Supporting Information

Reyes-Centeno et al. 10.1073/pnas.1323666111

The "Negrito" Hypothesis

The ethnographic term "Negrito" broadly refers to Southeast Asian populations exhibiting a phenotype of short stature, dark skin color, and tufted hair and implies a common origin hypothesis (1). Alongside Australians, Papuans, Melanesians, and Dravidian-speaking Indian populations, the "Negrito" have been hypothesized to be isolated, "relic" descendants of a first dispersal out of Africa and into Asia (2). Following a biogeographical approach, the designation of "relic" is in reference to the ecological context of populations that have become isolated as a result of occupying geographical refugia or exploiting specific ecological niches. To date, the most comprehensive genetic study exploring diversity of modern human populations in Asia sampled seven "Negrito" populations, including the Agta, Aeta, and Iraya from the northern Philippines; the Mamanwa and Ati from the southern Philippines; and the Jehai and Kensiu from Malaysia (3). This study found that "Negrito" population affinities are with geographically proximate populations rather than with other "Negrito" groups. The study therefore challenges a simple common origin hypothesis for the "Negrito" and implies other evolutionary mechanisms accounting for their common phenotype. Nonetheless, the Mamanwa's ancient association with Australians and highland Melanesians has been interpreted as evidence for an early, southern route dispersal into Southeast Asia (4). Likewise, the Aeta remain candidate descendants of a first dispersal within the multiple dispersals with isolation (MDI) model, alongside Melanesians and Papuans (5).

To assess affinities of our Aeta/Agta sample, as well as our Papuan and Melanesian samples, we conducted a principal component analysis (PCA) using the SNPRelate R package (6) and a discriminant analysis of principal components (DAPC) using the adegenet R package (7). We used the same data and groupings as in the main text (3, 4, 8–13) (Table 1 and Table S3). DAPC is a multivariate method, free of assumptions about Hardy–Weinberg equilibrium or linkage disequilibrium. It has been shown to generally perform better than the STRUCTURE method (14) and is also analogous to an ADMIXTURE method (15) in that a number, K, of clusters can be specified to assess population structure. We identified the best supported grouping of individuals running a K-means clustering of principal components (16) and used a Bayesian Information Criterion (BIC) approach to assess the best supported number of clusters. For the genetic dataset, we found $K = 5$ to be the best supported model (Fig. S2*A*) and therefore used this in the DAPC. Although results were less clear for the cranial phenotype dataset, with BIC results approximately equivalent for $K = 5$–8 clusters (Fig. S2*B*), we also used $K = 5$ as the best-supported model. For the genetic dataset, the DAPC along the first two axes revealed three major clusters within the five supported by the $K = 5$ model (Fig. S3*A*). They included (*i*) AU-NG-ME, (*ii*) ((JP-NE)-CA-(NI-SI)), and (*iii*) EA-SA. This clustering pattern is also observed along the first two PCs in a standard PCA (Fig. S3*B*). The Aeta/Agta were not classified into the AU-NG-ME cluster (Table S5), as might be expected if they shared an ancient association with those populations in a similar fashion as the Mamanwa. Instead, the Aeta/Agta classified primarily with the Japanese and Central Asians. As foreseen by the BIC results of the cranial phenotype data, classification was much more mixed in this case, with individuals classified across less coherent clusters (Table S5). Nevertheless, in the clusters where the Aeta/Agta were classified the most, Japanese and Central Asians were also strongly represented.

To more robustly assess the association of our Aeta/Agta sample, we conducted a TreeMix analysis (17) on the genetic data. The TreeMix method relaxes the assumptions of branching models of biological evolution, incorporating the possibility that populations did not remain isolated after their separation. Accordingly, evolutionary trees are constructed considering the possibility of gene flow between populations after their split. A maximum-likelihood tree was initially inferred from allele frequencies, with migration events added to populations that showed a poor fit to this tree. We modeled several scenarios allowing a number of migration events from 0 to 10, until (*i*) 99% of the variance in relatedness was explained and (*ii*) further migration events did not significantly increase the variance explained by the model. The trees were forced to have a root in Africa. Interestingly, the topography of the maximum-likelihood tree places the Aeta/Agta in a branch with Australians, Papuans, and Melanesians (Fig. S4*A*). It also reveals a strong likelihood of admixture between Japanese (JP) and Aeta/Agta (NE), with an inferred migration from the former to the latter.

