European Journal of Human Genetics (2017) 25, 246–252

www.nature.com/ejhg

ARTICLE

Lack of gene–language correlation due to reciprocal female but directional male admixture in Austronesians and non-Austronesians of East Timor

Sibylle M Gomes[1,6], Mannis van Oven[2,6], Luis Souto[1], Helena Morreira[1], Silke Brauer[2], Martin Bodner[3], Bettina Zimmermann[3], Gabriela Huber[3], Christina Strobl[3], Alexander W Röck[3], Francisco Côrte-Real[4], Walther Parson[3,5] and Manfred Kayser*[,2]

Nusa Tenggara, including East Timor, located at the crossroad between Island Southeast Asia, Near Oceania, and Australia, are characterized by a complex cultural structure harbouring speakers from two different major linguistic groups of different geographic origins (Austronesian (AN) and non-Austronesian (NAN)). This provides suitable possibilities to study gene–language relationship; however, previous studies from other parts of Nusa Tenggara reported conflicting evidence about gene–language correlation in this region. Aiming to investigate gene–language relationships including sex-mediated aspects in East Timor, we analysed the paternally inherited non-recombining part of the Y chromosome (NRY) and the maternally inherited mitochondrial (mt) DNA in a representative collection of AN- and NAN-speaking groups. Y-SNP (single-nucleotide polymorphism) data were newly generated for 273 samples and combined with previously established Y-STR (short tandem repeat) data of the same samples, and with previously established mtDNA data of 290 different samples with, however, very similar representation of geographic and linguistic coverage of the country. We found NRY and mtDNA haplogroups of previously described putative East/Southeast Asian (E/SEA) and Near Oceanian (NO) origins in both AN and NAN speakers of East Timor, albeit in different proportions, suggesting reciprocal genetic admixture between both linguistic groups for females, but directional admixture for males. Our data underline the dual genetic origin of East Timorese in E/SEA and NO, and highlight that substantial genetic admixture between the two major linguistic groups had occurred, more so via women than men. Our study therefore provides another example where languages and genes do not conform due to sex-biased genetic admixture across major linguistic groups.
European Journal of Human Genetics (2017) **25,** 246–252; doi:10.1038/ejhg.2016.101; published online 3 August 2016

INTRODUCTION

The island of Timor is located in a subregion of Island Southeast Asia sometimes referred to as Wallacea. In 1859, Timor was divided by European colonizers into a western half (except the district of Oecusse), that is, Dutch Timor, and an eastern half (together with Oecusse), that is, Portuguese Timor. After decolonization from the Netherlands in 1949, West Timor became an Indonesian territory together with other parts of the former Dutch East Indies colony (except West Papua), whereas East Timor remained a Portuguese colony. After abandonment from Portugal in 1975, East Timor was occupied by Indonesia, but in 2002 obtained full independence to become the Democratic Republic of Timor-Leste.

The Lesser Sunda Island, or Nusa Tenggara, including East Timor, have an extraordinary high linguistic diversity with speakers of two very different major linguistic groups, that is, Austronesian (AN) and non-Austronesian (NAN) languages.[1] AN languages are assumed to have originated from Taiwan about 6–5 thousand years ago (kya) and distributed via the Austronesian expansion through Island Southeast Asia before they arrived in Near Oceania, that is, the Bismarck Archipelago, about 3.5 kya and further spread into Remote Oceania.[2–4] NAN languages are assumed to have originated from Near Oceania, most likely New Guinea.[1] Hence, it is widely assumed that AN speakers are of East/Southeast Asian (E/SEA) origin, while NAN speakers are of Near Oceanian (NO) origin.

In East Timor, >20 AN languages or dialects and four NAN languages are spoken. NAN languages are spoken by one third (~370 000 speakers) of the population[5] while the remainder speaks AN languages. Within East Timor, there are two non-adjacent regions with NAN speakers: Bunak is spoken in the border area with West Timor, whereas Makasae, Makalero and Fataluku are spoken in the eastern part of the country (Figure 1). AN languages are spoken in the central part of East Timor, its island Atauro, as well as in the district of Oecusse (an exclave within West Timor).[6]

The presence of linguistically very different AN- and NAN-speaking groups of very different geographic origins makes Nusa Tenggara attractive for scientific research, particularly for studying genetic and linguistic structure and the co-inheritance of cultural and genetic traits. As gene flow between two linguistically different groups is often sex-biased (ie, occurring predominantly via males rather than via females, or *vice versa*), analysing paternally inherited genetic diversity

[1]Department of Biology, University of Aveiro, Campus de Santiago, Aveiro, Portugal; [2]Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands; [3]Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria; [4]Faculty of Medicine, University of Coimbra, Coimbra, Portugal; [5]Forensic Science Program, The Pennsylvania State University, University Park, PA, USA
[6]These authors contributed equally to this work.
*Correspondence: Dr Manfred Kayser, Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, PO Box 2040, 3000 CA Rotterdam, The Netherlands. Tel: +31 10 7038073; Fax: +31 10 7044575; E-mail: m.kayser@erasmusmc.nl
Received 11 June 2015; revised 13 June 2016; accepted 21 June 2016; published online 3 August 2016

separately from maternally inherited diversity, that is, NRY (non-recombining part of the Y chromosome) and mitochondrial (mt) DNA, is vital for such studies. A previous mtDNA and NRY study of various islands across Nusa Tenggara demonstrated absence of correlation between genes and languages due to genetic admixture between AN- and NAN-speaking groups.[7,8] However, an NRY study on the island of Sumba did detect shared evolution of languages and genes.[9] From these contrasting findings it was speculated that the geographic scale (a larger island region *vs* a single island) may play a role in gene–language evolution.[7]

