average of the $n(\alpha)$ of the varieties spoken in the nearby tows/villages.

The average is exponentially weighted according to

$$p(x, x_\alpha) = \frac{e^{-\frac{k(x,x_\alpha)}{d}}}{\sum_{\alpha=1}^{60} e^{-\frac{k(x,x_\alpha)}{d}}} \quad \rightarrow \quad B(x) = \sum_{\alpha=1}^{60} p(x, x_\alpha)\, n(\alpha), \tag{1}$$

where $k(x, x_\alpha)$ is the great-circle distances in Km between a generic point of the Island with geodesic coordinates x and the town/village with coordinates x_α. The proximity parameter d tunes the degree of localization of the intensity, in fact, it substantially cuts off the average those town/villages whose distance from x exceeds d (to be more precise, cuts off the average those town/villages whose distance from x exceeds by d the geographical distance between position x and the closest town/village). Fig. 2 shows the intensity $B(x)$ of Bantu words introgression corresponding to the choice $d = 200$ Km.

Fig. 2 definitely highlights a heterogeneous distribution of the intensity of Bantu words contamination of Malagasy dialects, with a clear gradient from North (higher Bantu languages influence) to South (lower). In particular, the South-Eastern region is the one with the lowest intensity of Bantu vocabulary introgression. We conclude on the basis of this unequal distribution that the penetration of Bantu terms took place in Madagascar and not in South-Eastern Africa. If the second hypothesis were true, all Malagasy dialects would show, more or less, the same rate of Bantu words introgression, with small random fluctuations here and there, uncorrelated with respect to the geographical location. On the contrary, given the very large dataset with 60 different Malagasy, which uniformly cover the entire island, there is no way to explain the observed regular North-South gradient by statistical fluctuations.

Moreover, we can deduce that the arrival of Bantu-speaking immigrants which followed the Austronesian colonization mostly took place in the North and North-West of Madagascar. On the contrary, as we will show in the next section, the landing of Austronesian colonizers took place in the South or South-East.

We have to say that the result of Fig. 2 very weakly depends on the choice of the parameter d when it varies in a reasonable range, which means that it has to be reasonably smaller than the linear dimensions of Madagascar (a thousand of Km), and sufficiently large in order that the average contains more than a single Malagasy dialect (few dozens of Km).

Mirroring conclusions concerning the African-Indonesian genetic admixture were reached by studies concerning the genetics of the Malagasy population, for example in (Pierron et al., 2017) it is stated: “The distribution of African and Asian ancestry across the island reveals that the admixture was sex biased and happened heterogeneously across Madagascar, suggesting independent colonization of Madagascar from Africa and Asia rather than settlement by an already admixed population.”

The arguments used in linguistics to support the competing hypothesis of a centennial stopover on the South-Eastern coast of Africa are essentially of two types. The first is that Africa was part of the Indonesian maritime trading network prior to colonization of Madagascar, and therefore it is not clear why the Indonesian colons should have stopped in Madagascar given that the African continent offered much more appetizing opportunities for colonization and/or trading. The natural objection to this argument is that unlike East-Africa which was already settled by Bantu-speaking populations with an advanced culture, Madagascar was probably empty or possibly inhabited by pre-Neolithic populations.

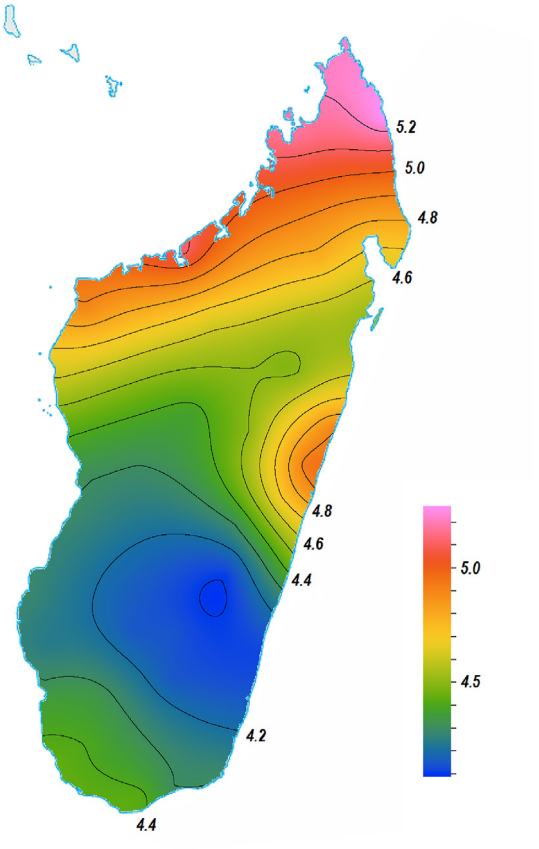

Fig. 2. The intensity $B(x)$ of Bantu words introgression corresponding to the proximity parameter $d = 200$ Km. The black iso-introgression lines connect locations of equal $B(x)$. The figure shows a heterogeneous distribution of the intensity of Bantu contamination of Malagasy dialects, with a clear gradient from North (higher Bantu languages influence) to South (lower). This unequal distribution implies that the penetration of Bantu terms took place in Madagascar and not in South-Eastern Africa. Moreover, one can deduce that the arrival of Bantu-speaking immigrants, which followed the Austronesian colonization, mostly took place in the North and North-West of Madagascar. This figure and the followings make use of map CET-R1, a perceptually uniform colour map created by P. Kovesi (see colorcet.com and (Kovesi, 2015) for further information). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