Following these exploratory analyses, we placed Papuans and Melanesians as descendants of the first dispersal and Agta/Aeta as descendants of the second dispersal for the MDI model. Because we grouped the Aeta and Agta as one population, our results are not directly comparable to those of ref. 5. However, we similarly interpret these analyses to suggest that the Aeta/Agta might have descended from an early southern route dispersal but have been strongly admixed with a subsequent dispersal. Because living Aeta/Agta speak an Austric language and given the inferred migration from Japan, such admixture might largely be consequent of the Holocene Austronesian expansion of mainland Asian populations into the Pacific (1).

1. Endicott P (2013) Revisiting the "negrito" hypothesis: A transdisciplinary approach to human prehistory in southeast Asia. *Hum Biol* 85(1–3):7–20.
2. Mirazón Lahr M (1996) *The Evolution of Modern Human Diversity: A Study of Cranial Variation* (Cambridge Univ Press, Cambridge, UK).
3. Abdulla MA, et al.; HUGO Pan-Asian SNP Consortium; Indian Genome Variation Consortium (2009) Mapping human genetic diversity in Asia. *Science* 326(5959): 1541–1545.
4. Pugach I, Delfin F, Gunnarsdóttir E, Kayser M, Stoneking M (2013) Genome-wide data substantiate Holocene gene flow from India to Australia. *Proc Natl Acad Sci USA* 110 (5):1803–1808.
5. Rasmussen M, et al. (2011) An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* 334(6052):94–98.
6. Zheng X, et al. (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28(24):3326–3328.
7. Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet* 11(1):94.
8. Cavalli-Sforza LL (2005) The Human Genome Diversity Project: Past, present and future. *Nat Rev Genet* 6(4):333–340.
9. Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. *Nature* 461(7263):489–494.
10. Xing J, et al. (2010) Toward a more uniform sampling of human genetic diversity: A survey of worldwide populations by high-density genotyping. *Genomics* 96(4):199–210.
11. Xing J, et al. (2009) Fine-scaled human genetic structure revealed by SNP microarrays. *Genome Res* 19(5):815–825.
12. Bryc K, et al. (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci USA* 107(2):786–791.
13. Reich D, et al. (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet* 89(4):516–528.
14. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155(2):945–959.
15. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655–1664.
16. Liu N, Zhao H (2006) A non-parametric approach to population structure inference using multilocus genotypes. *Hum Genomics* 2(6):353–364.
17. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 8(11):e1002967.

