RECONSTRUCTING SOCIAL PREHISTORY FROM GENOMIC DATA IN THE INDO-PACIFIC REGION

MURRAY COX

A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF SCIENCE

UNIVERSITY OF OTAGO, DUNEDIN, NEW ZEALAND

SUBMITTED 11 JANUARY 2016
ACCEPTED 12 APRIL 2017

ABSTRACT

Reconstructing the evolutionary history of our species has traditionally been the purview of archeology and linguistics, but is now increasingly influenced by genetics. However, the information held in our DNA cannot be read like a book, but must instead be extracted using population genetic theory, advanced statistical methods and computational tools that can handle large genome-scale datasets. In this series of published studies, these approaches have been applied to reconstruct human prehistory, with a special focus on the social features of past communities in the Indo-Pacific region. They reveal that marriage between Asian women and Melanesian men was favored during the spread of farming populations in the Neolithic period, that Madagascar was settled by a small number of Indonesian families with close female relatives, and that extremely complex marriage rules continue to define and structure small traditional communities in the Indo-Pacific region even today. These studies are largely unique in moving beyond a traditional emphasis in molecular anthropology of identifying and dating human migrations to instead reveal key aspects of the social rules by which those communities lived.

CONTENTS

1

EXTENDED SUMMARY

The history of our species is recorded in the DNA we carry inside us. Coupled with archeology and linguistics, human genetic diversity allows us to reconstruct prehistoric processes, such as the movements and interactions of people, the diseases they overcame, and increasingly, even the social choices they made.

Although often described as a book, the DNA record is no simple read. Past events typically imposed only subtle changes on patterns of genetic diversity. Consequently, complex statistical and computational methods, often borrowed from physics and mathematics, must be applied to vast genome-scale datasets to extract historical information. My biggest contributions have been made in this area (http://massey.genomicus.com).

Leveraging a diverse background in genetics, anthropology, statistics and computer science, I specialize in integrating methods drawn from diverse fields with population genetic theory, and implementing these re-imagined methods as new computational tools that can handle the 'big data' of modern genetics, particularly in the areas of simulation modeling and statistical inference. This in turn has led to major new discoveries in such areas as genetic dating, determining the cultural processes involved in population interactions, and reconstructing past social behaviors, including male dominance and the role of marriage rules in structuring small traditional communities.

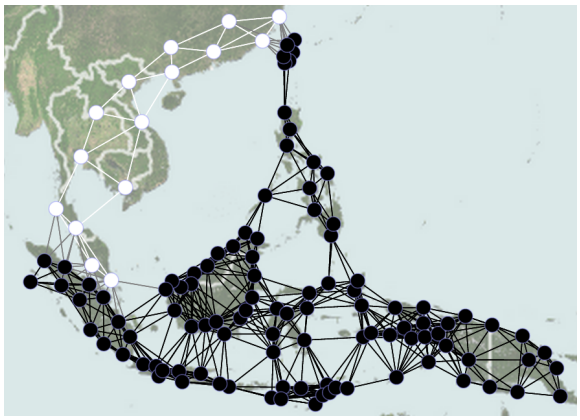

Figure 1: A representative computer model showing communities (nodes) and connections between them (lines) across Island Southeast Asia. Modified from Vallée et al (2016).

Much of this work has focused on a long series of case studies across a broad swathe of the Indo-Pacific region (Figure 1). This part of the world has a deep history spanning its initial settlement 50,000 years ago, through the arrival of farming populations during the Neolithic period, to historic era interactions with India and the Middle East. Today, populations across this

region speak thousands of mutually unintelligible languages and harbor a diverse array of cultural practices (Figure 2). Many groups are still quite isolated, relatively untouched in key ways by modern western society, thus allowing us to treat parts of the Indo-Pacific region as a 'living laboratory' in which we can compare and contrast patterns of genetic diversity across islands with very different languages, cultural practices and historical trajectories. When coupled with new theoretical, statistical and computational approaches (Figure 3), these case studies have allowed us to reveal major new aspects of the Indo-Pacific past.

Most research in contemporary molecular anthropology is still heavily focused on determining and dating prehistoric migrations. In contrast, one of the special features of my work is its aim of reconstructing the social aspects of past communities from the signals left behind in the genetic record. Here are just three representative examples:

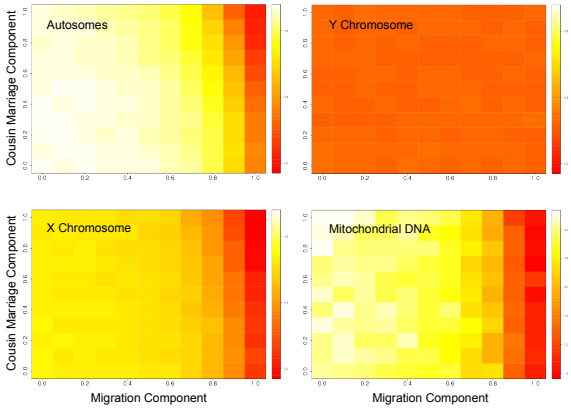

Figure 2: Simulations showing changes in genetic diversity on the autosomes, X chromosome, Y chromosome and mitochondrial DNA under different parameters of a complex marriage model. Modified from Guillot et al (2015).

- During the spread of Neolithic farming populations, groups with ultimate Asian ancestry encountered and interacted with existing Melanesian groups. Comparison of autosomal and sex-linked genetic loci reveals a cultural bias favoring marriage between Asian women and Melanesian men.

- Madagascar, an island off the east coast of Africa, was settled by migrants from Indonesia who spoke a sister language to one still used in southern Borneo today. A computer simulation of the settlement process, focused on maternally inherited genetic diversity from both Madagascar and Indonesia, reveals that only a small number of Indonesians settled the island during the mid first century AD, and they were dominated by a cohort of closely related Indonesian women.

- Most traditional communities are characterized by complex marriage rules, which structure interactions between villages, increasing trade and inhibiting warfare. Statistical modeling of genome-scale data reveals that these rules have been persistent

and common, but importantly, may also have been strategically ignored when necessary.

These snapshots highlight common themes of my research: its emphasis on extremely large genetic datasets; a prominent role for modeling, computer simulation and statistical inference; and the development and implementation of new software (http://massey.genomicus.com/software.html) specifically targeted to questions of major anthropological interest. My emphasis on teasing out past social history from modern genetic data is also a striking point of difference from most modern molecular anthropology research.

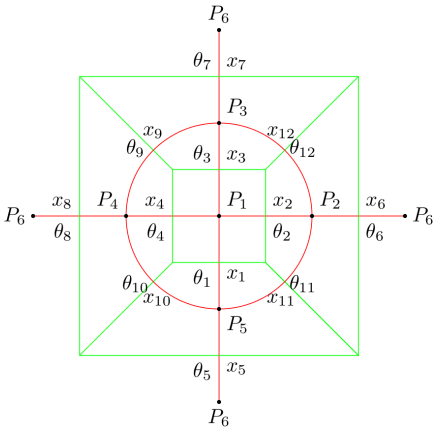

Figure 3: Topological representation showing links between communities and boundaries of community territories. From Amaris et al (2015).

Since completing my PhD, I have established a substantial body of literature, with over 50 research papers published in this subject area alone over the last decade. While my scientific contributions should be judged from the totality of my research portfolio, the following ten papers highlight key aspects of my research program, emphasizing its breadth and impact. To underscore the progression of my research career to date, these papers are simply presented in reverse chronological order.

For reprints, please visit http://massey.genomicus.com/publications.html

- CHAPTER 3 *Statistical modeling of genome-scale data reveals that complex marriage rules were persistent and common in small traditional communities, but importantly, may also have been ignored when necessary.*

  Guillot, E.G., M.L. Hazelton, T.M. Karafet, J.S. Lansing, H. Sudoyo, and **M.P. Cox**. 2015. Relaxed observance of traditional marriage rules allows social connectivity without loss of genetic diversity. *Molecular Biology and Evolution* 32:2254-2262. http://doi.org/10.1093/molbev/msv102

- CHAPTER 4 *A new method of wavelet decomposition was developed and applied to Indonesian genome-scale data to reconstruct and date*

past population mixing processes, including multiple contact events.

Sanderson, J., H. Sudoyo, T.M. Karafet, M.F. Hammer and **M.P. Cox**. 2015. Reconstructing past admixture processes from local genomic ancestry using wavelet transformation. *Genetics* 200:469-481.
http://doi.org/10.1534/genetics.115.176842

- CHAPTER 5 *A survey of male- and female-linked genetic diversity in ~3000 individuals from 70 Indonesian communities reveals the region's history, including strong sex-specific dispersal patterns.*

  Tumonggor, M.K., T.M. Karafet, B. Hallmark, J.S. Lansing, H. Sudoyo, M.F. Hammer, and **M.P. Cox**. 2013. The Indonesian archipelago: An ancient genetic highway linking Asia and the Pacific. *Journal of Human Genetics* 58:165-173.
  http://doi.org/10.1038/jhg.2012.154

- CHAPTER 6 *Computer simulations show that Madagascar was settled by a small number of Indonesian migrants, who were closely related along the female line, during the mid first millenium AD.*

  Cox, M.P., M.G. Nelson, M.K. Tumonggor, F.-X. Ricaut, and H. Sudoyo. 2012. A small cohort of Island Southeast Asian women founded Madagascar. *Proceedings of the Royal Society B* 279:2761-2768.
  http://doi.org/10.1098/rspb.2012.0012

- CHAPTER 7 *A model based on the tenets of structural anthropology explains the observed sex bias favoring marriage between Asian women and Melanesian men during the Neolithic expansion.*

  Lansing, J.S., **M.P. Cox**, T.A. de Vet, S.S. Downey, B. Hallmark and H. Sudoyo. 2011. An ongoing Austronesian expansion in Island Southeast Asia. *Journal of Anthropological Archaeology* 30:262-272.
  http://doi.org/10.1016/j.jaa.2011.06.004

- CHAPTER 8 *Comparisons of autosomal and sex-linked genetic loci reveal a bias in marriages between Asian women and Melanesian men during the spread of farming populations in the Neolithic.*

  Cox, M.P., T.M. Karafet, J.S. Lansing, H. Sudoyo, and M.F. Hammer. 2010. Autosomal and X-linked single nucleotide polymorphisms reveal a steep Asian-Melanesian ancestry cline in eastern Indonesia and a sex bias in admixture rates. *Proceedings of the Royal Society B* 277:1589-1596.
  http://doi.org/10.1098/rspb.2009.2041

- CHAPTER 9 *A simulation study based on the population genetic framework of coalescent theory shows that a widely-used genetic dating method is not robust to key demographic features of human prehistory.*

  Cox, M.P. 2008. Accuracy of molecular dating with the rho statistic: Deviations from coalescent expectations under a range of demographic models. *Human Biology* 80:335-357.
  http://www.jstor.org/stable/41466549

- CHAPTER 10 *Computational analysis of a region on the human X chromosome showed introgression from an archaic hominin ancestor into modern humans before any archaic genomes had been published.*

  Cox, M.P., F.L. Mendez, T.M. Karafet, M. Metni Pilkington, S.B. Kingan, G. Destro-Bisol, B.I. Strassmann, and M.F. Hammer. 2008. Testing for archaic hominin admixture on the X-chromosome: Model likelihoods for the modern human *RRM2P4* region from summaries of genealogical topology under the structured coalescent. *Genetics* 178:427-437.
  http://doi.org/10.1534/genetics.107.080432

- CHAPTER 11 *A statistical study of male diversity showing that male dominance, where men with higher social status have more children, was less common than expected during recent Indonesian prehistory.*

  Lansing, J.S., J.C. Watkins, B. Hallmark, **M.P. Cox**, T.M. Karafet, H. Sudoyo, and M.F. Hammer. 2008. Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations. *Proceedings of the National Academy of Sciences USA* 105:11645-11650.
  http://doi.org/10.1073/pnas.0710158105

- CHAPTER 12 *One of the first molecular 'community level', or microgeographic, surveys showing that communities and the languages they speak co-evolve over extremely small spatial and temporal scales.*

  Lansing, J.S., **M.P. Cox**, S.S. Downey, B.M. Gabler, B. Hallmark, T.M. Karafet, P. Norquest, J.W. Schoenfelder, H. Sudoyo, J.C. Watkins, and M.F. Hammer. 2007. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proceedings of the National Academy of Sciences USA* 104:16022-16026.
  http://doi.org/10.1073/pnas.0704451104

2

PUBLICATIONS

The following pages report papers published since the completion of my PhD up to the submission of this thesis. (References of 'in press' papers that were published after submission of this thesis were updated at the time of acceptance). The >50 publications that relate directly to the subject area of this application are bolded.