Timor is highly suitable for studying gene–language relationship on a restricted geographic scale because it harbours various AN and NAN groups within the island. Recently, Tumonggor *et al.*[10] found no correlation between genes and languages in West Timor. Here, we investigate gene–language relationship in East Timor via NRY and mtDNA analyses by combining previously reported mtDNA data with new analyses of NRY-haplogroup diversity. Gomes *et al.*[11] recently analysed mtDNA diversity in East Timor for the purpose of unveiling human settlement history. Two earlier studies catalogued NRY short tandem repeat (Y-STR) diversity in East Timor using a different set of samples.[12,13] For the latter sample set, we report here newly generated data for a battery of NRY single-nucleotide polymorphisms (Y-SNPs), which we analysed because Y-SNPs allow tracing events in the distant past, whereas Y-STRs are more suitable for recent past events due to their ~100 000 times higher mutation rates. From these two sets of East Timorese samples, we carefully selected samples with overlapping geographic and linguistic coverage of the country, and used their NRY and mtDNA data to investigate gene–language correlation and genetic–geographic population substructure.

MATERIALS AND METHODS

Samples

Two different sets of East Timorese samples were used in this study, one with previously reported mtDNA data,[11] and another with previously described Y-STRs data.[12,13] For the latter sample set, Y-SNP data were newly generated here. Examination of both sample sets revealed a strong overlap of East Timorese groups, their languages and geographic locations (Supplementary Material 1). Hence, these two sets of individuals are similarly representative of the East Timorese population regarding geographic distribution as well as language affiliations, so that comparing the NRY data from one set with the mtDNA data from the other set was justified. Language classification into AN or NAN speakers was based on Hull.[6] Samples with a maternal line (mtDNA dataset) and paternal line (NRY dataset) within East Timor, and without AN to NAN (or *vice versa*) language shifts within three generations, making a language classification in AN or NAN languages possible, were included in this study. Of the 324 East Timor samples described in the previous mtDNA study,[11] 290 fulfilled our sample selection criteria (Supplementary Material 2) and were used for mtDNA data analysis in the present study. Of the East Timor samples described in the previous Y-STR studies,[12,13] 273 fulfilled our sample selection criteria (Supplementary Material 3) and were used for *de novo* Y-SNP genotyping and included here for NRY data analysis together with the previously described Y-STR data[12,13] of the same samples. Samples were grouped according to the two major linguistic groups (AN *vs* NAN languages) and geography (districts) (Supplementary Material 1). This study was approved by the Universidade Nacional de Timor Lorosa'e (UNTL) and supported by the Portuguese Embassy at Dili, Timor-Leste. Y-SNP genotyping and sample/data storage at Erasmus MC were carried out under approval of the local Medical Ethics Committee (METC).

Genetic data generation

Generation of mtDNA and Y-STR data was described in earlier publications.[11–13] In brief, mtDNA control-region sequences (nps 16024-16569, 1-576) and Y-STR data from the following 12 markers, DYS19, DYS385a, DYS385b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438 and DYS439, were used here. For the present study, we newly analysed 273 samples (for which Y-STR data were established previously[12,13]) for 35 NRY binary markers (for marker description see Karafet *et al.*[14]) in a hierarchical fashion according to the phylogenetic relationships of the Y-SNPs. First, the Y-SNPs M4, P34, M38, M119, M122, M134, M208, M214, M217, M226, M230, M254 and M353 were genotyped in singleplex by standard PCR followed by restriction-fragment length polymorphism analysis, as previously described.[15] To further increase the phylogenetic resolution of the results thus obtained, relevant subsets of the samples were additionally analysed by employing a set of published multiplex assays[16] based on the single-base primer extension (*SNaPshot*) principle, targeting the following Y-SNPs: M9, P132, P256, M214, M74, M173, P202, M254 and M226 (multiplex A); M9, P79, M353, M177 and P117 (multiplex B); M175, M119, M110, M268, M95, M88, M122, M324, M7 and M134 (multiplex C); M130/RPS4Y, M38, M208, P33, P54, M217, M347 and M210 (multiplex D). Further information about