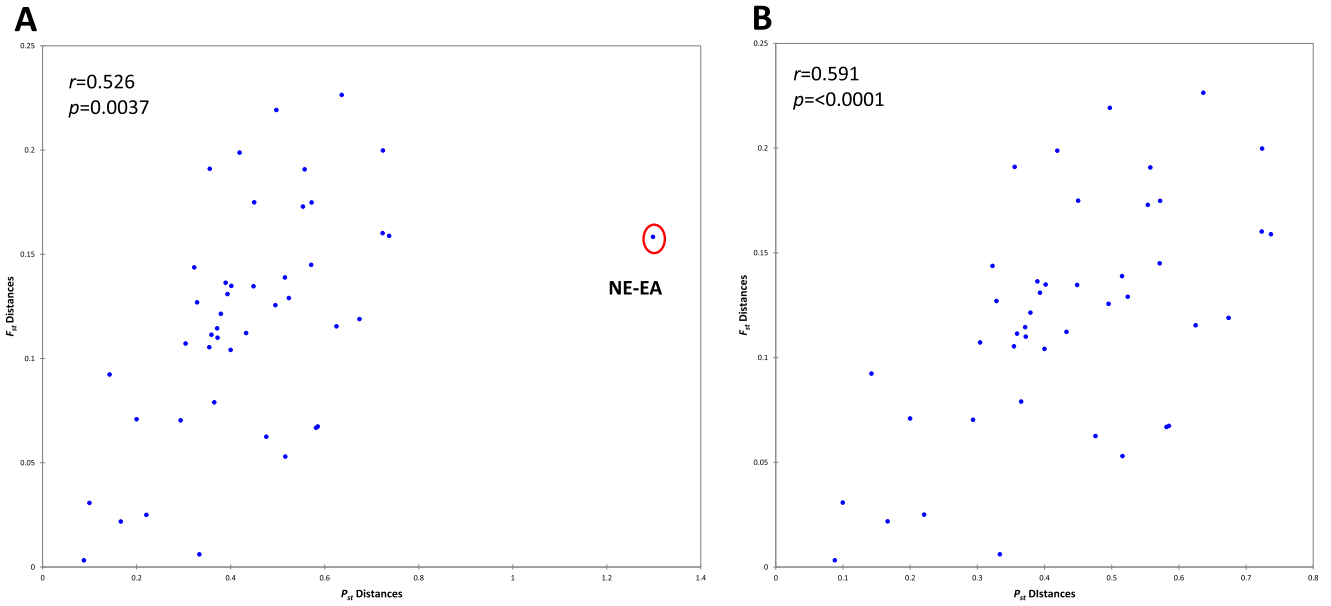

Fig. S1. F_{st}–P_{st} correlation. (*A*) Regression of F_{st} and P_{st} values, with presence of outlier (EA-NE population pair) highlighted. (*B*) Regression of F_{st} and P_{st} values after removal of outlier. Reported values are Pearson correlation, *r*, and two-tailed probability, *p*, after 10,000 permutations.

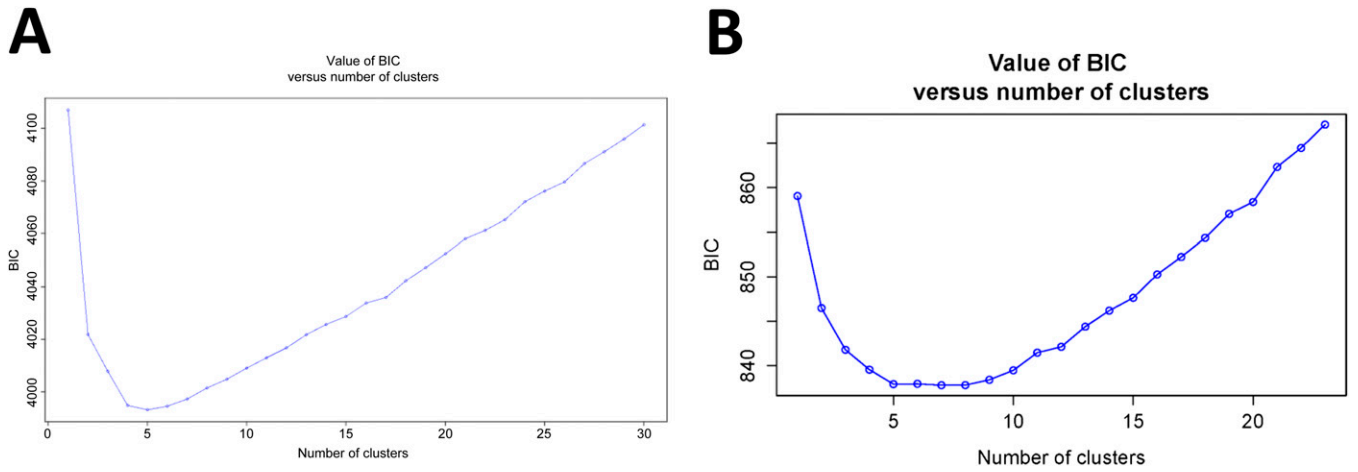

Fig. S2. BIC of *K* clusters for genetic (*A*) and cranial phenotype (*B*) datasets.