For reprints, please visit http://massey.genomicus.com

70. **Cox, M.P.** 2017. Calibrating the clock. In D.H. O'Rourke, *A Companion to Anthropological Genetics*. Wiley: New Jersey. In Press.

69. **Cox, M.P.** 2017. The genetic history of human populations in Island Southeast Asia during the Late Pleistocene and Holocene. In P. Bellwood, *First Islanders: The Prehistory of Island Southeast Asia*. Wiley: New Jersey. http://www.wiley.com/WileyCDA/WileyTitle/productCd-1119251540.html

68. Bradshaw, R.E., Y. Guo, A. Sim, M.S. Kabir, P. Chettri, I.K. Ozturk, L. Hunziker, R.L. Ganley, and **M.P. Cox**. 2016. Genome-wide gene expression dynamics of the fungal pathogen *Dothistroma septosporum* throughout its infection cycle of the gymnosperm host *Pinus radiata*. *Molecular Plant Pathology* 17:210-224. http://doi.org/10.1111/mpp.12273

67. Kusuma, P., **M.P. Cox**, N. Brucato, H. Sudoyo, T. Letellier, and F.-X. Ricaut. 2016. Western Eurasian genetic influences in the Indonesian Archipelago. *Quaternary International* 416:243-248. http://doi.org/10.1016/j.quaint.2015.06.048

66. Amaris, A.J.R., and **M.P. Cox**. 2015. A flexible theoretical representation for the temporal dynamics of structured populations as paths on polytope complexes. *Journal of Mathematical Biology* 71:735-766. http://doi.org/10.1007/s00285-014-0841-4

65. Berry, D., **M.P. Cox**, and B. Scott. 2015. Draft genome sequence of the filamentous fungus *Penicillium paxilli* (ATCC 26601). *Genome Announcements* 3:e00071-15. http://doi.org/10.1128/genomeA.00071-15

64. Burgess, S.A., **M.P. Cox**, S.H. Flint, D. Lindsay, and P.J. Biggs. 2015. Draft genome sequences of three strains of *Geobacillus stearothermophilus* isolated from a milk powder manufacturing

plant. *Genome Announcements* 3:e00939-15.
http://doi.org/10.1128/genomeA.00939-15

63. Duchemin, W., P.-Y. Dupont, M.A. Campbell, A.R.D. Ganley, and **M.P. Cox**. 2015. HyLiTE: Accurate and flexible analysis of gene expression in hybrid and allopolyploid species. *BMC Bioinformatics* 16:8.
http://doi.org/10.1186/s12859-014-0433-8

62. Dupont, P.-Y., C.J. Eaton, J.J. Wargent, S. Fechtner, P. Solomon, J. Schmid, R.C. Day, D.B. Scott, and **M.P. Cox**. 2015. Fungal endophyte infection of ryegrass reprograms host metabolism and alters development. *New Phytologist* 208:1227-1240.
http://doi.org/10.1111/nph.13614

61. Eaton, C.J., P.-Y. Dupont, P. Solomon, W. Clayton, B. Scott, and **M.P. Cox**. 2015. A core gene set describes the molecular basis of mutualism and antagonism in *Epichloë* spp. *Molecular Plant-Microbe Interactions* 28:218-231.
http://doi.org/10.1094/MPMI-09-14-0293-FI

60. Guillot, E.G., M.L. Hazelton, T.M. Karafet, J.S. Lansing, H. Sudoyo, and **M.P. Cox**. 2015. Relaxed observance of traditional marriage rules allows social connectivity without loss of genetic diversity. *Molecular Biology and Evolution* 32:2254-2262.
http://doi.org/10.1093/molbev/msv102

59. Karmin, M., L. Saag, M. Vicente, M.A.W. Sayres, M. Järve, U.G. Talas, S. Rootsi, A.-M. Ilumäe, R. Mägi, M. Mitt, L. Pagani, T. Puurand, Z. Faltyskova, F. Clemente, A. Cardona, E. Metspalu, H. Sahakyan, B. Yunusbayev, G. Hudjashov, M. DeGiorgio, E.-L. Loogväli, C. Eichstaedt, M. Eelmets, G. Chaubey, K. Tambets, S. Litvinov, M. Mormina, Y. Xue, Q. Ayub, G. Zoraqi, T.S. Korneliussen, F. Akhatova, J. Lachance, S. Tishkoff, K. Momynaliev, F.-X. Ricaut, P. Kusuma, H. Razafindrazaka, D. Pierron, **M.P. Cox**, G.N.N. Sultana, R. Willerslev, C. Muller, M. Westaway, D. Lambert, V. Skaro, L. Kovačević, S. Turdikulova, D. Dalimova, R. Khusainova, N. Trofimova, V. Akhmetova, I. Khidiyatova, D.V. Lichman, J. Isakova, E. Pocheshkhova, Z. Sabitov, N.A. Barashkov, P. Nymadawa, E. Mihailov, J.W.T. Seng, I. Evseeva, A.B. Migliano, S. Abdullah, G. Andriadze, D. Primorac, L. Atramentova, O. Utevska, L. Yepiskoposyan, D. Marjanović, A. Kushniarevich, D.M. Behar, C. Gilissen, L. Vissers, J. Veltman, E. Balanovska, M. Derenko, B. Malyarchuk, A. Metspalu, S. Fedorova, A. Eriksson, A. Manica, F. Mendez, T.M. Karafet, K. Veeramah, N. Bradman, M.F. Hammer, L.P. Osipova, O. Balanovsky, E.K. Khusnutdinova, K. Johnsen, M. Remm, M.G. Thomas, C. Tyler-Smith, P.A. Underhill, E. Willerslev, R. Nielsen,

M. Metspalu, R. Villems, and T. Kivisild. 2015. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Research* 25:459-466.
http://doi.org/10.1101/gr.186684.114

58. Kusuma, P., **M.P. Cox**, D. Pierron, H. Razafindrazaka, N. Brucato, L. Tonasso, H.L. Suryadi, T. Letellier, H. Sudoyo, and F.-X. Ricaut. 2015. Mitochondrial DNA and the Y chromosome suggest the settlement of Madagascar by Indonesian sea nomad populations. *BMC Genomics* 16:191.
http://doi.org/10.1186/s12864-015-1394-7

57. Nicholson, M.J., C.J. Eaton, C. Stärkel, B.A. Tapper, **M.P. Cox**, and B. Scott. 2015. Molecular cloning and functional analysis of gene clusters for the biosynthesis of indole-diterpenes in *Penicillium crustosum* and *P. janthinellum*. *Toxins* 7:2701-2722.
http://doi.org/10.3390/toxins7082701

56. Sanderson, J., H. Sudoyo, T.M. Karafet, M.F. Hammer and **M.P. Cox**. 2015. Reconstructing past admixture processes from local genomic ancestry using wavelet transformation. *Genetics* 200:469-481.
http://doi.org/10.1534/genetics.115.176842

55. **Cox, M.P.**, T. Dong, G. Shen, Y. Dalvi, D.B. Scott, and A.R.D. Ganley. 2014. An interspecific fungal hybrid reveals cross-kingdom rules for allopolyploid gene expression patterns. *PLoS Genetics* 10:e1004180.
http://doi.org/10.1371/journal.pgen.1004180

54. Guillot, E.G., and **M.P. Cox**. 2014. SMARTPOP: Inferring the impact of social dynamics on genetic diversity through high speed simulations. *BMC Bioinformatics* 15:175.
http://doi.org/10.1186/1471-2105-15-175

53. Lansing, J.S., S.A. Cheong, L.Y. Chew, **M.P. Cox**, M.-H.R. Ho, and W.A. Arthawiguna. 2014. Regime shifts in Balinese *subaks*. *Current Anthropology* 55:232-239.
http://www.jstor.org/stable/10.1086/675429

52. Prangle, D., P. Fearnhead, **M.P. Cox**, P.J. Biggs, and N.P. French. 2014. Semi-automatic selection of summary statistics for ABC model choice. *Statistical Applications in Genetics and Molecular Biology* 13:67-82.
http://doi.org/10.1515/sagmb-2013-0012

51. Schardl, C.L., C.A. Young, N. Moore, N. Krom, P.-Y. Dupont, J. Pan, S. Florea, J.S. Webb, J. Jaromczyk, J.W. Jaromczyk, **M.P. Cox**,

and M.L. Farman. 2014. Genomes of plant-associated Clavicipitaceae. *Advances in Botanical Research* 70:291-327. http://doi.org/10.1016/B978-0-12-397940-7.00010-0

50. Tumonggor, M.K., T.M. Karafet, S. Downey, J.S. Lansing, P. Norquest, H. Sudoyo, M. Hammer, and **M.P. Cox**. 2014. Isolation, contact and social behavior shaped genetic diversity in West Timor. *Journal of Human Genetics* 59:494-503. http://doi.org/10.1038/jhg.2014.62

49. Bradshaw, R.E., J.C. Slot, G.G. Moore, P. Chettri, P.J.G.M. de Wit, K.C. Ehrlich, A.R.D. Ganley, M.A. Olson, A. Rokas, I. Carbone, and **M.P. Cox**. 2013. Fragmentation of an aflatoxin-like gene cluster in a forest pathogen. *New Phytologist* 198:525-535. http://doi.org/10.1111/nph.12161

48. **Cox, M.P.** 2013. Southeast Asia Islands and Oceania: Human Genetics. In *Encyclopedia of Global Human Migration*. I. Ness and P. Bellwood (eds), Wiley-Blackwell: Malden, MA, USA and Oxford, UK. Volume 1, Prehistory, Chapter 37, pp. 293-301. http://doi.org/10.1002/9781444351071.wbeghm837

47. **Cox, M.P.**, B.R. Holland, M.C. Wilkins, and J. Schmid. 2013. Reconstructing past changes in locus-specific recombination rates. *BMC Genetics* 14:11. http://doi.org/10.1186/1471-2156-14-11