The second argument used in linguistics is the presence of Bantu names for domestic animals and some cultivated plants. There is, indeed, a clear evidence of this presence, but we don't know why this should be a proof of a stopover in Africa. When the Indonesian colonizers reached Madagascar, they were probably unable to carry many of the domestic animals and some cultivated plants directly from Indonesia.

An interesting contribution to the understanding of this point comes from (Cox et al., 2012), where it is proposed a "scenario in which Madagascar was settled approximately 1200 years ago by a very small group of women (approx. 30), most of Indonesian descent (approx. 93%). This highly restricted founding population raises the possibility that Madagascar was settled not as a large-scale planned colonization event from Indonesia, but rather through a small, perhaps even unintended, transoceanic crossing."

According to this scenario but also to much less extreme ones, some cultivated plants and many domestic animals were likely introduced in a second time from the South-Eastern coast of Africa. This also includes those plants and animals which reached Africa from South-Eastern Asia, as *banana* which is named *akondro* in official Malagasy, a word of Bantu origin. This fact is taken as an argument for an African stopover, nevertheless, it should be noticed that the word *akondro* is used along with Austronesian words as, for example, *kida* in then South-Westtof Madagascar and *katakata* in North-West (see (Donohue and Denham 2009), page 331, for other Austronesian words in the same class 19).

We would like to mention that not only genetics but also history and anthropology points to a direct colonization of Madagascar. For example, Philippe Beaujard argued, on the basis of the evidence of the absence of Swahili-like pottery in the first times of Malagasy presence on the Red Island, that the Indonesian colonizers came into contact with Bantu speakers only after their arrival in Madagascar (Beaujard, 2011). He also writes "Another archaeological element that argues in favor of the direct arrival of the Pre-Malagasy on the island is the presence of pottery with Arca seashell-impressed designs on the oldest sites in Madagascar." In addition, he argues that a number of linguistic data, such as the transfer of the term *lambo* (from Malay, *lembu* = *ox*) onto the bushpig have a difficult explanation if colonization of Madagascar was preceded by a stopover on the African coast.

Finally, it is interesting to observe that the movement of human beings (including sailing) is mathematically described by anomalous diffusion. This fact also points to a direct colonization of Madagascar, since an African stopover is compatible with Brownian motion but it is not with Lévy flights (see (Viswanathan et al., 2011), chapter 7).

In the next section, we consider more quantitative evidence of a direct colonization, as attested by *Diversity*.

4. Diversity also points to a direct colonization of Madagascar

Diversity measures the degree of genetic or linguistic heterogeneity in a given geographical region for a species or a language.

The idea, is that the homeland of a biological species (or genera, or other group) or of a language (or family, or other group) corresponds to the loci with the greatest Diversity. This idea was proposed about one century ago in biology (Vavilov, 1926) and in linguistics (Sapir, 1916), and it is nowadays widely accepted. A well known example is the genetic Diversity of Modern Humans, which is much higher in Africa, where our species originated, than in the other continents.

Therefore, the center of dispersal of Malagasy variants, likely coinciding with the landing spot, can be inferred by finding the site with the largest Diversity.

There are different ways to assess Diversity, but a meaningful quantity should be defined by comparing geographical and lexical distances. Larger Diversity means larger lexical distances given comparable geographical distances. In (Serva *et al.*, 2012) the authors gave a definition which allowed them to assign a value of Diversity to each locus where a Swadesh list was collected, and the clear conclusion was that the largest Diversity could be observed for towns/villages in the South and South-East of the Island.