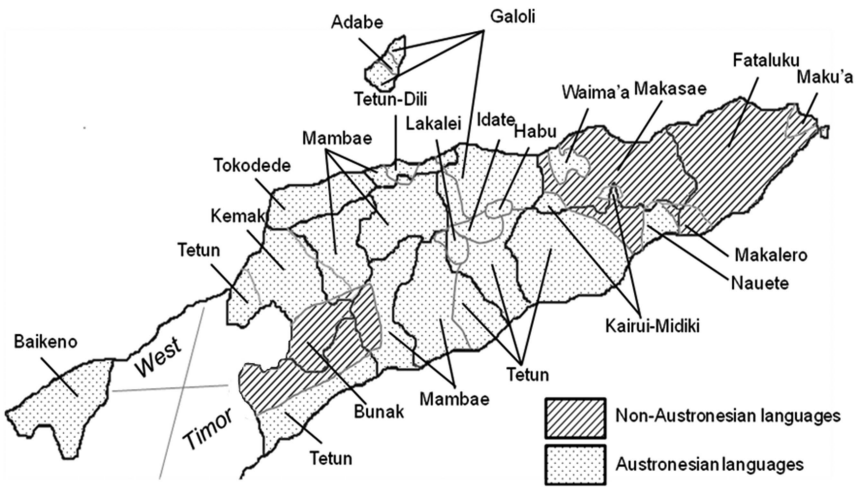
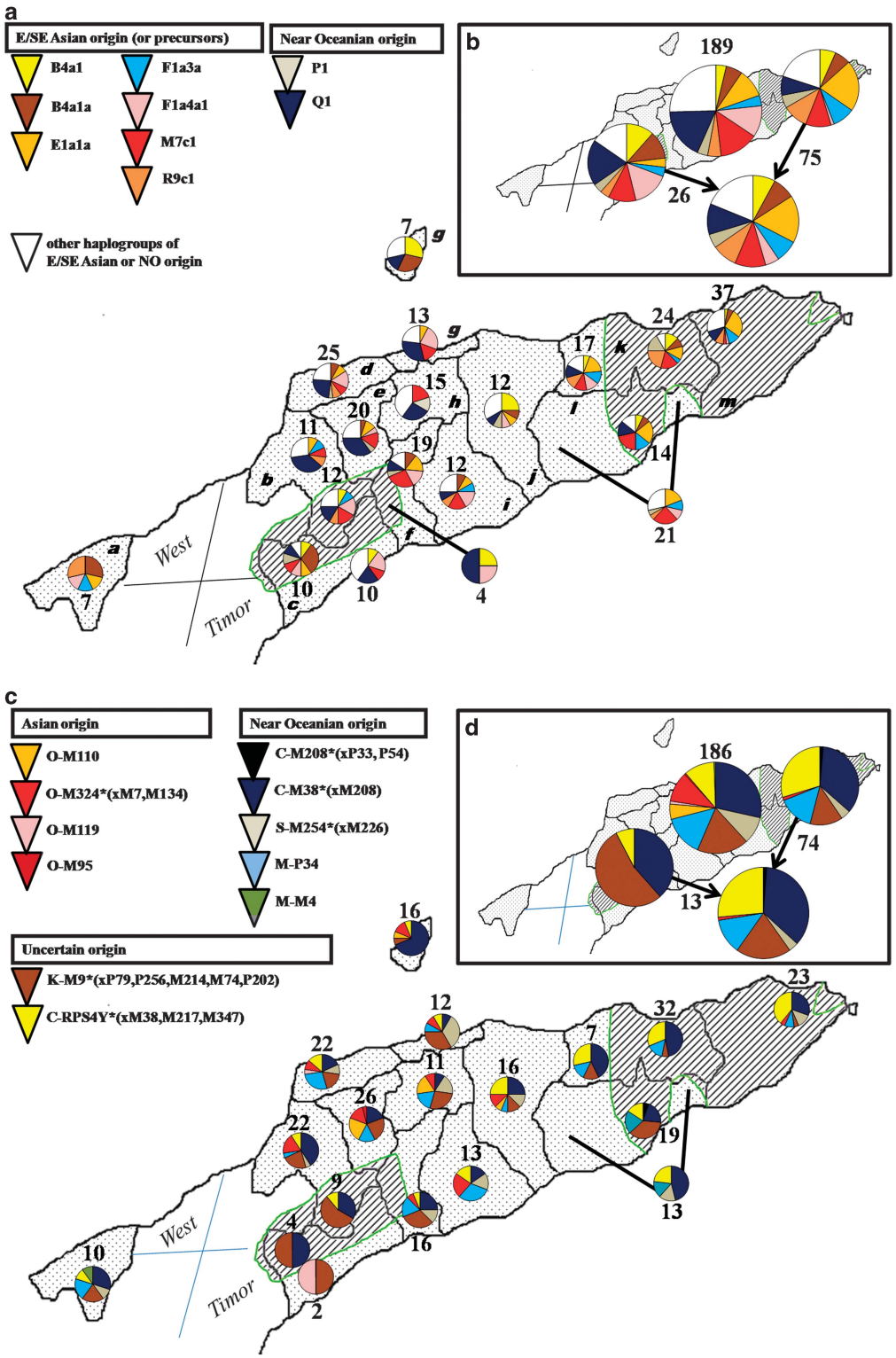


Figure 1 Map of East Timor. Languages are indicated, black lines represent district borders. Grey lines show regions according to language if these regions differ from district borders (modified after Durand,[34] Wurm and Hattori[35]). A simplified distribution of Austronesian (AN, dotted area) and non-Austronesian (NAN, striped area) languages is shown (map modified after Fox[29] and McWilliam[28]). Language classification was according to Hull,[6] Adabe was classified as Austronesian (Hull[36] and Antoinette Schapper, personal communication).

the Y-SNP and Y-STR markers is available in Supplementary Material 4. NRY haplogroups were designated according to PhyloTree's minimal Y tree (version 9-Apr-2014).[17] Assignment of NRY and mtDNA haplogroups to their putative regional geographic origins in E/SEA or NO as established before was done according to previous publications.[7,11,15,18–21] The Y-chromosome genotypes for all samples analysed, including Y-SNP and Y-STR data, are available in Supplementary Material 5. The mtDNA sequence data are available from GenBank (KJ655583-KJ655889, KJ676774-KJ676790) as described previously.[11]

Statistical analyses

Diversity indices as well as F_{ST} (for Y-SNP haplogroups and mtDNA haplogroups) and R_{ST}/Φ_{ST} (for Y-STR haplotypes and mtDNA sequences) values were computed and an analysis of molecular variance (AMOVA) was performed with groupings according to language and geography (see Supplementary Material 1), all using Arlequin v.3.5.1.2.[22] Multidimensional scaling (MDS) analyses of pairwise R_{ST}, Φ_{ST} and F_{ST} distances were performed with the software SPSS Statistics (IBM, Armonk, NY, USA) using the PROXSCAL



algorithm;[23] stress values were evaluated according to Sturrock and Rocha[24] and all found to be statistically significant. The multicopy Y-STR marker DYS385 was disregarded in the R_{ST}-based analysis. As the Y-STR locus DYS389II includes DYS389I, we subtracted the number of repeats observed at DYS389I from the number of repeats observed at DYS389II in order not to count the DYS389I values twice.