46. Chettri, P., K.C. Ehrlich, J.W. Cary, J. Collemare, **M.P. Cox**, S.A. Griffiths, M.A. Olson, P.J.G.M. de Wit, and R.E. Bradshaw. 2013. Dothistromin genes at multiple separate loci are regulated by AflR. *Fungal Genetics and Biology* 51:12-20. http://doi.org/10.1016/j.fgb.2012.11.006

45. Guillot, E.G., M.K. Tumonggor, J.S. Lansing, H. Sudoyo, and **M.P. Cox**. 2013. Climate change influenced female population sizes through time across the Indonesian archipelago. *Human Biology* 85:135-152. http://www.jstor.org/stable/10.13110/humanbiology.85.1-3.0135

44. Schardl, C.L., C.A. Young, U. Hesse, S.G. Amyotte, K. Andreeva, P.J. Calie, D.J. Fleetwood, D.C. Haws, N. Moore, B. Oeser, D.G. Panaccione, K.K. Schweri, C.R. Voisey, M.L. Farman, J.W. Jaromczyk, B.A. Roe, D.M. O'Sullivan, B. Scott, P. Tudzynski, Z. An, E.G. Arnaoudova, C.T. Bullock, N.D. Charlton, L. Chen, **M.P. Cox**, R.D. Dinkins, S. Florea, A.E. Glenn, A. Gordon, U. Güldener, D.R. Harris, W. Hollin, J. Jaromczyk, R.D. Johnson, A.K. Khan, E. Leistner, A. Leuchtmann, C. Li, J.-G. Liu, J. Liu, M. Liu, W. Mace, C. Machado, P. Nagabhyru, J. Pan, J. Schmid,

K. Sugawara, U. Steiner, J.E. Takach, E. Tanaka, J.S. Webb, E.V. Wilson, J. Wiseman, R. Yoshida, and Z. Zeng. 2013. Plant-symbiotic fungi as chemical engineers: Multi-genome analysis of the Clavicipitaceae reveals dynamics of alkaloid loci. *PLoS Genetics* 9:e1003323.
http://doi.org/10.1371/journal.pgen.1003323

43. Tumonggor, M.K., T.M. Karafet, B. Hallmark, J.S. Lansing, H. Sudoyo, M.F. Hammer, and **M.P. Cox**. 2013. The Indonesian archipelago: An ancient genetic highway linking Asia and the Pacific. *Journal of Human Genetics* 58:165-173.
http://doi.org/10.1038/jhg.2012.154

42. Wehi, P., **M.P. Cox**, T. Roa and H. Whaanga. 2013. Marine resources in Māori oral tradition: He kai moana, he kai mā te hinengaro. *Journal of Marine and Island Cultures* 2:59-68.
http://doi.org/10.1016/j.imic.2013.11.006

41. **Cox, M.P.**, M.G. Nelson, M.K. Tumonggor, F.-X. Ricaut, and H. Sudoyo. 2012. A small cohort of Island Southeast Asian women founded Madagascar. *Proceedings of the Royal Society B* 279:2761-2768.
http://doi.org/10.1098/rspb.2012.0012

40. de Wit, P.J.G.M., A. van der Burgt, B. Ökmen, I. Stergiopoulos, A.L. Aerts, A.H.A. Bahkali, H. Beenen, P. Chettri, **M.P. Cox**, E. Datema, R.P. de Vries, B. Dhillon, A.R. Ganley, Y. Guo, R.C. Hamelin, B. Henrissat, M.S. Kabir, M.K. Jashni, G. Kema, S. Klaubauf, A. Lapidus, A. Levasseur, E. Lindquist, S. Lucas, R. Mehrabi, R.A. Ohm, T.J. Owen, A. Salamov, A. Schwelm, E. Schijlen, H. Sun, H.A. van den Burg, R.C.H.J. van Ham, S. Zhang, S.B. Goodwin, I.V. Grigoriev, J. Collemare, and R.E. Bradshaw. 2012. The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. *PLoS Genetics* 8:e1003088.
http://doi.org/10.1371/journal.pgen.1003088

39. Eaton, C.J., I.E. Cabrera, J.A. Servin, S.J. Wright, **M.P. Cox**, and K.A. Borkovich. 2012. The guanine nucleotide exchange factor RIC8 regulates conidial germination through G_α proteins in *Neurospora crassa*. *PLoS ONE* 7:e48026.
http://doi.org/10.1371/journal.pone.0048026

38. Ricaut, F.-X.*, **M.P. Cox***, M. Lacan, C. Keyser, F. Duranthon, B. Ludes, J. Guilaine, and E. Crubézy. 2012. A time series of prehistoric mitochondrial DNA reveals western European genetic diversity was largely established by the Bronze Age. *Advances*

in Anthropology 2:14-23. *Joint first authors
http://doi.org/10.4236/aa.2012.21002

37. Wilder, J.A., **M.P. Cox**, A.M. Paquette, R. Alford, A.W. Satyagraha, A. Harahap, and H. Sudoyo. 2011. Genetic continuity across a deeply divergent linguistic contact zone in North Maluku, Indonesia. *BMC Genetics* 12:100.
http://doi.org/10.1186/1471-2156-12-100

36. Eaton, C.J., **M.P. Cox**, and D.B. Scott. 2011. What triggers grass endophytes to switch from mutualism to pathogenism? *Plant Science* 180:190-195.
http://doi.org/10.1016/j.plantsci.2010.10.002

35. Lansing, J.S., and **M.P. Cox**. 2011. The domain of the replicators: Selection, neutrality and cultural evolution. *Current Anthropology* 52:105-125.
http://www.jstor.org/stable/10.1086/657643

34. Lansing, J.S., **M.P. Cox**, T.A. de Vet, S.S. Downey, B. Hallmark and H. Sudoyo. 2011. An ongoing Austronesian expansion in Island Southeast Asia. *Journal of Anthropological Archaeology* 30:262-272.
http://doi.org/10.1016/j.jaa.2011.06.004

33. Russell, A.L.*, **M.P. Cox***, V.A. Brown, and G.F. McCracken. 2011. Population growth of Mexican free-tailed bats (*Tadarida brasiliensis mexicana*) predates human agricultural activity. *BMC Evolutionary Biology* 11:88. *Joint first authors
http://doi.org/10.1186/1471-2148-11-88

32. **Cox, M.P.**, C.J. Eaton, and D.B. Scott. 2010. Exploring molecular signaling in plant-fungal symbioses using high throughput RNA sequencing. *Plant Signaling and Behavior* 5:1353-1358.
http://doi.org/10.4161/psb.5.11.12950

31. **Cox, M.P.**, and M.F. Hammer. 2010. A question of scale: Human migrations writ large and small. *BMC Biology* 8:98.
http://doi.org/10.1186/1741-7007-8-98

30. **Cox, M.P.**, T.M. Karafet, J.S. Lansing, H. Sudoyo, and M.F. Hammer. 2010. Autosomal and X-linked single nucleotide polymorphisms reveal a steep Asian-Melanesian ancestry cline in eastern Indonesia and a sex bias in admixture rates. *Proceedings of the Royal Society B* 277:1589-1596.
http://doi.org/10.1098/rspb.2009.2041

29. **Cox, M.P.**, and J.S. Lansing. 2010. Invited Commentary on "Farming and language in Island Southeast Asia: Reframing Austronesian history". *Current Anthropology* 51:241-242.
http://www.jstor.org/stable/10.1086/650991

28. **Cox, M.P.**, D.A. Peterson, and P.J. Biggs. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11:485.
http://doi.org/10.1186/1471-2105-11-485

27. Eaton, C.J., **M.P. Cox**, B. Ambrose, M. Becker, U. Hesse, C.L. Schardl, and D.B. Scott. 2010. Disruption of signaling in a fungal-grass symbiosis leads to pathogenesis. *Plant Physiology* 153:1780-1794.
http://doi.org/10.1104/pp.110.158451

26. Hammer, M.F., A.E. Woerner, F.L. Mendez, J.C. Watkins, **M.P. Cox**, and J.D. Wall. 2010. The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nature Genetics* 42:830-831.
http://doi.org/10.1038/ng.651

25. Karafet, T.M., B. Hallmark, **M.P. Cox**, H. Sudoyo, S. Downey, J.S. Lansing, and M.F. Hammer. 2010. Major east-west division underlies Y chromosome stratification across Indonesia. *Molecular Biology and Evolution* 27:1833-1844.
http://doi.org/10.1093/molbev/msq063

24. Razafindrazaka, H., F.-X. Ricaut, **M.P. Cox**, M. Mormina, J.-M. Dugoujon, L.P. Randriamarolaza, E. Guitard, L. Tonasso, B. Ludes, and E. Crubézy. 2010. Complete mitochondrial DNA sequences provide new insights into the Polynesian motif and the peopling of Madagascar. *European Journal of Human Genetics* 18:575-581.
http://doi.org/10.1038/ejhg.2009.222

23. Ricaut, F.-X., T. Thomas, M. Mormina, **M.P. Cox**, M. Bellatti, R. Foley, and M. Mirazón Lahr. 2010. Ancient Solomon Islands mtDNA: Assessing Holocene settlement and the impact of European contact. *Journal of Archaeological Science* 37:1161-1170.
http://doi.org/10.1016/j.jas.2009.12.014

22. **Cox, M.P.**, D.A. Morales, A.E. Woerner, J. Sozanski, J.D. Wall, and M.F. Hammer. 2009. Autosomal resequence data reveal Late Stone Age signals of population expansion in sub-Saharan African foraging and farming populations. *PLoS ONE* 4:e6366.
http://doi.org/10.1371/journal.pone.0006366

21. Lansing, J.S., **M.P. Cox**, S.S. Downey, M.A. Janssen, and J.W. Schoenfelder. 2009. A robust budding model of Balinese water temple networks. *World Archaeology* 41:112-133.
http://doi.org/10.1080/00438240802668198

20. Ricaut, F.-X., H. Razafindrazaka, **M.P. Cox**, J.-M. Dugoujon, E. Guitard, C. Sambo, M. Mormina, M. Mirazón Lahr, B. Ludes,

and E. Crubézy. 2009. A new deep branch of Eurasian mtDNA macrohaplogroup M reveals additional complexity regarding the settlement of Madagascar. *BMC Genomics* 10:605. http://doi.org/10.1186/1471-2164-10-605

19. **Cox, M.P.** 2008. Accuracy of molecular dating with the rho statistic: Deviations from coalescent expectations under a range of demographic models. *Human Biology* 80:335-357. http://www.jstor.org/stable/41466549

18. **Cox, M.P.** 2008. The genetic environment of Melanesia: Clines, clusters and contact. In V.T. Koven (ed.), *Population Genetics Research Progress*, Chapter 2. Nova Science Publishers: New York, pp 45-83. https://www.novapublishers.com/catalog/product_info.php?products_id=7023

17. **Cox, M.P.**, F.L. Mendez, T.M. Karafet, M. Metni Pilkington, S.B. Kingan, G. Destro-Bisol, B.I. Strassmann, and M.F. Hammer. 2008. Testing for archaic hominin admixture on the X-chromosome: Model likelihoods for the modern human *RRM2P4* region from summaries of genealogical topology under the structured coalescent. *Genetics* 178:427-437. http://doi.org/10.1534/genetics.107.080432