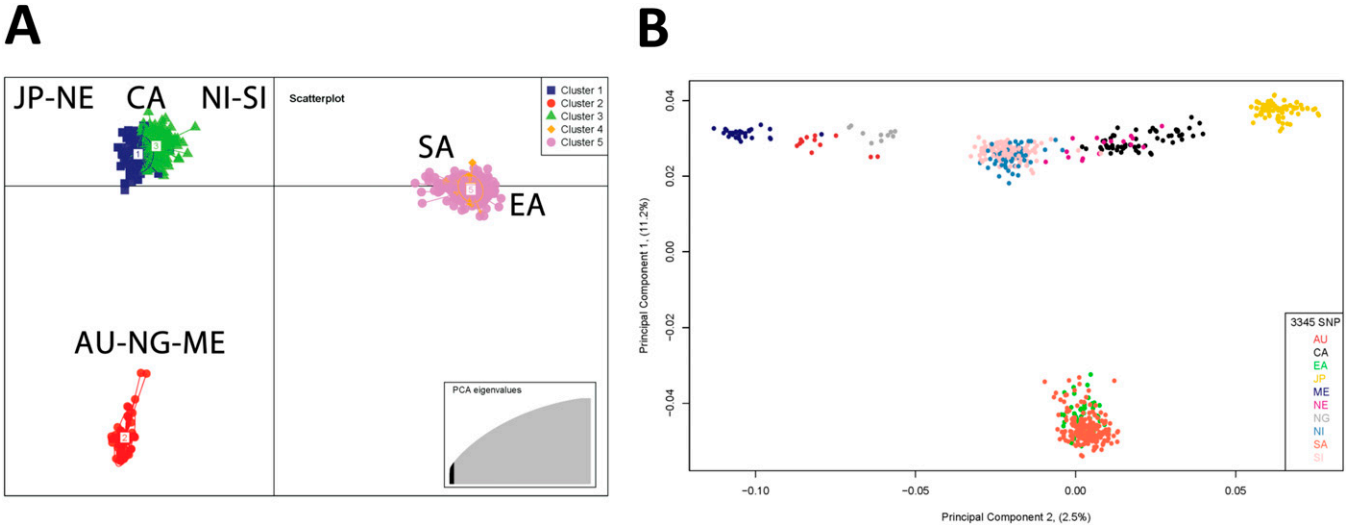

Fig. S3. Genetic DAPC and PCA scatterplots. (*A*) $K = 5$ DAPC scatterplot along the first two axes. (*B*) PCA of SNP genetic data along the first two axes, capturing ~13.7% of total variance.

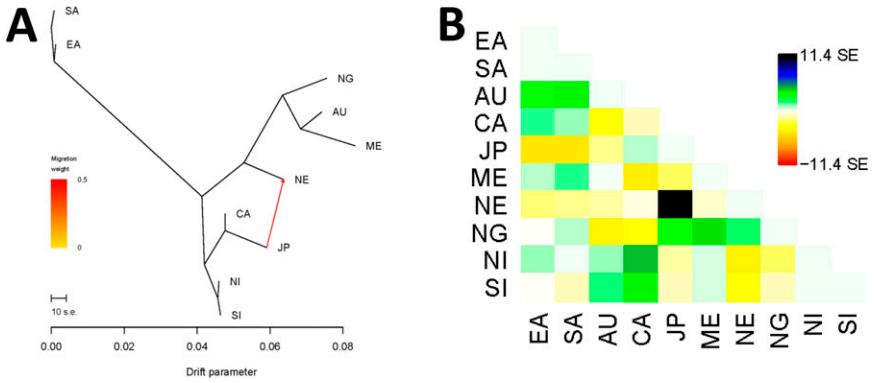

Fig. S4. TreeMix Analysis. (*A*) Maximum-likelihood tree of population affinity. Red arrow indicates migration event and directionality. (*B*) Residual values from the tree. Values above zero indicate pairs of populations that are candidates for admixture events.

Table S1. Population divergence values

Population	AU	CA	EA	JP	ME	NE	NG	NI	SA	SI
AU	0	35,339	100,422	45,489	19,359	25,110	19,645	41,286	106,548	32,642
CA	1,262.1	0	70,440	10,248	42,428	12,718	28,151	8,949	77,815	8,406
EA	3,586.5	2,515.7	0	96,054	113,476	69,835	91,255	74,738	4,306	67,586
JP	1,624.6	366	3,430.5	0	49,496	14,431	33,457	29,123	101,514	23,486
ME	691.4	1,515.3	4,052.7	1,767.7	0	28,711	19,314	50,590	116,348	38,301
NE	896.8	454.2	2,494.1	515.4	1,025.4	0	16,355	21,011	75,592	15,725
NG	701.6	1,005.4	3,259.1	1,194.9	689.8	584.1	0	37,176	96,986	27,608
NI	1,474.5	319.6	2,669.2	1,040.1	1,806.8	750.4	1,327.7	0	82,225	1,266
SA	3,805.3	2,779.1	153.8	3,625.5	4,155.3	2,699.7	3,463.8	2,936.6	0	74,609
SI	1,165.8	300.2	2,413.8	838.8	1,367.9	561.6	986	45.2	2,664.6	0

Below diagonal: generations (*T*); above diagonal: *T* expressed in calendar years, assuming generations of 28 y and rounded to the nearest whole number. AU, Australia; CA, Central Asia; EA, East Africa; JP, Japan; ME, Melanesia; NE, Philippines Aeta/Agta "Negrito"; NG, New Guinea; NI, North India; SA, South Africa; and SI, South India.