RESULTS

MtDNA variation according to linguistic and geographic affinities

Thirty-two different haplogroups among 189 AN speakers and 24 among 101 NAN speakers were detected (Supplementary Material 6; Figure 2a and b), of which four (Q1, M7c1, F1a4a1 and E1a1a) were frequent with ≥5% in speakers of both major linguistic groups. MtDNA haplogroups of previously described putative E/SEA origin and those of putative NO origin were found in both AN and NAN speakers (Supplementary Material 6; Figure 2a and b). Surprising under the gene–language hypothesis is that maternal lineages of previously described putative NO origin (R14, Q1 and P1, haplogroup origin assignment in Supplementary Material 6) were more frequent in AN (23%) than in NAN speakers (16%) (Table 1). This difference was mostly due to the NO haplogroup Q1 (AN 17%, NAN 11%) (Supplementary Material 6; Figure 2b), and agrees with findings from West Timor.[10] Maternal lineages of previously described putative E/SEA origin (B4, B5, D5, D6, F, M7, M10, M21, M73, N21 and R9, haplogroup origin assignment in Supplementary Material 6) accounted for 53% of the AN speakers, as expected, but unexpectedly under the gene–language hypothesis also for 59% of the NAN speakers (Table 1; Supplementary Material 6; Figures 2a and b). MtDNA haplogroups that were exclusively found in AN speakers (B4*, B5b1c, D6a, E2, F1a1a, F1a2, F3b1a, M*, M73a, Q3, R*, R14) or in NAN speakers (B4c2, M10, M21b and R9c1b2) occurred at a combined frequency of <12%, respectively (Supplementary Material 6; Figures 2a and b).

The distribution of East Timorese NAN speakers is geographically divided into the western part (Bunak) and the eastern part of the country (Fataluku, Makasae, Makalero). Comparisons of west NAN, east NAN and AN groups revealed a broad lineage spectrum for all of these groups. However, when the west NAN and east NAN groups were lumped together, respectively, they showed a largely similar mtDNA haplogroup composition compared with all AN groups combined (Figure 2b), with the exceptions of E1a1a (putative E/SEA origin) and Q1 (putative NO origin) (Figure 2b). However, these two haplogroups rather differed on the geographic, not necessarily linguistic, dimension and showed opposite distributions across East Timor with E1a1a more frequently found in eastern groups (27.0–12.5%, AN and NAN speakers) than in western groups (10.0–0%, AN and NAN speakers), and Q1 more frequently seen in the west (36.4–10.0%, AN and NAN speakers without Oecusse) than in the east (14.3–0%, AN and NAN speakers) of the country (Figure 2a; Supplementary Material 7).

An AMOVA based on Φ_{ST} from mtDNA control-region sequences (Table 2) provided no support for population groupings according to linguistics (AN *vs* NAN) (0.4% of total variation explained among groups, not significant) and also not according to linguistics and geography considered together (west NAN *vs* AN *vs* east NAN) (0.7%, not significant), whereas a small support (1.34%, $P<0.05$) was achieved for a grouping according to geography only (West *vs* Middle *vs* East). Very similar AMOVA results were obtained based on F_{ST} from mtDNA haplogroups (Table 2). However, no clear pattern

Table 1 Origin of mtDNA and NRY haplogroups in Austronesians and non-Austronesians from East Timor

Genetic system	*Linguistic affiliation*	n	*E/SE Asian origin (%)*	n	*Near Oceanian origin (%)*	n	*Unknown origin (%)*	*Overall*
mtDNA	Non-Austronesian	79	78.2 (59.4[a])	22	21.8 (15.8[b])	0	0	101
	Austronesian	134	70.9 (52.9[a])	52	27.5 (23.3[b])	3	1.6	189
	Total	213	73.4 (55.2[a])	74	25.5 (20.7[b])	3	1.0	290
NRY	Non-Austronesian	1	1.1	46	52.9	40	46.0	87
	Austronesian	33	17.7	99	53.2	54	29.0	186
	Total	34	12.5	145	53.1	94	34.4	273

Abbreviations: E/SE, East/Southeast; mtDNA, mitochondrial DNA; NRY, non-recombining part of the Y chromosome.
[a]Without E* when classified as East Indonesian origin with East Asian precursor, for haplogroup assignment according to origin and references see Supplementary Material 6 (mtDNA haplogroups) and Supplementary Material 9 (NRY haplogroups).
[b]Without B4a1a1 when classified as East Indonesian/Melanesian/Near Oceanian origin with East Asian precursor, for haplogroup assignment according to origin and references see Supplementary Material 6 (mtDNA haplogroups) and Supplementary Material 9 (NRY haplogroups).