Recently, an improved definition (Serva and Pasquini, 2020) was proposed, which fulfill some important requirements. First, the Diversity $D(x)$ can be computed for every Malagasy geodesic coordinate x and not only for the 60 different loci

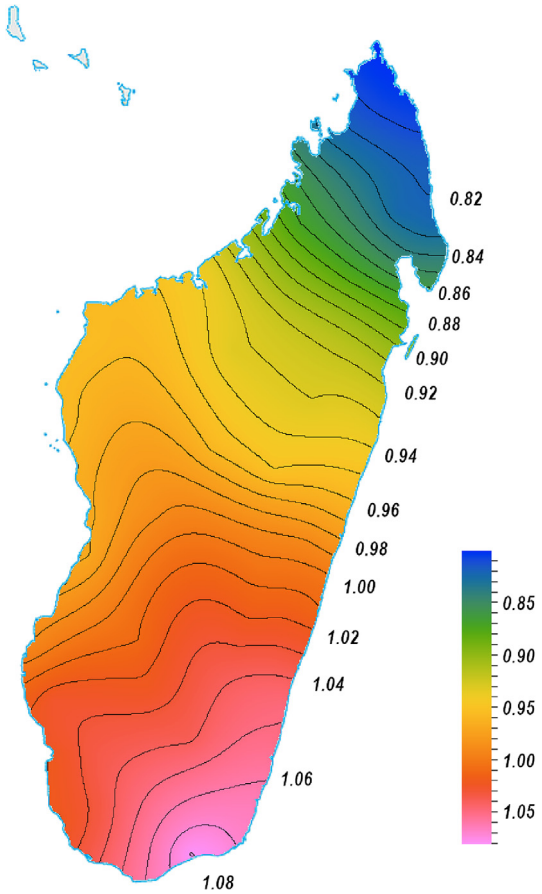

Fig. 3. Diversity $D(x)$ computed with the same proximity parameter of Fig. 2. The black iso-Diversity lines connect locations of equal $D(x)$. At variance with the intensity $B(x)$, the Diversity $D(x)$ progressively decreases going from South or South-East to North, in particular, the extreme North sees a severe reduction of Diversity while South or to South-East show the largest values. Therefore, this picture points to South or to South-East as the dispersal center of Malagasy variants and, consequently, as the landing spot of the Asian settlers. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

(towns or villages) where the dialects were collected. Moreover, Diversity becomes a locally determined quantity as usually occurs in Biology: the towns/villages further away from the geodesic coordinate x contribute exponentially less to the formation of $D(x)$ than the nearest towns/villages.

Given that in (Serva and Pasquini, 2020) a refined definition of Diversity was used and also given that the research took advantage of a much larger dataset, the same qualitative results as in (Serva *et al.*, 2012) was found, but with a much higher resolution. As it can be appreciated in Fig. 3, that we have generated here using the same proximity parameter of Fig. 2, there is a very clear reduction of Diversity going from South or South-East to North.

Therefore, the landing spot was likely in the Southern or South-Eastern coast of Madagascar. This landing spot was also proposed by historians and anthropologists (see, for example (Beaujard, 2012),).

After that, the language started a process of diversification, which led to modern varieties. The relationships among these varieties probably reflect the historical process of internal colonization, which is a phenomenon not yet deciphered.

The identification of the South-Eastern coast of Madagascar as the landing spot of the Asian ancestors of nowadays Malagasy people is corroborated by other observations. One of the major currents in the Indian Ocean is the South Equatorial Current that goes from Sumatra to Madagascar. When Mount Krakatoa erupted in 1883, pumice was transported to the east coast of Madagascar, where the Mananjary river empties into the sea (between Farafangana and Mahanoro). During the Second World War, pieces of wreckage from ships sailing between Java and Sumatra which had been bombed by the Japanese air force also arrived in South-Eastern coast (Faublée, 1970). According to these facts, the ancestors of nowadays Malagasy people probably passed by the easily navigable Sunda strait and, with the help of the South Equatorial Current, they reached, intentionally or unintentionally, the South-East coast of Madagascar.

The South-East landing spot hypotheses adds fuel to the all-in-one-voyage conjecture, since an intermediate stopover in the East African coast would have more likely implied landing on the North-Western coast.

5. Comorians and other Bantu-speaking immigrants reached Madagascar in different waves

In this section we try to detail the settlement of Madagascar by Africans, separately considering the linguistic contribution of different populations. Namely, we consider apart the Comoros Sabaki languages (Swahili group); the East-Africa Sabaki languages (Swahili group) and the East-Africa non-Sabaki languages (included Makhuwa dialects). To attain this goal, we have to consider separately the three Bantu language groups and compute, for each of them, the number of terms which are lent to each of the Malagasy dialects α.

Let us call $n_c(\alpha)$ (Comoros), $n_s(\alpha)$ (East Africa Sabaki) and $n_b(\alpha)$ (East Africa non-Sabaki) these numbers which compose three sets of 60 items. Each of the three numbers $n_c(\alpha)$, $n_s(\alpha)$ and $n_b(\alpha)$, is obviously smaller or equal to $n(\alpha)$, since the terms lent by a single group to each Malagasy variety α form a sub-ensemble of those lent from by the entire collection of 20 Bantu languages. Moreover, the inequality $n(\alpha) \leq n_c(\alpha) + n_s(\alpha) + n_b(\alpha)$ is also satisfied for any α, since a term in the variety α may find correspondence in more than a single group.