16. **Cox, M.P.**, A.E. Woerner, J.D. Wall, and M.F. Hammer. 2008. Intergenic DNA sequences from the human X chromosome reveal high rates of global gene flow. *BMC Genetics* 9:e76. http://doi.org/10.1186/1471-2156-9-76

15. Downey, S.S., B. Hallmark, **M.P. Cox**, P. Norquest, and J.S. Lansing. 2008. Computational feature-sensitive reconstruction of language relationships: Developing the ALINE distance for comparative historical linguistic reconstruction. *Journal of Quantitative Linguistics* 15:340-369. http://doi.org/10.1080/09296170802326681

14. Hagelberg, E., **M.P. Cox**, W. Schiefenhövel, and I. Frame. 2008. A genetic perspective on the origins and dispersal of the Austronesians: Mitochondrial DNA variation from Madagascar to Easter Island. In A. Sanchez-Mazas, R. Blench, M.D. Ross, I. Peiros, and M. Lin (eds.), *Past Human Migrations in East Asia: Matching Archaeology, Linguistics and Genetics*, Chapter 16. Routledge: London, pp 356-375. http://doi.org/10.4324/9780203926789

13. Hammer, M.F., F.L. Mendez, **M.P. Cox**, A.E. Woerner, and J.D. Wall. 2008. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genetics* 4:e1000202. http://doi.org/10.1371/journal.pgen.1000202

12. Lansing, J.S., J.C. Watkins, B. Hallmark, **M.P. Cox**, T.M. Karafet, H. Sudoyo, and M.F. Hammer. 2008. Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations. *Proceedings of the National Academy of Sciences USA* 105:11645-11650.
http://doi.org/10.1073/pnas.0710158105

11. Metni Pilkington, M., J.A. Wilder, F.L. Mendez, **M.P. Cox**, A. Woerner, T. Angui, S. Kingan, Z. Mobasher, C. Batini, G. Destro-Bisol, H. Soodyall, B.I. Strassmann, and M.F. Hammer. 2008. Contrasting signatures of population growth for mitochondrial DNA and Y chromosomes among human populations in Africa. *Molecular Biology and Evolution* 25:517-525.
http://doi.org/10.1093/molbev/msm279

10. Russell, A.L., S.M. Goodman, and **M.P. Cox**. 2008. Coalescent analyses support multiple mainland-to-island dispersals in the evolution of Malagasy *Triaenops* bats (Chiroptera: Hipposideridae). *Journal of Biogeography* 35:995-1003.
http://doi.org/10.1111/j.1365-2699.2008.01891.x

9. Wall, J.D., **M.P. Cox**, F.L. Mendez, A. Woerner, T. Severson, and M.F. Hammer. 2008. A novel DNA sequence database for analyzing human demographic history. *Genome Research* 18:1354-1361.
http://doi.org/10.1101/gr.075630.107

8. **Cox, M.P.**, A.J. Redd, T.M. Karafet, C.A. Ponder, J.S. Lansing, H. Sudoyo, and M.F. Hammer. 2007. A Polynesian motif on the Y chromosome: Population structure in Remote Oceania. *Human Biology* 79:525-535.
http://www.jstor.org/stable/41466509

7. Garrigan, D., S.B. Kingan, M. Metni Pilkington, J.A. Wilder, **M.P. Cox**, H. Soodyall, B. Strassmann, G. Destro-Bisol, P. de Knijff, A. Novelletto, J. Friedlaender, and M.F. Hammer. 2007. Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. *Genetics* 177:2195-2207.
http://doi.org/10.1534/genetics.107.077495

6. Lansing, J.S., **M.P. Cox**, S.S. Downey, B.M. Gabler, B. Hallmark, T.M. Karafet, P. Norquest, J.W. Schoenfelder, H. Sudoyo, J.C. Watkins, and M.F. Hammer. 2007. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proceedings of the National Academy of Sciences USA* 104:16022-16026.
http://doi.org/10.1073/pnas.0704451104

5. Woerner, A.E., **M.P. Cox**, and M.F. Hammer. 2007. Recombination-filtered genomic datasets by information

maximization. *Bioinformatics* 23:1851-1853.
http://doi.org/10.1093/bioinformatics/btm253

4. **Cox, M.P.** 2006. Extreme patterns of variance in small populations: Placing limits on human Y-chromosome diversity through time in the Vanuatu Archipelago. *Annals of Human Genetics* 71:390-406.
http://doi.org/10.1111/j.1469-1809.2006.00327.x

3. **Cox, M.P.** 2006. Minimal hierarchical analysis of global human Y-chromosome SNP diversity by PCR-RFLP. *Anthropological Science* 114:69-74.
http://doi.org/10.1537/ase.050712

2. **Cox, M.P.**, and M. Mirazón Lahr. 2006. Y-chromosome diversity is inversely associated with language affiliation in paired Austronesian- and Papuan-speaking communities from Solomon Islands. *American Journal of Human Biology* 18:35-50.
http://doi.org/10.1002/ajhb.20459

1. **Cox, M.P.** 2005. Indonesian mitochondrial DNA and its opposition to a Pleistocene era origin of proto-Polynesians in Island Southeast Asia. *Human Biology* 77:179-188.
http://www.jstor.org/stable/41466315

3

RELAXED OBSERVANCE OF TRADITIONAL MARRIAGE RULES ALLOWS SOCIAL CONNECTIVITY WITHOUT LOSS OF GENETIC DIVERSITY

Guillot, E.G., M.L. Hazelton, T.M. Karafet, J.S. Lansing, H. Sudoyo, and **M.P. Cox**. 2015. Relaxed observance of traditional marriage rules allows social connectivity without loss of genetic diversity. *Molecular Biology and Evolution* 32:2254-2262.
http://doi.org/10.1093/molbev/msv102

ABSTRACT Marriage rules, the community prescriptions that dictate who an individual can or cannot marry, are extremely diverse and universally present in traditional societies. A major focus of research in the early decades of modern anthropology, marriage rules impose social and economic forces that help structure societies and forge connections between them. However, in those early anthropological studies, the biological benefits or disadvantages of marriage rules could not be determined. We revisit this question by applying a novel simulation framework and genome-wide data to explore the effects of Asymmetric Prescriptive Alliance, an elaborate set of marriage rules that has been a focus of research for many anthropologists. Simulations show that strict adherence to these marriage rules reduces genetic diversity on the autosomes, X chromosome and mitochondrial DNA, but relaxed compliance produces genetic diversity similar to random mating. Genome-wide data from the Indonesian community of Rindi, one of the early study populations for Asymmetric Prescriptive Alliance, are more consistent with relaxed compliance than strict adherence. We therefore suggest that, in practice, marriage rules are treated with sufficient flexibility to allow social connectivity without significant degradation of biological diversity.

SUPPLEMENTARY MATERIALS
http://mbe.oxfordjournals.org/content/32/9/2254/suppl/DC1

Relaxed Observance of Traditional Marriage Rules Allows Social Connectivity without Loss of Genetic Diversity

Elsa G. Guillot,[1] Martin L. Hazelton,[1] Tatiana M. Karafet,[2] J. Stephen Lansing,[3] Herawati Sudoyo,[4] and Murray P. Cox*[,1]

[1]Statistics and Bioinformatics Group, Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

[2]ARL Division of Biotechnology, University of Arizona

[3]Complexity Institute, Nanyang Technological University, Singapore

[4]Eijkman Institute for Molecular Biology, Jakarta, Indonesia

***Corresponding author:** E-mail: m.p.cox@massey.ac.nz.

Associate editor: Connie Mulligan

Abstract

Marriage rules, the community prescriptions that dictate who an individual can or cannot marry, are extremely diverse and universally present in traditional societies. A major focus of research in the early decades of modern anthropology, marriage rules impose social and economic forces that help structure societies and forge connections between them. However, in those early anthropological studies, the biological benefits or disadvantages of marriage rules could not be determined. We revisit this question by applying a novel simulation framework and genome-wide data to explore the effects of Asymmetric Prescriptive Alliance, an elaborate set of marriage rules that has been a focus of research for many anthropologists. Simulations show that strict adherence to these marriage rules reduces genetic diversity on the autosomes, X chromosome and mitochondrial DNA, but relaxed compliance produces genetic diversity similar to random mating. Genome-wide data from the Indonesian community of Rindi, one of the early study populations for Asymmetric Prescriptive Alliance, are more consistent with relaxed compliance than strict adherence. We therefore suggest that, in practice, marriage rules are treated with sufficient flexibility to allow social connectivity without significant degradation of biological diversity.

Key words: **mating systems, Asymmetric Prescriptive Alliance, genetic diversity, Indonesia, Approximate Bayesian Computation.**

Article

Introduction

Human societies are characterized by a myriad of often elaborate marriage rules. Describing the diversity of these rules and their central role in the organization of human communities was once a major focus of anthropological research (Mascie-Taylor and Boyce 1988). Following the seminal works of Van Wouden (1935) and Lévi-Strauss (1949) in the 1930s and 1940s, studies of marriage systems proliferated during the 1950s and 1960s (Lévi-Strauss 1965; Gilbert and Hammel 1966; Jacquard 1967, 1970; MacCluer et al. 1971), but interest declined as anthropologists recognized that they lacked the necessary tools to formally test the hypotheses they had developed. Today, access to large genetic data sets and fast computer simulation provides a springboard to revisit many of these historically unanswered questions.

Marriage rules are universal, yet extraordinarily diverse (Lévi-Strauss 1949). Although enforcement varies, all communities, both traditional and westernized, impose at least some constraints on who individuals can or cannot marry. Many marriage rules are famously intricate, such as Asymmetric Prescriptive Alliance (APA), which is characterized by a complex, but clearly defined, intergenerational framework (fig. 1) (Needham 1964; Beatty 1990; Forth 2009). Men are required to marry their mother's brother's daughter (MBD), and women move from "wife-giver" to "wife-taker" communities in what Van Wouden (1935) describes as a "circulating connubium." Consequently, although women move in one direction around the network of communities, bride wealth flows in the opposite direction. Hence marriage to the MBD's clan creates asymmetric alliance ties between patrilines. APA is particularly common in the small islands of eastern Indonesia (Forth 1990), but found globally, it is unclear why such intricate marriage rules emerged, almost certainly independently, in multiple places at multiple times.