Figure 2 Distribution of mtDNA and NRY haplogroups across East Timor. (**a**) Map as in Figure 1 with population samples analysed for mtDNA and haplogroup results represented as pie charts placed at their respective district locations. District names according to letter code: *a*—Oecusse, *b*—Bobonaro, *c*—Cova-Lima, *d*—Liquiçá, *e*—Ermera, *f*—Ainaro, *g*—Dili and Dili-Atauro, *h*—Aileu, *i*—Manufahi, *j*—Manatuto, *k*—Baucau, *l*—Viqueque, *m*—Lautém. On the eastern tip of Timor the districts of Baucau and Viqueque show regions of Austronesian (AN) and non-Austronesian (NAN) speakers. The AN languages Waima'a, Kairui and Midiki were clustered to simplify the graphic (for comparison see Figure 1). Lautém and its offshore island Jaco only include NAN languages. The green line in district Lautém indicates a former AN language region (language: Makuva or Maku'a) which belongs today to NAN speakers of Fataluku. Only mtDNA haplogroups observed with >4% are shown as initially described by Gomes *et al*.[11] For practical reasons, B4a1a1 and B4a1a3a were joined to B4a1a*, P1, P1d and P1e to P1* and R9c1, R9c1a and R9c1b2 to R9c1* to reduce the number of haplogroups. Haplogroups according to PhyloTree, Build 16.[37] Numbers indicate per-group sample size. Putative regional geographic origins of the haplogroups as previously described are indicated. (**b**) mtDNA haplogroup data for population samples are lumped together according to the two major linguistic groups AN and NAN shown as pie charts, with east NAN and west NAN shown separately in addition. Numbers indicate sample size. (**c**) Map as in Figure 1 with population samples analysed for NRY DNA and haplogroup results represented as pie charts placed at their respective district locations. For district names see Figures 2a. Y-SNP-based haplogroups according to PhyloTree's minimal Y tree (version 9-Apr-2014).[17] Numbers indicate per-group sample size. Putative regional geographic origins of the haplogroups as previously described are indicated. (**d**) NRY haplogroup data for population samples are lumped together according to AN and NAN shown as pie charts, with east NAN and west NAN shown separately in addition. Numbers indicate sample size.

Table 2 AMOVA results from mtDNA analysis with groupings according to linguistics and geography

	Groups according to linguistics[a]			*Groups according to linguistics and geography*[b]			*Groups according to geography*[c]		
	Var components	*% of variation*	*Fixation indices*	*Var components*	*% of variation*	*Fixation indices*	*Var components*	*% of variation*	*Fixation indices*
Based on F_{ST} distances from mtDNA haplogroups									
Among groups	0.00195 Va	0.42	0.00421	0.00401 Va	0.86	0.00865	0.00753 Va	1.62	0.01621*
Among districts within groups	0.00612 Vb	1.32	0.01326*	0.00492 Vb	1.06	0.01069	0.00174 Vb	0.37	0.00381*
Within East Timor population	0.45535 Vc	98.26	0.01742*	0.45535 Vc	98.08	0.01924*	0.45535 Vc	98.00	0.01996*
Based on Φ_{ST} from mtDNA CR haplotypes									
Among groups	0.03043 Va	0.4	0.004	0.05157 Va	0.68	0.00682	0.10156 Va	1.34	0.01342*
Among districts within groups	0.08969 Vb	1.19	0.0119	0.07686 Vb	1.02	0.01024	0.03263 Vb	0.43	0.00437
Within East Timor population	7.43114 Vc	98.41	0.0159*	7.43114 Vc	98.3	0.01699*	7.43114 Vc	98.23	0.01774*

Abbreviations: CR, control region; mtDNA, mitochondrial DNA.
*P-value <0.05.
[a]Groups: Non-Austronesian (all) *vs* Austronesian (all).
[b]Groups: West non-Austronesian (NAN-Ainaro, NAN-Bobonaro, NAN-Cova-Lima) *vs* east non-Austronesian (NAN-Baucau, NAN-Viqueque, Lautém) *vs* Austronesian (all).
[c]Groups: West (Oecusse, Liquiçá, Ermera, Cova Lima_AN, Bobonaro_AN, Cova Lima_NAN, Bobonaro_NAN) *vs* Middle (Ainaro_AN, Aileu, Dili, Manufahi, Manatutu, Dili-Atauro, Ainaro_NAN) *vs* East (Baucau_AN, Viqueque_AN, Baucau_NAN, Viqueque_NAN, Lautém).

according to linguistics and geography was seen in the MDS plot (Supplementary Material 8a and b); the main cluster observed indicates limited genetic differentiation and high gene flow between AN and NAN groups throughout East Timor. Overall, these data provide no mtDNA evidence for gene–language relationship and little evidence for genetic–geographic population sub-substructure in East Timor.

NRY variation according to linguistic and geographic affinities

On the basis of the data of 35 binary NRY markers, the samples were classified into 11 haplo-/paragroups (Supplementary Material 9). The NRY-haplogroup frequencies in AN and NAN speakers are presented in Supplementary Material 9 and Figure 2c and d. For the NRY haplogroups C-M208*, C-M38*, S-M254*, K-M9*, M-P34 and M-M4* a putative NO origin was previously described, and for O-M110, O-M119*, O-M324* and O-M95* a putative E/SEA origin (see Materials and methods). For the unresolved lineages assigned to the higher-level paragroups K-M9* and C-RPS4Y*, it is currently impossible to provide a definitive origin in either E/SEA or NO, because both K-M9 and C-RPS4Y contain subclades of putative E/SEA origin (eg, O-M175 and C-M217, respectively) as well as others of putative NO origin (eg, M-P256 and C-M38).