We can compute the associated intensity as in (1) and obtain $B_c(x)$, $B_s(x)$ and $B_b(x)$ which also satisfy $B(x) \leq B_c(x) + B_s(x) + B_b(x)$. The result can be appreciated in Fig. 4.

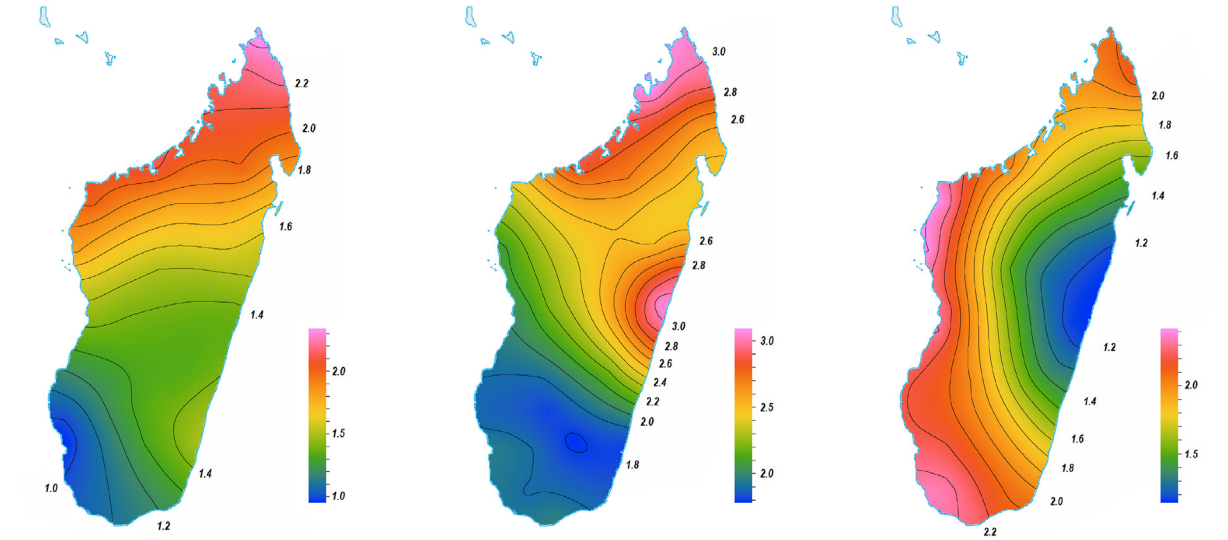

Fig. 4. From left to right the intensities $B_s(x)$ (East-Africa Sabaki), $B_b(x)$ (East-Africa non-Sabaki) and $B_b(x)$ (Comoros Sabaki) with $d = 200$ Km. East-Africa languages, both Sabaki and non-Sabaki, form a very similar pattern, with decreasing intensity of introgression from North to South (the difference consists in a further point of introgression in the Eastern coast for non-Sabaky languages). The Comorian group, on the contrary, forms a totally different pattern with decreasing intensity from West to East. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

The figure shows that East-Africa languages, both Sabaki and non-Sabaki form a very similar pattern, with decreasing intensity of introgression from North to South (the difference consists in a further point of introgression in the Eastern coast for non-Sabaky languages). The Comorian group, on the contrary, forms a different pattern with decreasing intensity from West to East. Notice that the global pattern in Fig. 2 is largely dominated by the 16 East-Africa languages, since the 4 Comorian languages contribute only for one fourth of the total amount of Bantu loans.

The different pattern generated by the Comoros suggests that the contribution of languages of this archipelago belongs to a different wave of introgression, separated in space and, possibly, time from the East African wave.

The Comoros archipelago consists of four islands which lie between the North-West of Madagascar and the South-Eastern coast of Africa. Three of them, Ngazidja (Grande-Comore), Mwali (Mohéli) and Ndzuani (Anjouan), became independent in 1975, and they currently form a single state entity named Udzima wa Komori (Union des Comores). Maore (Mayotte), on the contrary, chose to remain under French administration, however, on a cultural, ethnical and even religious level, it is a part of the same whole as the three other islands.

The four Sabaki varieties spoken in the archipelago are collectively called Shimasiwa, which means *language of the archipelago*, or Shikomori which means *language of the Comoros*. They are the mother languages of the vast majority of inhabitants of the four islands, excepting for Mayotte, where about one third of the natives speaks a Malagasy language (two different varieties) rather than a Bantu language. Nevertheless, although the Malagasy presence in the Comoros dates back to at least one millennium (Allibert, 2015), the Malagasy language presently spoken in Mayotte (Kibosy) seems to be the output of recent events, not much older than a couple of centuries (see (Gueunier, 2004) by the authority on Malagasy dialects on Mayotte Noël Gueunier and also (Serva and Pasquini, 2021a)).