Marriage rules appear to be important for both social and biological reasons. The connections they create often underpin stable long-term trade and support networks (Winterhalder and Smith 2000; Marlowe 2003; Huber et al. 2011; Henrich et al. 2012), but because marriage is also the primary institution leading to offspring, marriage rules should strongly affect patterns of genetic diversity, particularly in small traditional communities. Research on this question has been surprisingly limited. Departures from expected genetic patterns have sometimes been attributed to social factors (Watkins 2004; Chaix et al. 2007; Moorad et al. 2011; Heyer et al. 2012), such as reduced Y chromosome diversity under polygyny (Lansing et al. 2008), where few men are permitted to marry and reproduce, or excessive

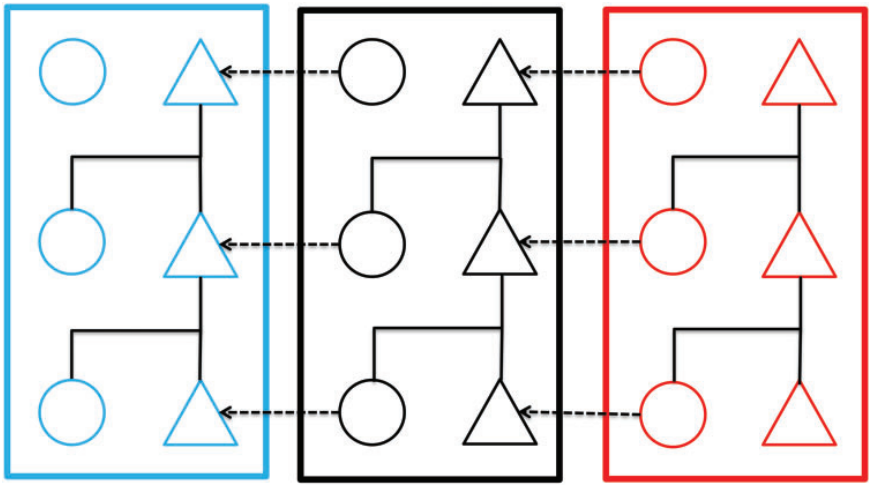

FIG. 1. Kinship under APA. Each box represents a clan; each row represents a generation. The red clan acts as wife-giver to the black clan, which in turn acts as wife-giver to the blue clan. Dashed arrows represent marriage alliances with women moving from their natal clan to the community of their husband.

mitochondrial DNA (mtDNA) diversity under patrilocality (Tumonggor et al. 2013), where women preferentially move between villages. Comparison of X chromosome and autosomal diversity has also revealed other social constructs, such as sex-specific patterns of postmarital dispersal (Ségurel et al. 2008; Verdu et al. 2013) and sex-biased admixture (Cox et al. 2010). Theoretical studies are even rarer. The effects of polygyny (Guillot and Cox 2014) and marriage alliances around ring communities (Billari et al. 2007) have been modeled. Both studies required special methods, as most marriage rules cannot be studied using standard population genetic theory, such as the coalescent, which assumes random mating, asexual populations, and the crucial trait of exchangeability (Kingman 1982). In consequence, the biological effects of marriage rules remain largely unexplored.

Here, we ask whether marriage rules affect patterns of genetic diversity. We determine whether marriage systems, such as APA, produce biological benefits or disadvantages; whether asymmetry is an important biological feature of such marriage systems; and whether the effects of marriage rules can be meaningfully detected in real-world genetic data. To address these questions, we employ new modeling software, SMARTPOP, which we designed specifically to simulate the intricate marriage rules observed in human communities (Guillot and Cox 2014). Finally, we also use genome-wide single nucleotide polymorphism (SNP) data from Rindi, a community in eastern Indonesia that practices APA (Forth 1981), to estimate the historic degree of compliance with this mating system.

Results

The first major question is whether and how marriage rules affect patterns of genetic diversity. We explored this for APA by simulating data across a grid of values for π_{MBD}, the probability that the MBD rule is followed, and π_{mig}, the probability that a woman migrates according to the wife-giver/wife-taker scheme. In the anthropological literature, APA typically follows a classificatory rule that defines all female siblings of the same generation within a clan as equal (Maybury-Lewis 1965), thus obliging men to marry a woman from the wife-giving group, but not specifically their cousin ($\pi_{MBD} = 0$, $\pi_{mig} = 1$). However, unlike this more general case, in Rindi the genealogical cousin is explicitly preferred (Forth 1981). Our framework, which separates the migration and cousin marriage aspects of the APA rule system, allows us to represent the entire gamut of APA-type systems.

Under a strict interpretation of APA ($\pi_{MBD} = 1$, $\pi_{mig} = 1$), genetic diversity shows reproducible reductions in genetic diversity compared with random mating ($\pi_{MBD} = 0$, $\pi_{mig} = 0$) (fig. 2). Nonlinear regressions show that APA influences genetic diversity on the autosomes, X chromosome and mtDNA, but not the paternally inherited Y chromosome (table 1). This is expected because the system modeled here is patrilocal—women move between communities to marry, whereas men remain in their natal clan (fig. 1). This holds true whether women follow the APA rules or not. However, following the MBD rule negatively affects genetic diversity on the autosomes and X chromosome, and following the wife-giver/wife-taker migration rule negatively affects diversity on the autosomes, X chromosome, and mtDNA. As shown by the regression coefficients, this loss of genetic diversity is driven more by constraints on migration than cousin mating. We therefore conclude that strong adherence to APA is detrimental to the maintenance of genetic diversity.

The equivalent symmetric migration scheme (Symmetric Prescriptive Alliance [SPA]) decreases genetic diversity in a similar manner to APA across the parameter space. We therefore suggest that symmetry versus asymmetry in the movements of women has little biological effect.

The second question is the extent to which marriage rules can be followed stringently. Within a strict APA system, not all men can follow the rule to marry their MBD, as an

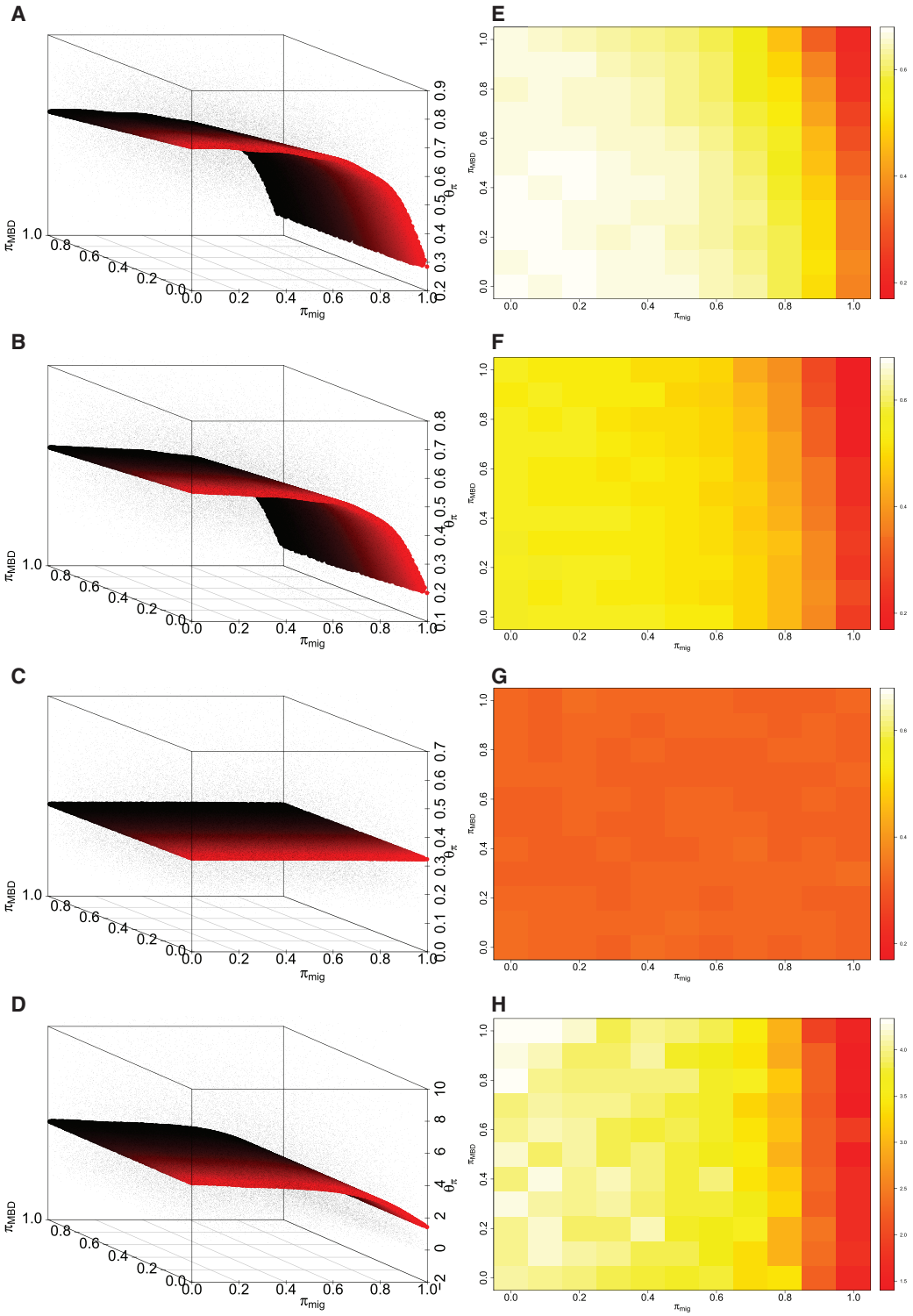

Fig. 2. Genetic diversity (θ_π) across a grid of random values for π_{MBD} and π_{mig} under APA for (*A*,*E*) autosomes, (*B*,*F*) X chromosome, (*C*,*G*) Y chromosome, and (*D*,*H*) mtDNA over 50,000 simulations (3,000 individuals, 20 demes). (*A*)–(*D*) present fitted values (sheet with black to red color range for better visualization in 3D) from the GAMs, together with simulated values (dots). (*E*)–(*H*) are projections of simulated mean diversity values across the grid of parameters.

Table 1. Local GAM to Fit Mean Pairwise Diversity θ_π as a Function of π_{MBD} and π_{mig}.

Genomic Region	R^2	$P(\pi_{MBD})$	$P(\pi_{mig})$
Autosomes	**0.43**	***0***	***0***
X chromosome	**0.35**	***0***	***0***
Y chromosome	**7.0×10^{-5}**	**0.10**	**0.24**
mtDNA	**0.091**	**0.35**	***0***

NOTE.—Significant values for the correlation with π_{MBD} and π_{mig} are shown in italics (50,000 simulations).

appropriate individual of the right sex is not always available. The effects of deviating from the MBD rule have been discussed previously (Kunstadter et al. 1963; Ackerman 1964; McFarland 1970; MacCluer et al. 1971; Fredlund 1976; Mascie-Taylor and Boyce 1988), but not determined objectively due to the unavailability of computer simulation as a readily accessible tool in the 1950s–1970s. In our simulations, we can measure the rate at which MBD marriage actually occurs, as a function of π_{MBD} and π_{mig} (fig. 3). In the extreme case where rules are followed strictly, on average only 30% of marriages can be made to a MBD due to the unavailability of an appropriate partner. As π_{MBD} decreases (x axis), more individuals marry a random partner rather than following the rule, leading to lower rates of MBD marriages. When there is compliance with the migration rule, but not the cousin alliance rule, MBD marriages still occur at moderate frequency (~12%) due to the random chance of marrying the right cousin. However, when the migration rule is not followed, women move to a nonprescribed clan, where no appropriate cousin is present for them to marry. Therefore, as π_{mig} decreases (colored lines), the actual rate of MBD marriage decreases rapidly as well.