In East Timorese AN and NAN speakers, the proportion of paternal lineages with putative NO origin was highly similar with 53.2 and 52.9%, respectively (Table 1; Figure 2c), which appears surprising for the AN speakers under the gene–language hypothesis. In contrast, the proportion of putative E/SEA paternal lineages among AN and NAN speakers differed and was significantly higher in AN (17.7%) than in NAN speakers where these lineages were absent except for a single individual (1.1%). Although the very low frequency of E/SEA NRY lineages in NAN speakers is expected under a gene–language hypothesis, the fact that the AN speakers carried merely 17.7% E/SEA NRY lineages appeared surprising. Western NAN speakers, on average, had a higher frequency of K-M9* and a lower frequency of C-RPS4Y* (the two NRY haplogroups with unknown geographic origin) than eastern NAN speakers (Figure 2d). However, this difference appears to be driven by geography rather than linguistics as indicated by the similar frequencies in the respective AN-speaking geographic neighbours (Figure 2c). The NO origin haplogroup M-P34, which was seen with considerable frequency in eastern NAN (as well as in almost all AN groups), was absent from both western NAN groups as well as one neighbouring AN group, but sample size in all three groups was relatively low, limiting meaningful conclusions (Supplementary Material 10; Figure 2c and d).

AMOVA based on F_{ST} for NRY-haplogroup frequencies and R_{ST} for Y-STR haplotypes (Table 3) provided no support for a population grouping according to linguistics (AN *vs* NAN) (Y-SNP 1.26% and Y-STR 1.88%, both not significant) nor according to geography for Y-SNPs (1.84%, not significant) whereas for Y-STRs it was small (4%, $P<0.05$). A grouping based on linguistics and geography considered together (west NAN *vs* AN *vs* east NAN) received a small support (Y-SNP 2.82% $P<0.05$ and Y-STR 6.33% $P<0.05$), in contrast to the mtDNA findings with no support. This is in line with the MDS analysis (Supplementary Material 8c and d) where the western NAN groups and the eastern NAN groups cluster away from each other and from the AN groups (with the exception of Viqueque_NAN, Supplementary Material 8c, and of Lautem see Supplementary Material 8d), which was not seen for mtDNA. These statistical analyses provide no NRY evidence for gene–language relationship and limited (slightly higher than for mtDNA) evidence for genetic–geographic population sub-substructure in East Timor. However, we like to emphasize that in contrast to mtDNA, where the proportions of E/SEA and NO lineages in AN and NAN speakers were rather similar, respectively, for NRY this was only seen for NO lineages, whereas the E/SEA lineages that were relatively frequent in AN (17.7%) were absent from NAN with the exception of a single individual.

DISCUSSION

The island of East Timor is located within the contact zone of AN and NAN languages. According to linguists, AN languages have influenced and expanded into NAN language territories and *vice versa* within East Timor.[25–27] Evidence of a slow ongoing tendency to language admixture of AN and NAN languages caused by loan words, language shifts and similar Austronesian cultural behaviour (independent of their speakers' linguistic affiliations) have been noted.[28,29] The NRY and mtDNA data not only clearly show the dual genetic origin of East Timorese in E/SEA and NO, in line with linguistic evidence, they also demonstrate that considerable mixing between members of both major linguistic groups had occurred during the population history of East Timor.

Table 3 AMOVA results from NRY analysis with groupings according to linguistics and geography

	Groups according to linguistics[a]			*Groups according to linguistics and geography*[b]			*Groups according to geography*[c]		
	Var components	*% of variation*	*Fixation indices*	*Var components*	*% of variation*	*Fixation indices*	*Var components*	*% of variation*	*Fixation indices*
Based on F_{ST} distances from Y-SNP haplogroups									
Among groups	0.00520 Va	1.26	0.01264	0.01170 Va	2.82	0.02822*	0.00756 Va	1.84	0.01842
Among districts within groups	0.01701 Vb	4.14	0.04188**	0.01364 Vb	3.29	0.03385*	0.01403 Vb	3.42	0.0348*
Within East Timor population	0.38916 Vc	94.6	0.054**	0.38916 Vc	93.89	0.06112**	0.38916 Vc	94.74	0.05258**
Based on R_{ST} distances from Y-STR haplotypes									
Among groups	0.14421 Va	1.88	0.01884	0.49519 Va	6.33	0.06327*	0.30704 Va	4.01	0.04006*
Among districts within groups	0.54013 Vb	7.06	0.07192**	0.36156 Vb	4.62	0.04932**	0.38783 Vb	5.06	0.05271**
Within East Timor population	6.97005 Vc	91.06	0.0894**	6.97005 Vc	89.05	0.10946**	6.97005 Vc	90.93	0.09065**

Abbreviations: Y-SNPs, NRY single-nucleotide polymorphisms; Y-STR, NRY short tandem repeat.
*P-value <0.05.
**P-value <0.0001 (obtained after 1000 permutations).
[a]Groups: Non-Austronesian (all) *vs* Austronesian (all).
[b]Groups: West non-Austronesian (NAN-Bobonaro, NAN-Cova-Lima) *vs* east non-Austronesian (NAN-Baucau, NAN-Viqueque, Lautém) *vs* Austronesian (all).
[c]Groups: West (Oecusse, Liquiçá, Ermera, Bobonaro_AN, Cova Lima_NAN, Bobonaro_NAN) *vs* Middle (Ainaro_AN, Aileu, Dili, Manufahi, Manatutu, Dili-Atauro) *vs* East (Baucau_AN, Viqueque_AN, Baucau_NAN, Viqueque_NAN, Lautém).