In (Serva and Pasquini, 2021a), it was found that Kiantalaotsy, the oldest of the two Malagasy dialects spoken in Mayotte, is almost equally lexically close to all Western and South-Western Sakalava dialects of Madagascar (Besalampy and Maintirano Sakalavas being the closest). The region where these dialects are spoken is also the region with higher intensity of introgression of the Comorian languages (Fig. 4, right panel). This is probably a consequence of the fact that Kiantalaotsy originated by commercial contacts leaded by Antalaotse, Muslim traders scattered along all the West coast of Madagascar and Mayotte, where the presence of an Antalaotse outpost is attested at the very beginning of the XIX*th* century or even before. The commercial network of Antalaotse traders between the Comoros and the Western coast of Madagascar, active till very recent times, probably vehiculated the introgression of Comorian terms into the Malagasy varieties starting from the West.

6. Methods

The complete information we used for the research consists in 80 Swadesh lists of 207 terms. The lists correspond to 60 different varieties of Malagasy (collected by one of the authors, by far the most complete and extensive dataset for this language) and 20 Bantu languages (15 from Mozambique, 4 from the Comoros and one from Tanzania). Each of the 207 words in each list translates the same corresponding concept in a given language.

The point is: given a Malagasy variety α, how many of the 207 terms of this language also appear in at least one of the 20 Bantu languages. This research can be done using cognate techniques developed in Lexicostatistics starting from the fifties of the last century. We prefer to adopt an automatic research with has the advantage to be more objective and much faster to use.

First, we have to decide when two words, corresponding to the same concept in two different languages are the same, where *same* means that one language has borrowed it from the other, or they have a common origin. In order to avoid false positive or negative, we skip those concepts whose words are typically too short for an objective comparison (match or mismatch can be due by chance). Therefore, we skip the 23 preposition, adverbs, conjunctions and personal pronouns listed from 1 to 16 and from 201 to 207 in Swadesh lists (see Dataset S1 in Supplementary Material). We also skip the concepts *snow* and *ice* which are poorly represented in Malagasy and African languages and the concepts *mother* and *father*, whose terms are often randomly borrowed from European languages. Thus, we remain with 60 + 20 reduced Swadesh lists of 180 terms.

Given two words α_i and β_i corresponding to the same concept i in the language α (a Malagasy variety) and in the language β (a Bantu language), we define their Normalized Levenshtein Distance $D(\alpha_i, \beta_i)$ as

$$D(\alpha_i, \beta_i) = \frac{D_L(\alpha_i, \beta_i)}{L(\alpha_i, \beta_i)}, \tag{2}$$

where $D_L(\alpha_i, \beta_i)$ is the Levenshtein distance between the two words (the minimum number of operations: deletions, insertions and replacement, needed to transform one word in the other) and $L(\alpha_i, \beta_i)$ is the number of characters of the longer of the two (Serva and Petroni, 2008) (see also (Serva and Pasquini, 2020)).

This Normalized Levenshtein Distance can take any rational value between 0 (identical words) and 1 (totally different words). Words may partially change when borrowed (*beefsteak* becomes *bifteck* in French), so we have to decide when two words are the "same", *i.e.*, we have to decide a threshold for the Normalized Levenstein Distance, below which the two terms are the "same". In this paper we use the threshold $\sigma = 0.45$, nevertheless any σ between 0.3 and 0.5 leads to the same qualitative results. But if we go below 0.3 there is a risk to exclude positive matches while a threshold larger than 0.5 lead to false positives, *i.e.*, to identify as 'same' two words which indeed are similar only by chance.

If at least one among the 20 Bantu languages indexed by $\beta = 1, \ldots., 20$ has a $D_L(\alpha_i, \beta_i)$ equal or smaller than σ, we assume that the corresponding Malagasy word is borrowed. This procedure can be repeated for all the 180 concepts ($i = 1, \ldots \ldots 180$) so that the number $n(\alpha)$ of borrowed words in language α is determined by sum. In symbols:

$$n(\alpha_i) = 1 \quad \text{if} \quad \sup_{\beta=1,\ldots,20} D(\alpha_i, \beta_i) \leq \sigma \tag{3}$$

and $n(\alpha_i) = 0$ otherwise. Then

$$n(\alpha) = \sum_{1=1}^{180} n(\alpha_i). \tag{4}$$

Analogous quantities can be separately computed for each sub-group of Bantu languages. We consider apart the 4 Comoros Sabaki languages, the 4 Sabaki East-Africa languages and the 12 non-Sabaki East-Africa languages. Let us call $n_c(\alpha)$ (Comoros), $n_s(\alpha)$ (East Africa Sabaki) and $n_b(\alpha)$ (East Africa non-Sabaki) these numbers which compose three sets of 60 items. Each of the three numbers $n_c(\alpha)$, $n_s(\alpha)$ and $n_b(\alpha)$ is obviously smaller or equal to $n(\alpha)$ since the terms lent by a single group to each Malagasy variety α form a sub-ensemble of those lent from by the entire collection of the 20 Bantu languages. Moreover, the inequality $n(\alpha) \leq n_c(\alpha) + n_s(\alpha) + n_b(\alpha)$ is also satisfied for any α since a term in the variety α may find correspondence in more than a single group.