The third question is whether the effects of marriage rules can be inferred in practice. We applied Approximate Bayesian Computation (ABC) to genome-wide data from Rindi, a population on the eastern Indonesian island of Sumba in which APA has been well studied. Figure 4 shows posterior distributions for N, π_{MBD}, and π_{mig}. The population size is estimated at nearly 7,000 individuals (mode = 6,608, 95% credible region 5,086–10,630). We estimate that the migration rule is followed in just over half of cases (mode = 0.59, 95% credible region 0.02−0.93). The mating parameter also shows moderate compliance with the rule (mode = 0.56), but again has a wide 95% credible region (0.03−0.98). Despite considerable uncertainty in these values, the modes of the posterior densities suggest that the Rindi community has not followed either of the extreme cases—strict APA or random mating. In practice, therefore, this community would fall near the center of the graphs in figure 2, where genetic diversity does not differ markedly from random mating.

Cross-validation reveals the accuracy of parameter inference by testing simulated cases with known values (see supplementary material, Supplementary Material online, for details). Population size ($E_{pred} = 0.94$) and the migration parameter ($E_{pred} = 1.24$) show moderate linear relationships

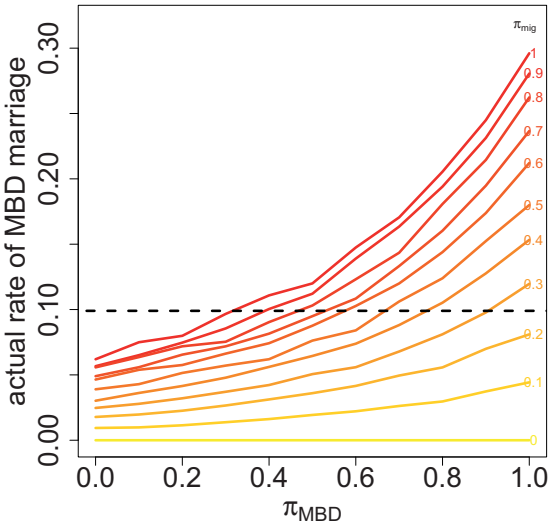

FIG. 3. Actual rate of MBD marriage observed in simulations as a function of π_{MBD} (x axis) and π_{mig} (color scale) (25,000 observations during 50 simulations for each data point). The dashed line represents the rate of MBD marriage observed by Forth (1981) in Rindi (10%); the observed migration rate to the prescribed clan is 26%.

between estimated and real values, indicating that the inference procedure has reasonable statistical power to infer these parameters. Less power is available to infer the mating parameter ($E_{pred} = 1.53$).

Finally, using simulated data sets, we asked how genomic sequence data, without the ascertainment bias of SNP genotyping chips, would improve statistical inference. Although predictions for population size ($E_{pred} = 0.54$) are improved, values for the migration ($E_{pred} = 0.94$) and mating parameters ($E_{pred} = 1.51$) suggest that it will always prove challenging to infer mating systems from genomic data, even when they are unbiased. Larger sample sizes and targeted clan sampling designs may help ameliorate these issues.

Discussion

Marriage rules are a ubiquitous feature of all traditional societies. From early in the twentieth century, an extensive body of anthropological literature has attempted to determine their purpose. Although little consensus was reached on the details, anthropologists developed a broadly held view that marriage rules help structure connections within and between communities, and that they therefore play a fundamental role in social cohesion. However, any social rules that affect marriage also have a direct impact on offspring, and hence, the genetics of communities. In small communities, which were the only type that existed throughout most of human history, genetic diversity is easily lost through genetic drift, which in turn can lead to reduced individual fitness, lower reproductive success, increased levels of genetic disorders, and ultimately, community extinction (Ober et al. 1999; Winata et al. 1995).

Before this study, it was not known whether marriage rules help or hinder the maintenance of genetic diversity. We are

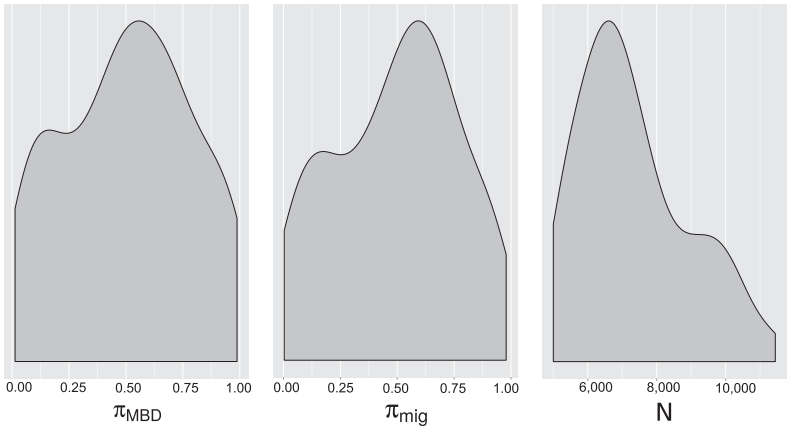

Fig. 4. ABC for Rindi, a population practicing APA on Sumba in eastern Indonesia. Posterior distributions are shown for *N*, π_{MBD}, and π_{mig}.

able to address this question through two recent advances. First, the availability of new simulation tools that can simultaneously model marriage rules and population genetics. Most population genetic methods do not explicitly model two sexes (men and women), let alone specific marriage rules. New computer programs, such as SMARTPOP (Guillot and Cox 2014), now make extensive modeling of social rules and population genetics possible. Second, marriage rules act at very small scales: Either within communities or across a small cluster of communities. Community-level sampling (as opposed to regional collections from schools or medical clinics) is now starting to become more common.

We apply these advances to address the role of APA in mediating patterns of genetic variation in Rindi, the well-studied APA community on Sumba, a small island in eastern Indonesia. APA describes a specific form of cousin marriage, which also implies a regular intergenerational movement of women between clans. Our theoretical simulations show that APA reduces genetic diversity, but only when it is followed strictly. In particular, the migration component of the APA rules elicits a dramatic decline in diversity if followed by more than 80% of women. More variable compliance with the marriage rules leads to genetic diversity that does not differ markedly from random mating.

Here, we make the necessary simplification that the alliance model has been relatively stable through time. Although these simulations assume adherence to APA during the recent past, other scenarios are of course possible. For instance, a recent shift to APA would weaken the effect of the rules on genetic diversity. Occasional reassortment of clans, a process known as fusion–fission (Smouse et al. 1981; Chaix et al. 2007), could also variously increase or decrease genetic diversity among the groups. As genomic data sets improve, it should become feasible to model such complex social dynamics. The simpler scenario modeled here provides an obligatory first step.

The two parameters that underpin APA, the MBD rule and wife-giver/wife-taker migration, together with the population size, were inferred for Rindi from genome-wide SNP chip data. The results show that the size of the Rindi population is large, which is consistent with female immigration due to patrilocal postmarriage residence patterns (Guillot et al. 2013). Patterns of genetic diversity in Rindi seem most consistent with intermediate values of π_{MBD} and π_{mig}, thus arguing against random mating or a strict adoption of the APA rules.

Forth (1981) undertook a detailed ethnographic study of Rindi in the 1970s. He observed the actual proportion of MBD marriages (10%), as well as the proportion of marriages to the prescribed clan (26%). There is no simple relationship between the modeled parameters and the observed rate of cousin marriage, as an appropriate spouse may not be available even if the rules are followed strictly by the community. Using the simulated relationships in figure 3, we deduce that Forth's observed rates imply a theoretical compliance with the MBD rule of approximately 90% for those individuals who do marry into the prescribed clan. Our overall estimates of π_{MBD} and π_{mig} from the genomic data, 56% and 59%, respectively, differ from those Forth observed. However, we note that our values represent long-term averages, perhaps suggesting that adherence to the APA rules once varied from Forth's observations in the late twentieth century. Our estimates are most consistent with only moderate long-term compliance with the migration and marriage rules, which in turn would help the community to maintain genetic diversity.

Marriage rules are therefore perhaps best viewed as convenient ideologies: Revered more in theory than in practice. Nevertheless, these results show that marriage rules have important biological outcomes for communities, and that strict adherence can be biologically disadvantageous. We argue that the flexibility with which marriage rules are implemented in practice is therefore not so much a problem as the key point. Although small human communities almost certainly do not think in genetic terms, there are both social and biological reasons to overlook violations of marriage rules. The moderate observance rate in Rindi suggests relatively weak enforcement of marriage rules in this community. Elsewhere, strict compliance can be driven by strong sanctions against transgressors, often mediated through belief systems, and not

uncommonly leading to the ultimate sanction, death (Lansing 2006).

This study shows how modern computer simulations can provide new insight into old anthropological questions. The stochastic behavior of individuals, such as instances where the required spouse is unavailable or a different spouse is chosen for an alternative social reason, appears to be a dominant feature of traditional marriage systems. The effects of deviating from a strict interpretation of marriage rules can now be modeled, as can other community choices, such as symmetric versus asymmetric migration. The addition of a genetic element to these models further allows exploration of the effects of marriage systems on biological diversity. Symmetric migration between communities in an SPA setting produces much the same biological outcome as APA. Hence, the preference for APA over SPA in eastern Indonesia (Forth 1990) may be better explained by socioeconomic factors, such as the long-term stability of asymmetric wife-giver/wife-taker exchange, which creates enduring networks of relationships between patrilocal kin groups (Van Wouden 1935; Lévi-Strauss 1965).

The addition of a statistical inference framework to our theoretical work allows us to estimate the long-term biological effects of marriage rules on specific communities such as Rindi. Although the statistical power of the analyses presented here is relatively low, this partly reflects the need for summary statistics that are able to circumvent a small sample size and the ascertainment bias found in current genotyping chips. Unbiased data from whole-genome sequencing will become increasingly common in coming years and the approach presented here is ready to take full advantage of these new data. However, power analyses show that reconstructing mating systems from any sort of genetic data will always be a challenging undertaking.

We do, however, show that genetic evidence has the potential to reconstruct aspects of the social systems by which communities historically lived. Marriage rules are ubiquitous, but we suggest that it is unlikely they were followed strictly. The majority of these violations probably had prosaic local causes. In many cases, the individual required by the marriage rule may not have been available to marry. Alternately, reduced genetic diversity in small communities quickly leads to the accumulation of genetic disorders. Although communities presumably had little understanding of genetic inheritance, they may have linked social behaviors, such as adherence to marriage rules, to unfavorable biological outcomes. Certainly, reduced genetic diversity under a strict interpretation of the APA marriage rules suggests that there was little biological incentive for communities to enforce marriage rules strongly, at least for long periods of time. Whether this holds true across the wide gamut of marriage rules recorded globally by anthropologists is now a question that can feasibly be revisited.

Materials and Methods

Ethics

Biological samples were collected by J.S.L., H.S., and a team from the Eijkman Institute for Molecular Biology, with the assistance of Indonesian Public Health clinic staff, following protocols for the protection of human subjects established by both the Eijkman Institute and the University of Arizona institutional review boards. Permission to conduct research in Indonesia was granted by the State Ministry of Research and Technology.

Sampling and Genetic Screening

Genetic markers were screened in 28 consenting, closely unrelated and apparently healthy individuals from Rindi, a community on the eastern Indonesian island of Sumba. Apart from excluding immediate relatives, individuals were approached randomly during the course of a medical visit. Participant interviews confirmed ethnic, linguistic, and geographic affiliations with Rindi for at least two generations into the past. mtDNA markers are as described elsewhere (Tumonggor et al. 2013). Autosomal ($n = 664{,}475$), X chromosome ($n = 16{,}034$), and Y chromosome ($n = 266$) SNPs were screened in 24 individuals using the Illumina HumanOmniExpress-24 BeadChip (GeneByGene, Houston, TX). SNP chip genotype data for Rindi are available from the authors on request.