However, our results suggest that genetic mixing between members of AN and NAN language groups was not equal for men and women as we see clear differences for the maternal and the paternal genetic ancestry of East Timorese, which sheds light on further details of the admixture history. For both AN and NAN speakers, we noted that maternal ancestry was clearly more E/SEA than NO, whereas the paternal ancestry was clearly more NO than E/SEA. Overall, these data suggest that the genetic admixture between the two major linguistic groups was mostly driven by AN-speaking women of E/SEA origin, and NAN-speaking men of NO origin. The same sex-biased admixture scenario had been concluded previously for Near Oceania,[21] including for the admixture that gave rise to the occupation of Remote Oceania.[20,21,30] Notably, for Near Oceania it is widely assumed that NAN-speaking Papuans arrived much earlier, that is, >35 kya, than Austronesians did about 3.5 kya. The sex-biased admixture of more Asian women and more Papuan males was likely the result of the subsequent arrival of the Austronesians into the residence of Papuans, together with the matrilocality and matrilinearity of the arriving early Austronesian societies and perhaps in combination with polygyny of the resident NAN-speaking Papuan societies.[18,31] That we observed the same in East Timor, as well as across East Indonesia previously[7,8] might indicate that NAN speakers arrived in Nusa Tenggara from New Guinea before AN speakers did from Asia (or at the same time), which is in line with views of some[6,25,26] but not all linguistic scholars.[28]

Moreover, our observation in East Timor that AN and NAN speakers have similar proportions of E/SEA and NO maternal lineages together with similar proportions of NO paternal lineages, but strongly different proportions of E/SEA paternal lineages (18% in AN and 1% in NAN), suggests reciprocal admixture for women between both linguistic groups, but directional admixture for men, with more NAN men mixing into AN groups but fewer AN men into NAN groups. This restricted admixture behaviour of men but not women implies a higher mobility of women than men in line with the assumed patrilocality of most East Timorese groups today.[32,33]

Notably, the strong imbalance in the E/SEA NRY lineages in AN and NAN speakers observed here for East Timor was not previously observed in a study carried out across many islands of the Nusa Tenggara region, where rather similar average proportions of E/SEA NRY lineages in AN and NAN speakers were found (AN 27%, NAN 25%).[7,8] Notably, the NRY-SNP resolution used in this study is nearly identical to that of Mona *et al.*[7] Future systematic studies within other Nusa Tenggara islands where AN and NAN speakers live, such as Alor and Pantar, will show if East Timor stands out with this finding or not. As our main conclusions were derived from the observed NRY and mtDNA haplogroups together with the previously described knowledge of their putative regional geographic origin, any potential errors in the previous geographic origin assignments of haplogroups will consequently impact on our study.

Overall, our study provides an example where language and genetic information, at least those inherited uniparentally, are not congruent because of sex-biased genetic admixture between major linguistic groups. This once again underlines the suitability of using mtDNA and NRY for investigating human population history as these data can detect sex-mediated events more clearly than bi-parentally inherited autosomal markers or X-chromosomal data including genome-wide diversity studies.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank all East Timorese who provided their samples and data to this study. This study was financially supported in part by FCT (Foundation of Science and Technology Portugal), COMPETE (*Programa Operacional Temático Factores de Competitividade)*, FEDER (European Community Fund) with the Project PTDC/CS-ANT/108558/2008, the FCT fellowship SFRH/BD/63165/2009. It received additional support by the Erasmus MC University Medical Center Rotterdam, the Austrian Science Fund (FWF) [L397 and P22880-B12], the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 285487, the intramural funding programme of the Medical University Innsbruck for young scientists MUI-START Project 2013042025, and the Austrian *Theodor Körner Fonds zur Förderung von Wissenschaft und Kunst*.

1 Pawley A: Prehistoric migration and colonisation processes in Oceania: a view from historical linguistics and archaeology; in: Lucassen J, Lucassen L, Manning P (eds): *Migration History in World History: Multidisciplinary Approaches*. Brill Academic Publishers: Leiden (The Netherlands), 2010, pp 77–112.