Table S2. Effective population size (N_e)

Population	N_e
AU	4,784
CA	6,057
EA	12,167
JP	5,692
ME	3,626
NE	2,304
NG	2,462
NI	8,464
SA	13,174
SI	5,824

Table S3. Genetic and cranial samples

Population	Cranial subpopulations	Genetic subpopulations	Language family*	Genetic data (refs.)
AU	Australian	Australian	Australian	4
CA	Uyghur, Dungan, Kalmyk, Tarantchi	Uyghur, Kyrgystani	Eurasiatic, Dene-Caucasian	3, 8, 10
EA	Amhara, Karo, Habesha, Bouma, Glaba, Turkana, Igai, Koukou, Afar-Danakil, Nyangatom, Pouma	Alur, Bulala, Kaba, Mada, Hausa	Nilo-Saharan, Afro-Asiatic	10–12
JP	Japanese	Japanese	Eurasiatic	3, 8, 10, 11
ME	Solomon and Vanuatu Islanders	Papua New Guinea highlands	Indo-Pacific	4, 8, 13
NE	Agta, Aeta	Aeta, Agta	Austric	3
NG	Papua New Guinea, Torres Strait Islanders	Papua New Guinea "lowlands" (Bougainville)	Indo-Pacific	3, 8
NI	Bengali	Kashmiri Pandit, Vaish, Srivastava, Sahariya, Lodi, Satnami, Bhil, Tharu, Meghawal	Indo-European	4, 9
SA	Xhosa, Khoi, Nama, San, Sotho, Malabar, Zulu, Tswana	San, Bantu, !Kung, Pedi, Dogon, Bambara, Nguni, Sotho/Twana, Mbuti Pygmy, Hema, Luhya, Bamoun, Fang, Kongo, Xhosa	Khoisan, Niger-Kordofanian	10–12
SI	Maravar, Tamil	Vysya, Naidu, Velama, Kamsali, Chenchu, Kurumba, Hallaki, Dalit, Mala, Madiga, Irula	Dravidian	3, 4, 9–11

*Language families as defined by J. Greenberg (8).

Table S4. Geographical waypoints used in dispersal models

	Geographic coordinates	
Waypoints*	Latitude	Longitude
Bangkok	13.73	100.52
Cairo	30.06	31.24
Chennai	13.06	80.24
Colombo	6.93	79.86
Dhaka	23.71	90.41
Dubai	25.27	55.31
Jakarta	−6.21	106.84
Karachi	24.89	67.03

*Locations correspond to Fig. 1 of the main text.

Table S5. DAPC classification

	DAPC cluster									
	1		2		3		4		5	
Population	Genetics	Phenotype	Genetics	Phenotype	Genetics	Phenotype	Genetics	Phenotype	Genetics	Phenotype
AU	0	1 (5%)	12 (100%)	11 (55%)	0	6 (30%)	0	0	0	2 (10%)
CA	25 (45%)	7 (28%)	0	2 (8%)	31 (55%)	10 (40%)	0	3 (12%)	0	3 (12%)
EA	0	1 (4%)	0	2 (8%)	0	3 (12%)	0	7 (28%)	66 (100%)	12 (48%)
JP	107 (100%)	12 (39%)	0	4 (13%)	0	8 (26%)	0	5 (16%)	0	2 (6%)
ME	0	0	30 (100%)	3 (18%)	0	2 (12%)	0	8 (47%)	0	4 (23%)
NE	15 (94%)	9 (39%)	0	0	1 (6%)	13 (56%)	0	0	0	1 (4%)
NG	0	4 (13%)	10 (100%)	15 (48%)	0	4 (13%)	0	5 (16%)	0	3 (10%)
NI	0	5 (33%)	0	1 (7%)	61 (100%)	3 (20%)	0	3 (20%)	0	3 (20%)
SA	0	1 (5%)	0	2 (10%)	0	6 (30%)	41 (19%)	1 (5%)	174 (81%)	10 (50%)
SI	0	10 (38%)	0	0	141 (100%)	3 (12%)	0	8 (31%)	0	5 (19%)

Classification number and rate (in parentheses, approximate percent of total).

Other Supporting Information Files

Dataset S1 (XLSX)