Computer Simulations

Simulations were run with SMARTPOP v.2.0 (Guillot and Cox 2014), a free open-source C++ forward-in-time simulator, which was purpose-built to model the effects of marriage rules on human communities. Twenty demes of equal size (150 individuals in the regression, varying population sizes in the ABC study) were modeled, thus approximating the number of clans in the APA system recorded for Rindi (Forth 1981). Genetic data were simulated across four genomic regions: 200 unlinked loci on the autosomes ($n = 200$; 32 bp), 10 unlinked loci on the X chromosome ($n = 10$; 1,000 bp), a fully linked locus on the Y chromosome ($n = 1$; 10,000 bp), and a fully linked mtDNA locus ($n = 1$; 544 bp). This data structure was selected to mimic key features of the real data set as closely as possible, while still meeting nontrivial constraints imposed by runtime speed in the obligatory forward-in-time simulation setting. Slightly more individuals were sampled for mtDNA ($n = 28$) than nuclear loci ($n = 24$) to match the observed data.

Demes evolved through phases of migration, mating, and mutation at each generation. Due to considerable uncertainty surrounding human mutation rates (Scally and Durbin 2012) and relative insensitivity to exact values in this component of the analysis, average mutation rates were employed for the autosomes (2.5×10^{-7} mutations/site/generation), X chromosome (2.5×10^{-7} mutations/site/generation), Y chromosome (2.5×10^{-7} mutations/site/generation), and mtDNA (4.0×10^{-6} mutations/site/generation) (Soares et al. 2009; Lynch 2010) using a generation interval of 25 years (Fenner 2005). To simplify the computation, generations did not overlap, thus not allowing us to model intergenerational marriages. Migration and mating were implemented according to the marriage rules of the given model (see details below). For each simulation, the system

was allowed to reach equilibrium within a single large randomly mating population, before dispersal of structured demes that followed a particular set of marriage rules for 1,000 generations. Simulated data were strongly robust to these initialization parameters (see also Guillot and Cox 2014).

Model System

Although APA has been described by anthropologists in different ways (Needham and Elkin 1973), two integral components are 1) cousin mate prescription and 2) structured migration. Migration was implemented as a wife-giver/wife-taker system, in which a deme always takes wives from the same set of source populations and gives wives to a different set of sink populations (fig. 1). Each deme was permitted up to three wife-giver and three wife-taker clans, although for any given family, the mother's brother's deme is always the wife-giver clan. Mate choice is the prescription for a male to marry his MBD.

APA can be envisaged as a two-parameter system: π_{MBD}, the probability that the MBD rule is followed, and π_{mig}, the probability that a women migrates according to the wife-giver/wife-taker scheme. In its most stringent form ($\pi_{\mathrm{MBD}} = 1$, $\pi_{\mathrm{mig}} = 1$), women always move to their prescribed partner clan and marry their paternal cousin (if one exists). The opposite situation ($\pi_{\mathrm{MBD}} = 0$, $\pi_{\mathrm{mig}} = 0$) represents random patrilocal migration and random mating. Because a suitable cousin may not always be available to marry (e.g., in a family with no children of the required sex), we track the effective (i.e., actual) rate of MBD marriage in addition to π_{MBD}. As Rindi practices polygyny (Lansing et al. 2011), as do most other APA communities, simulations allow up to three wives per male.

For comparison, we also simulate SPA, where any two demes exchange wives at each generation (i.e., each clan acts simultaneously as wife-giver and wife-taker). This model is simulated as for APA, with changes only to the migration scheme.

Summary Statistics

As almost all traditional summary statistics ultimately reflect aspects of the folded site frequency spectrum (Achaz 2009), we use the site frequency spectrum itself as a summary statistic. For historical reasons, we also report several commonly used summaries, such as the site homozygosity H (Nei 1978), Watterson's nucleotide diversity θ_{W}, Tajima's mean pairwise diversity θ_{π}, and the observed number of singleton polymorphisms η_1. Summaries were calculated separately for the autosomes, X chromosome, Y chromosome and mtDNA, as mating systems are expected to affect each of these genetic regions in different ways. The known ascertainment bias of existing SNP chips (an inherent feature of their design; Clark et al. 2005) overrepresents polymorphic sites and underrepresents invariant sites. This bias was addressed by developing unbiased summaries that capture the relative frequencies of polymorphic sites on the autosomes and X chromosome (see supplementary material, Supplementary Material online, for details).

Summary statistics from the simulated data were compared with published observations of unbiased autosomal and X chromosome sequence data to confirm comparability in summary values (Hammer et al. 2008). Length normalized values of the mean pairwise divergence $\widehat{\theta_{\pi}}$ for southern Han Chinese ($\widehat{\theta_{\pi}^{A}} = 8.0 \times 10^{-4}$, $\widehat{\theta_{\pi}^{X}} = 5.8 \times 10^{-4}$) and Melanesian populations ($\widehat{\theta_{\pi}^{A}} = 7.8 \times 10^{-4}$, $\widehat{\theta_{\pi}^{X}} = 6.6 \times 10^{-4}$) are broadly consistent with simulated values ($\widehat{\theta_{\pi}^{A}} = 1.8 \times 10^{-4}$, $\widehat{\theta_{\pi}^{X}} = 1.0 \times 10^{-4}$). We note that the effective population sizes of southern Han Chinese and Melanesians (the geographically and historically closest populations to Rindi for which unbiased autosomal and X chromosome sequence data are available) are likely to be considerably greater than for the small communities that our simulations are intended to mimic. We suggest that the reduced levels of genetic diversity seen in the simulations can be attributed to this lower population size.

General Additive Model Regression

The effects of π_{MBD} and π_{mig} on θ_{π} and H were modeled for the autosomes, X chromosome, Y chromosome, and mtDNA using a general additive model (GAM), which accommodates local and global nonlinear effects (Hastie and Tibshirani 1990). Regressions were fitted to simulated values using the formula $\theta_{\pi} \sim \pi_{\mathrm{MBD}} + s(\pi_{\mathrm{mig}})$, where $s(\pi_{\mathrm{mig}})$ is a smoothing spline function for the migration parameter. GAM regressions were calculated using the R package MGCV (Wood 2011).

Approximate Bayesian Computation

The fit between genomic data from Rindi and simulations was determined using ABC (Beaumont et al. 2010; Sunnåker et al. 2013). This likelihood-free statistical inference method estimates model parameters by comparing outcomes from simulations with real data. Three parameters were inferred: The population size N, π_{MBD}, and π_{mig}. All priors were drawn from continuous uniform distributions with $\pi_{MBD} \in [0, 1]$, $\pi_{mig} \in [0, 1]$, and $N \in [5,000, 12,000]$. From 1×10^5 simulations, 0.1% were accepted using a rejection algorithm (Beaumont et al. 2002). ABC was performed using the R packages abc and abctools (Csillery et al. 2012; Nunes and Prangle 2015).

Different sets of summary statistics were explored and the optimal set selected that returned the lowest prediction error (a measure of distance between estimated and true values for each parameter) (Csillery et al. 2012). The Y chromosome data were discarded due to the limited number of SNPs screened by the HumanOmniExpress chip ($n = 266$) and insensitivity of the Y chromosome to π_{MBD} and π_{mig} observed in initial simulations. Cross-validation was used to confirm the accuracy of the inference method.

Finally, to determine the potential role of larger data sets available in the future, the power of the ABC framework was determined for simulated full sequence data without the ascertainment bias inherent with SNP chips. ABC was performed using the same parameters as described above and

cross-validated over 1,000 simulations to generate prediction errors for N, π_{MBD}, and π_{mig}.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Royal Society of New Zealand through a Rutherford Fellowship (RDF-10-MAU-001) and Marsden Grant (11-MAU-007) to M.P.C. E.G.G. was funded by a doctoral scholarship from the Institute of Fundamental Sciences, Massey University. Computational resources were provided by Massey University and the New Zealand eScience Infrastructure (NeSI). The authors thank Gregory Forth (University of Alberta) and three anonymous reviewers for their constructive comments.

References

Achaz G. 2009. Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183(1):249–258.
Ackerman C. 1964. Structure and statistics: the Purum case. *Am Anthropol.* 66(1):53–65.
Beatty A. 1990. Asymmetric alliance in Nias, Indonesia. *Man* 25(3):454–471.
Beaumont MA, Nielsen R, Robert C, Hey J, Knowles L, Hickerson M, Scott A. 2010. In defence of model-based inference in phylogeography. *Mol Ecol.* 19:436–446.
Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.
Billari F, Fent T, Prskawetz A, Aparicio Diaz B. 2007. The "Wedding-Ring": an agent based model based on social interaction. *Demogr Res.* 17:59–82.
Chaix R, Quintana-Murci L, Hegay T, Hammer MF, Mobasher Z, Austerlitz F, Heyer E. 2007. From social to genetic structures in central Asia. *Curr Biol.* 17(1):43–48.
Clark AG, Hubisz MJ, Bustamante CD, Williamson SH. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* 15:1496–1502.
Cox MP, Karafet TM, Lansing JS, Sudoyo H, Hammer MF. 2010. Autosomal and X-linked single nucleotide polymorphisms reveal a steep Asian-Melanesian ancestry cline in eastern Indonesia and a sex bias in admixture rates. *Proc R Soc Lond B Biol Sci.* 277(1687):1589–1596.
Csillery K, Francois O, Blum MGB. 2012. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol Evol.* 3:475–479.
Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol.* 128:415–423.
Forth GL. 1990. From symmetry to asymmetry. An evolutionary interpretation of eastern Sumbanese relationship terminology. *Anthropos* 85(4):373–392.
Forth GL. 2009. Human beings and other people. *Bijdr Taal Land Volkenkd.* 165(4):493–514.
Forth GL. 1981. Rindi: an ethnographic study of a traditional domain in eastern Sumba. The Hague (The Netherlands): Kininklijk Instituut voor Taal-, Land- en Volkenkude.
Fredlund EV. 1976. Measuring marriage preference. *Ethnology* 15(1):35–45.
Gilbert JP, Hammel EA. 1966. Computer simulation and analysis of problems in kinship and social structure. *Am Anthropol.* 68:71–93.
Guillot EG, Cox MP. 2014. SMARTPOP: inferring the impact of social dynamics on genetic diversity through high speed simulations. *BMC Bioinformatics* 15:175.
Guillot EG, Tumonggor MK, Lansing JS, Sudoyo H, Cox MP. 2013. Climate change influenced female population sizes through time across the Indonesian archipelago. *Hum Biol.* 85(1–3):135–152.
Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD. 2008. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genet.* 4(9):8.
Hastie TJ, Tibshirani RJ. 1990. Generalized additive models. London: CRC Press.
Henrich J, Boyd R, Richerson PJ. 2012. The puzzle of monogamous marriage. *Philos Trans R Soc Lond B Biol Sci.* 367:657–669.
Heyer E, Chaix R, Pavard S, Austerlitz F. 2012. Sex-specific demographic behaviours that shape human genomic variation. *Mol Ecol.* 21(3):597–612.
Huber BR, Danaher WF, Breedlove WL. 2011. New cross-cultural perspectives on marriage transactions. *Cross Cult Res.* 45(4):339–375.
Jacquard A. 1967. La reproduction humaine en régime malthusien. Un modèle de simulation par la méthode de Monte-Carlo. *Population* 22(5):897–920.
Jacquard A. 1970. Panmixie et structure des familles. *Population* 25(1):69–76.
Kingman JFC. 1982. The coalescent. *Stoch Process Appl.* 13(3):235–248.
Kunstadter P, Buhler R, Stephan FF, Westoff CF. 1963. Demographic variability and preferential marriage patterns. *Am J Phys Anthropol.* 21(4):511–519.
Lansing JS. 2006. Perfect order: recognizing complexity in Bali. Princeton (NJ): Princeton University Press.
Lansing JS, Cox MP, De Vet TA, Downey S, Hallmark B, Sudoyo H. 2011. An ongoing Austronesian expansion in Island Southeast Asia. *J Anthropol Archaeol.* 30(3):262–272.
Lansing JS, Watkins JC, Hallmark B, Cox MP, Karafet TM, Sudoyo H, Hammer MF. 2008. Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations. *Proc Natl Acad Sci U S A.* 105(33):11645–11650.
Lévi-Strauss C. 1949. Les Structures Élémentaires de la Parenté. Paris: PUF.
Lévi-Strauss C. 1965. The future of kinship studies. Proceedings of the Royal Anthropological Institute of Great Britain and Ireland, 1965: 13–22.
Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A.* 107(3):961–968.
MacCluer JW, Neel JV, Chagnon NA. 1971. Demographic structure of a primitive population: a simulation. *Am J Phys Anthropol.* 35(2):193–207.
Marlowe FW. 2003. The mating system of foragers in the standard cross-cultural sample. *Cross Cult Res.* 37(3):282–306.
Mascie-Taylor N, Boyce AJ. 1988. Human mating patterns. Cambridge: Cambridge University Press.
Maybury-Lewis DHP. 1965. Prescriptive marriage systems. *Southwest J Anthropol.* 21(3):207–230.
McFarland DD. 1970. Effects of group size on the availability of marriage partners. *Demography* 7(4):475–476.
Moorad JA, Promislow DE, Smith KR, Wade MJ. 2011. Mating system change reduces the strength of sexual selection in an American frontier population of the 19th century. *Evol Hum Behav.* 32(2):147–155.
Needham R. 1964. Descent, category, and alliance in Sirionó society. *Southwest J Anthropol.* 20(3):229–240.
Needham R, Elkin AP. 1973. Prescription. *Oceania* 43(3):166–181.
Nei M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89(3):583–590.
Nunes M, Prangle D. 2015. abctools: tools for ABC analyses. Available from: http://cran.r-project.org/web/packages/abctools/index.html
Ober C, Hyslop T, Hauck WW. 1999. Inbreeding effects on fertility in humans: evidence for reproductive compensation. *Am J Hum Genet.* 64(1):225–231.
Scally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. *Genetics* 13:745–753.