2 Spriggs M: Chronology of the Neolithic Transition in Island Southeast Asia and the 21 Western Pacific: a view from 2003. *Rev Archaeol* 2003; **24**: 57–80.
3 Gray RD, Drummond AJ, Greenhill SJ: Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 2009; **323**: 479–483.
4 Kayser M: The human genetic history of Oceania: near and remote views of dispersal. *Curr Biol* 2010; **20**: R194–R201.
5 Schapper A, Huber J, van Engelenhoven A: The historical relation of the Papuan languages of Timor and Kisar; in: Hammarström H, van der Heuvel W (eds): *History, Contact and Classification of Papuan Languages: Language and Linguistics in Melanesia*. Port Moresby: Linguistic Society of Papua New Guinea, 2012; 194–242. Special Issue 2012 Part I.
6 Hull G: The Papuan languages of East Timor. *Studies in Languages and Cultures of East Timor* 2004; **6**: 23–99.
7 Mona S, Grunz KE, Brauer S *et al*: Genetic admixture history of eastern Indonesia as revealed by Y-chromosome and mitochondrial DNA analysis. *Mol Biol Evol* 2009; **26**: 1865–1877.
8 Mona S, Grunz KE, Brauer S *et al*: Corrigendum: genetic admixture history of eastern Indonesia as revealed by Y-Chromosome and mitochondrial DNA analysis. *Mol Biol Evol* 2009; **26**: 1865–1877.
9 Lansing JS, Cox MP, Downey SS *et al*: Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc Natl Acad Sci USA* 2007; **104**: 16022–16026.
10 Tumonggor MK, Karafet TM, Downey S *et al*: Isolation, contact and social behavior shaped genetic diversity in West Timor. *J Hum Genet* 2014; **59**: 494–503.
11 Gomes MS, Bodner M, Souto L *et al*: Human settlement history between Sunda and Sahul: a focus on East Timor (Timor-Leste) and the Pleistocenic mtDNA diversity. *BMC Genomics* 2015; **16**: 70.
12 Souto L, Gusmão L, Amorim A, Côrte-Real F, Vieira DN: Y-STR haplotype diversity in distinct linguistic groups from East Timor. *Am J Hum Biol* 2006a; **18**: 691–701.
13 Souto L, Gusmão L, Ferreira E, Amorim A, Côrte-Real F, Vieira DN: Y-chromosome STR haplotypes in East Timor: forensic evaluation and population data. *Forensic Sci Int* 2006b; **156**: 261–265.
14 Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF: New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res* 2008; **18**: 830–838.
15 Kayser M, Brauer S, Cordaux R *et al*: Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol Biol Evol* 2006; **23**: 2234–2244.
16 van Oven M, Brauer S, Choi Y *et al*: Human genetics of the Kula Ring: Y-chromosome and mitochondrial DNA variation in the Massim of Papua New Guinea. *Eur J Hum Genet* 2014a; **22**: 1393–1403.
17 van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MH: Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. *Hum Mutat* 2014b; **35**: 187–191.
18 Kayser M, Brauer S, Weiss G *et al*: Melanesian origin of Polynesian Y chromosomes. *Curr Biol* 2000; **10**: 1237–1246.
19 Kayser M, Brauer S, Weiss G, Schiefenhovel W, Underhill PA, Stoneking M: Independent histories of human Y chromosomes from Melanesia and Australia. *Am J Hum Genet* 2001; **68**: 173–190.
20 Kayser M, Brauer S, Weiss G *et al*: Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *Am J Hum Genet* 2003; **72**: 281–302.
21 Kayser M, Choi Y, van Oven M *et al*: The impact of the Austronesian expansion: evidence from mtDNA and Y-chromosome diversity in the Admiralty Islands of Melanesia. *Mol Biol Evol* 2008; **25**: 1362–1374.
22 Excoffier L, Lischer HEL: Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Res* 2010; **10**: 564–567.
23 Busing FMTA, Commandeur JJF, Heiser WJ: PROXSCAL: a multidimensional scaling program for individual differences scaling with constraints; in: Bandilla W, Faulbaum F (eds): *SoftStat '97: Advances in Statistical Software 6*. Stuttgart: Lucius & Lucius, 1997, pp 67–74.
24 Sturrock K, Rocha J: A multidimensional scaling stress evaluation table. *Field Methods* 2000; **12**: 49–60.
25 Ross M: Pronouns as a preliminary diagnostic of grouping Papuan languages; in: Pawley A, Attenborough R, Golson J, Hide R (eds): *Papuan Pasts: Cultural, Linguistic and Biological Histories of Papuan-Speaking Peoples*. Canberra (Australia): Pacific Linguistics, 2005, pp 15–65.
26 Wurm SA: Linguistic prehistory in the New Guinea area. *J Hum Evol* 1983; **12**: 25–35.
27 Schapper A: Finding Bunaq: the homeland and expansion of the Bunaq in central Timor; in: McWilliam A, Traube EG (eds): *Life and Land in Timor: Ethnographic Papers*. Canberra: ANU E Press, 2011, pp 163–186.
28 McWilliam A: Austronesians in Linguistic Disguise: Fataluku cultural fusion in East Timor. *J Southeast Asian Stud* 2007; **38**: 355–375.
29 Fox JJ: Tracing the path, recounting the past: historical perspectives on Timor; in: Fox JJ, Soares DB (eds): *Out of the Ashes: Destruction and Reconstruction of East Timor*. Adelaide: Crawford House Publishing, 2000, pp 1–29.
30 Wollstein A, Lao O, Becker C *et al*: Demographic history of Oceania inferred from genome-wide data. *Curr Biol* 2010; **20**: 1983–1992.
31 Hage P, Marck J: Matrilineality and the Melanesian origin of Polynesian Y chromosomes. *Curr Anthropol* 2003; **44**: 121–127.
32 Narciso V, Henriques P: 'Women and land in Timor-Leste: issues in gender and development'. *Indian J Gend Stud* 2010; **17**: 59.
33 van Wouden FAE: *Types of Social Structure in Eastern Indonesia*. The Hague: Martinus Nijhoff, 1968.
34 Durand F: *Timor-Leste - País no Cruzamento da Ásia e do Pacífico Um Atlas Histórico-Geográfico*. Lisbon: LIDEL, 2010; pp 46–47.
35 Wurm SA, Hattori S: *Language atlas of the Pacific area*. Canberra: The Australian Academy of the Human-Japan Academy, 1981; p 47.
36 Hull G: The languages of Timor 1772-1997: a literature review. *Studies in Language and Cultures of East Timor* 1998; **1**: 1–38.
37 van Oven M, Kayser M: Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* 2009; **30**: E386–E394.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)