7. Conclusions

We hope that this paper has shed new light on a still controversial aspect of the history of peopling of Madagascar concerning the role of Africa and the Bantu-speaking peoples. Our results may not be conclusive, however we think that we have added new clues in favor of a direct settlement of Madagascar by Indonesian seafarers without a preliminary stopover in Africa. Moreover, we acknowledge that we have not exhaustively dealt the competing hypotheses, this would require a space that is more suited to the size of a book than an article.

We do not summarize here the pros and cons of the two competing hypotheses, but we would like to make a simple general consideration. The most striking contrast between Malagasy genetics and Malagasy language consists in proportions: the genetic makeup of Malagasy is for one half African and for one half Indonesian (yet with a small prevalence of African genes); on the contrary, their language contains a mere 2% of African terms while the overwhelming majority is Indonesian.

This difference cannot be easily explained by a genetic interbreeding in East-Africa with contemporary introgression of Bantu terms, since these proportions would be less dissimilar if it were true. Moreover, geneticists have shown that while the Y chromosome is African in a larger proportion, the situation is reversed when the mitochondrial DNA is concerned. This means that with regard to the Malagasy ancestors, there are more African males than Asian ones, and more Asian females than African ones. If a crew of Asian colonizers had landed in Africa and mixed with locals, the situation would have been the opposite.

The observed phenomenology is, on the contrary, well compatible with an initial Indonesian settlement followed by a continuous stream (or different waves) of African immigrants. The suceeding Bantu-speaking immigrants adopted the language of their new land, and they likely were preponderantly males.

A very enlightening example in recent times is Argentina. The language is Castilian with very few Italian words (many concerning food), while genealogy and genetics show that Italian and Spanish ancestors contributed to the nowadays Argentinian population in almost equal proportions (possibly with an Italian prevalence). Argentina was colonized by Spain, the Italians arrived later, initially in several waves and later by a continuous stream, exactly what likely happened in Madagascar with Indonesians and Africans.

In short, even this simple reasoning leads us to believe that the hypothesis of a direct colonization of Madagascar could be true, after all the *Novacula Occami* states that “Non sunt multiplicanda entia sine necessitate”.

Conflicts of interest

None.

Credit author statement

Conceptualization: Maurizio Serva. Data curation: Michele Pasquini. Formal analysis: Michele Pasquini. Investigation: Maurizio Serva. Methodology: Maurizio Serva. Software: Michele Pasquini. Supervision: Maurizio Serva. Validation: Michele Pasquini. Visualization: Michele Pasquini. Writing – Original Draft: Maurizio Serva. Writing – Review & Editing: Michele Pasquini.

Data availability

A linguistic dataset (Dataset S1.pdf) and an auxiliary dataset (Table S1.pdf) are included as supplementary material.

Acknowledgments

We are grateful to Heriniaina Andry Raboanary, Toky Hajatiana Raboanary, Julien Amédée Raboanary, Mara Edouard Remanevy, Dimby Vaovolo, Barthélemy Manjakahery and Marius Mandimbitafika Sambizafy for advice and suggestions concerning the Malagasy varieties and to the Professors Aurelio Simango, Narciso Gastene and Davety Mpiuka for advice and suggestions concerning the Mozambican languages used in the present study.

The research has benefited from the logistical support of the Institut Supérieur Polytechnique de Madagascar (ISPM) and of the assistance of its teachers and students.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.langsci.2022.101497.

References

Adelaar, A., 1989. Malay influence on Malagasy; linguistic and culture - historical implications. Ocean Ling. 28, 1–46.
Adelaar, A., 1994. Malay and javanese loanwords in Malagasy, Tagalog and Siraya (Formosa). Bijdragen tot Taal-, Land- Volkenkunde 150, 49–64.
Adelaar, A., 2006. The Indonesian migrations to Madagascar: making sense of the multidisciplinary evidence. In: Simanjuntak, T., Pojoh, I.H.E., Hisyam, M. (Eds.), Austronesian Diaspora and the Ethnogenesis of People in Indonesian Archipelago. Lipi Press, Jakarta, pp. 205–232.
Adelaar, A., 2007. Language contact in the Austronesian far west. In: 3rd Conference on Austronesian Languages and Linguistics. School of Asia and African Studies, London.
Adelaar, A., 2016. Austronesians in Madagascar: a critical assessment of the works of Paul Ottino and Philippe Beaujard (Chapter 4). In: Early Exchange between Africa and the Wider Indian Ocean World. Palgrave Macmillan and Gwyn Campbell Editors, pp. 77–112.
Adelaar, A., 2017. A linguist's perspective on the settlement history of Madagascar. NUSA 61, 69–88.
Adelaar, A., 2022. Language contact in Africa. To appear in the oxford guide to the malayo-polynesian languages of South east Asia. In: Adelaar, A., Schapper, A. (Eds.), Oxford Guides to the World's Languages. OUP, Oxford.
Allibert, C., 2015. *L'archipel des Comores et son histoire ancienne. Essai de mise en perspective des chroniques, de la tradition orale et des typologies de céramiques locales et d'importation.* Afriques **06**, L'Afrique orientale et l'océan Indien: connexions, réseaux d'échanges et globalisation.
Blust, R., 2019. The Austronesian homeland and dispersal. Ann. Rev. Ling. 5, 417–434.
Beaujard, Ph, 2011. The first migrants to Madagascar and their introduction of plants: linguistic and ethnological evidence. Azania: The Journal of the British Institute of History and Archaeology in East Africa 46, 169–189.
Beaujard, Ph, 2012. *Les mondes de l'Océan Indien*, Tome 1. De la formation de l'état au premier système-monde afro-eurasien. Armand Colin, Paris.
Brucato, N., Kusuma, P., Cox, M.P., Pierron, D., Purnomo, G.A., Adelaar, A., Kivisild, T., Letellier, T., Sudoyo, H., Ricaut, F.-X., 2016. Malagasy genetic ancestry comes from an historical Malay trading post in SouthEast Borneo. Mol. Biol. Evol. 33, 2396–2400.
Brucato, N., Fernandes, V., Kusuma, P., Černý, V., Mulligan, C.J., Soares, P., Rito, T., Besse, C., Boland, A., Deleuze, J.-F., Cox, M.P., Sudoyo, H., Stoneking, M., Pereira, L., Ricaut, F.-X., 2019. Evidence of Austronesian genetic lineages in East Africa and South Arabia: complex dispersal from Madagascar and Southeast Asia. Genome Bio. Evolut. 11, 748–758.
Cox, M.P., Nelson, M.G., Tumonggor, M.K., Ricaut, F.-X., Sudoyo, H., 2012. A small cohort of Island SouthEast Asian women founded Madagascar. Proceedings of the Royal Society B 279, 2761–2768.
Dahl, O.C., 1951. Malgache et Maanjan: une comparaison linguistique. In: Olso: Egede Instituttet (Arne Gimnes Forlag).
Dahl, O.C., 1954. Le substrat bantou en malgache. Norsk Tidsskrift for Sprogvidenskap 17, 325–362.
Dahl, O.C., 1977. La subdivision de la famille Barito et la place du malgache. Acta Orient. 38, 77–134.
Dahl, O.C., 1991. Migration from Kalimantan to Madagascar. Norwegian University Press. Institute for Comparative Research in Human Culture, Oslo, Norway.
de Houtman van Gouda, F., 1603. Spraeckende Woord-Boeck Inde Maleysche Ende Madagascarsche Talen Met Vele Arabische Ende Turcsche Woorden. Jan Evertsz, Amsterdam.
Deschamps, H., 1960. Histoire de Madagascar. Éditions Berger-Levrault, Paris.
Donohue, M., Denham, T., 2009. Banana (Musa spp.) domestication in the Asia-Pacific region: linguistic and archaeobotanical perspectives. Ethnobot. Res. Appl. 7, 293–332.
Faublée, J., 1970. Les manuscrits arabico-malgaches du Sud-Est, leur importance historique. Rev. Fr. Hist. Outre Mer. 57, 268–287.
Gueunier, N.J., 2004. Le dialecte malgache de Mayotte (Comores): une discussion dialectologique et sociolinguistique. Faits de Langues 23, 397–420.
Heiske, M., Alva, O., Pereda-Loth, V., Van Schalkwyk, M., Radimilahy, C., Letellier, T., Rakotarisoa, J.-A., Pierron, D., 2021. Genetic Evidence and Historical Theories of the Asian and African Origins of the Present Malagasy Population. Human Molecular Genetics. https://doi.org/10.1093/hmg/ddab018. (advance article).
Hurles, M.E., Sykes, B.C., Jobling, M.A., Forster, P., 2005. The dual origin of the Malagasy in island SouthEast Asia and East Africa: evidence from maternal and paternal lineages. Am. J. Hum. Genet. 76, 894–901.
Kovesi, P., 2015. Good Colour Maps: How to Design Them arXiv:1509.03700 [cs.GR].
Kusuma, P., Cox, M.P., Pierron, D., Razafindrazaka, H., Brucato, N., Tonasso, L., Suryadi, H.L., Letellier, T., Sudoyo, H., Ricaut, F.-X., 2015. Mitochondrial DNA and the Y chromosome suggest the settlement of Madagascar by Indonesian sea nomad populations. BMC Genom. 16, 191.
Kusuma, P., Brucato, N., Cox, M.P., Pierron, D., Razafindrazaka, H., Adelaar, A., Sudoyo, H., Letellier, T., Ricaut, F.-X., 2016. Contrasting linguistic and genetic origins of the Asian source populations of Malagasy. Nature. Scientific Rep. 6, 26066.
Mariano, L., 1613. Rélation du voyage de decouverte fait à l'île Saint-Laurent dans les années 1613-1614, par le capitaine Paulo Rodrigues da Costa et les pères jésuites Pedro Freire at Luis Mariano à bord de la caravelle Nossa Senhora da Esperança. In: Grandidier, Alfred, Grandidier, Guillome (Eds.), *Collection des ouvrages anciens concernant Madagascar*, Paris, Comité de Madagascar, vol. 2, pp. 1–79.
Pierron, D., Razafindrazaka, H., Pagani, L., Ricaut, F.-X., Antao, T., Capredon, M., Sambo, C., Radimilahy, C., Rakotoarisoa, J.-A., Blench, R.M., Letellier, T., Kivisild, T., 2014. Genome-wide evidence of Austronesian-Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar. Proc. Natl. Acad. Sci. USA 111, 936–941.
Pierron, D., Heiske, M., Razafindrazaka, H., Rakoto, I., Rabetokotany, N., Ravololomanga, B., Rakotozafy, L.M.-A., Mialy Rakotomalala, M., Razafiarivony, M., Rasoarifetra, B., Andriamampianina Raharijesy, M., Razafindralambo, L., Ramilisonina, Fanony, F., Lejamble, S., Thomas, O., Abdallah, A.M., Rocher, C., Arachiche, A., Tonaso, L., Pereda-Loth, V., Schiavinato, S., Brucato, N., Ricaut, F.-X., Kusuma, P., Sudoyo, H., Ni, S., Boland, A., Deleuze, J.-F., Beaujard, Ph, Grange, Ph, Adelaar, S., Stoneking, M., Rakotoarisoa, J.-A., Radimilahy, C., Letellier, T., 2017. Genomic landscape of human Diversity across Madagascar. Proc. Natl. Acad. Sci. USA 114, E6498–E6506.
Razafindrazaka, H., Ricaut, F.-X., Cox, M.P., Mormina, M., Dugoujon, J.-M., Randriamarolaza, L.P., Guitard, E., Tonasso, L., Ludes, B., Crubézy, E., 2010. Complete mitochondrial DNA sequences provide new insights into the Polynesian motif and the peopling of Madagascar. Eur. J. Hum. Genet. 18, 575–581.

Sapir, E., 1916. Time perspective in aboriginal American culture, a study in method. Geological Survey Memoir 90, Anthropological Series Government Printing Bureau 13, 1916. Ottawa, ON.
Serva, M., 2012b. The settlement of Madagascar: what dialects and languages can tell us. PLoS One 7 (2), e30666.
Serva, M., Petroni, F., 2008. Indo-European languages tree by Levenshtein distance. EPL 81, 68005.
Serva, M., Pasquini, M., 2020. Dialects of Madagascar. PLoS One 5 (10), e0240170.
Serva, M., Pasquini, M., 2021a. Malagasy dialects in Mayotte. EPL 133, 68003.
Serva, M., Pasquini, M., 2021b. The Sabaki languages of Comoros. Indian Ocean Rev. Sci. Technol. 1.
Serva, M., Petroni, F., Volchenkov, D., Wichmann, S., 2012a. Malagasy dialects and the peopling of Madagascar. J. R. Soc. Interface 9, 54–67.
Soodyall, H., Jenkins, T., Stoneking, M., 1995. Polynesian mtDNA in the Malagasy. Nat. Genet. 10, 377–378.
Tibbetts, G., 1979. A Study of the Arabic Texts Containing Material on South-East Asia. E. J. Brill, Leiden, The Netherlands.
Tofanelli, S., Bertoncini, S., Castrí, L., Luiselli, D., Calafell, F., Donati, G., Paoli, G., 2009. On the origins and admixture of Malagasy: new evidence from high-resolution analyses of paternal and maternal lineages. Mol. Biol. Evol. 26, 2109–2124.
Vavilov, N.I., 1926. Centers of origin of cultivated plants. Trudi po Prikl. Bot. Genet. Selek. (Bulletin of Applied Botany and Genetics) 16, 139–248.
Viswanathan, G.M., da Luz, M.G.E., Raposo, E.P., Stanley, H.E., 2011. The Physics of Foraging: An Introduction to Random Searches and Biological Encounters. Cambridge University Press.