Ségurel L, Martínez-Cruz B, Quintana-Murci L, Balaresque P, Georges M, Hegay T, Aldashev A, Nasyrova F, Jobling MA, Heyer E, et al. 2008. Sex-specific genetic structure and social organization in Central Asia: insights from a multi-locus study. *PLoS Genet.* 4(9):14.

Smouse PE, Vitzthum VJ, Neel JV. 1981. The impact of random and lineal fission on the genetic divergence of small human groups: a case study among the Yanomama. *Genetics* 98:179–197.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet.* 84(6):740–759.

Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C. 2013. Approximate Bayesian computation. *PLoS Comput Biol.* 9(1):e1002803.

Tumonggor MK, Karafet TM, Hallmark B, Lansing JS, Sudoyo H, Hammer MF, Cox MP. 2013. The Indonesian archipelago: an ancient genetic highway linking Asia and the Pacific. *J Hum Genet.* 58(3):165–173.

Van Wouden FAE. 1935. Sociale Structuurtypen in de Groote Oost. Leiden (The Netherlands): Ginsberg.

Verdu P, Becker NSA, Froment A, Georges M, Grugni V, Quintana-Murci L, Hombert J-M, Van der Veen L, Le Bomin S, Bahuchet S, et al. 2013. Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. *Mol Biol Evol.* 30(4):918–937.

Watkins JC. 2004. The role of marriage rules in the structure of genetic relatedness. *Theor Popul Biol.* 66(1):13–24.

Winata S, Arhya IN, Moeljopawiro S, Hinnant JT, Liang Y, Friedman TB, Asher JH Jr. 1995. Congenital non-syndromal autosomal recessive deafness in Bengkala, an isolated Balinese village. *J Med Genet.* 32:336–343.

Winterhalder B, Smith EA. 2000. Analyzing adaptive strategies: human behavioral ecology at twenty-five. *Evol Anthropol.* 9(2):51–72.

Wood SN. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J R Stat Soc B.* 73(1):3–36.

4

RECONSTRUCTING PAST ADMIXTURE PROCESSES FROM LOCAL GENOMIC ANCESTRY USING WAVELET TRANSFORMATION

Sanderson, J., H. Sudoyo, T.M. Karafet, M.F. Hammer and **M.P. Cox**. 2015. Reconstructing past admixture processes from local genomic ancestry using wavelet transformation. *Genetics* 200:469-481.
http://doi.org/10.1534/genetics.115.176842

ABSTRACT Admixture between long-separated populations is a defining feature of the genomes of many species. The mosaic block structure of admixed genomes can provide information about past contact events, including the time and extent of admixture. Here, we describe an improved wavelet-based technique that better characterizes ancestry block structure from observed genomic patterns. Principal components analysis is first applied to genomic data to identify the primary population structure, followed by wavelet decomposition to develop a new characterization of local ancestry information along the chromosomes. For testing purposes, this method is applied to human genome-wide genotype data from Indonesia, as well as virtual genetic data generated using genome-scale sequential coalescent simulations under a wide range of admixture scenarios. Time of admixture is inferred using an approximate Bayesian computation framework, providing robust estimates of both admixture times and their associated levels of uncertainty. Crucially, we demonstrate that this revised wavelet approach, which we have released as the R package *adwave*, provides improved statistical power over existing wavelet-based techniques and can be used to address a broad range of admixture questions.

SUPPLEMENTARY MATERIALS
http://www.genetics.org/content/200/2/469/suppl/DC1

HIGHLIGHTED ARTICLE
GENETICS | INVESTIGATION

Reconstructing Past Admixture Processes from Local Genomic Ancestry Using Wavelet Transformation

Jean Sanderson,*,[1] Herawati Sudoyo,[†] Tatiana M. Karafet,[‡] Michael F. Hammer,[‡,§] and Murray P. Cox*,[2]
*Statistics and Bioinformatics Group, Institute of Fundamental Sciences, Massey University, Palmerston North 4442, New Zealand, [†]Eijkman Institute for Molecular Biology, Jakarta, Indonesia, [‡]Division of Biotechnology, Arizona Research Laboratories and [§]Department of Anthropology, University of Arizona, Tucson, Arizona 85721

ABSTRACT Admixture between long-separated populations is a defining feature of the genomes of many species. The mosaic block structure of admixed genomes can provide information about past contact events, including the time and extent of admixture. Here, we describe an improved wavelet-based technique that better characterizes ancestry block structure from observed genomic patterns. principal components analysis is first applied to genomic data to identify the primary population structure, followed by wavelet decomposition to develop a new characterization of local ancestry information along the chromosomes. For testing purposes, this method is applied to human genome-wide genotype data from Indonesia, as well as virtual genetic data generated using genome-scale sequential coalescent simulations under a wide range of admixture scenarios. Time of admixture is inferred using an approximate Bayesian computation framework, providing robust estimates of both admixture times and their associated levels of uncertainty. Crucially, we demonstrate that this revised wavelet approach, which we have released as the R package *adwave*, provides improved statistical power over existing wavelet-based techniques and can be used to address a broad range of admixture questions.

KEYWORDS wavelets; principal component analysis (PCA); admixture; local ancestry; dating

ADMIXTURE occurs when previously separated populations interact and merge. This process has been instrumental in human history, with most global groups showing at least some signals of population merger (Hellenthal *et al.* 2014). The admixture process produces "mosaic" genomes with alternating blocks of DNA from each ancestral population. Over time, recombination decreases the length of these ancestry blocks, and therefore the distribution of block sizes is informative about the time of admixture. However, the extent to which these patterns can provide additional information about historic admixture processes is still a young area of exploration.

A range of methods have been developed to partition the genome of an admixed individual into ancestry blocks based on raw genomic data (Falush *et al.* 2003; Price *et al.* 2009). Some methods assign ancestry directly. For instance, *HAPMIX* uses a hidden Markov model to estimate the break points of ancestry blocks, while other approaches define ancestry blocks using simple empirical criteria, such as strings of shared *vs.* nonshared polymorphisms (Pool and Nielsen 2009) or the differential presence of population-specific variants (Brown and Pasaniuc 2014). Another set of methods is more indirect. *ROLLOFF* (Moorjani *et al.* 2011), *LAMP* (Baran *et al.* 2012), and *ALDER* (Loh *et al.* 2013) all search for rapid changes in linkage disequilibrium to define the borders of ancestry blocks, while other approaches assign ancestry for predefined genomic windows using conditional random fields (Maples *et al.* 2013) or principal component analysis (PCA) (Gravel 2012).

These methods vary in their effectiveness. Simple empirical criteria perform surprisingly well for admixture between species (as for the mouse admixture zone studied by Pool and Nielsen 2009). Similarly, most of these methods tend to be highly accurate for recent admixture between well-separated human groups (such as African Americans or American Latinos). Indeed, in these settings, subtleties such as multiple waves of admixture have even be detected (Gravel 2012).

doi: 10.1534/genetics.115.176842
Manuscript received October 29, 2014; accepted for publication April 3, 2015; published Early Online April 7, 2015.
Available freely online through the author-supported open access option.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.176842/-/DC1.
[1]Present address: School for Health and Related Research, University of Sheffield, Sheffield, United Kingdom.
[2]Corresponding author: Institute of Fundamental Sciences, Massey University, Private Bag 11 222, Palmerston North 4442, New Zealand. E-mail: m.p.cox@massey.ac.nz.

However, reconstructing complex demographic features for much older admixture events (*i.e.*, thousands rather than hundreds of years in the past) remains extremely challenging (Moorjani *et al.* 2011). While methods have in principle been proposed to detect multiple waves of ancient admixture, in many realistic settings they are still restricted to single admixture events (Loh *et al.* 2013), although some evidence for multiple ancient admixture events has been presented for several Indian populations (Moorjani *et al.* 2011).

Other indirect methods look increasingly promising in this "old admixture" space. Approaches based on principal components analysis and wavelets have been employed with some success. PCA is a nonparametric data-reduction technique, which has been used widely to identify patterns of population structure in genetic data (Patterson *et al.* 2006; Novembre and Stephens 2008; McVean 2009; Bryc *et al.* 2010; Ma and Amos 2012). Dispersion of admixed individuals along the first principal component connecting ancestral populations can be used as a diagnostic for two-way admixture (Patterson *et al.* 2006; Mcvean 2009). For instance, *PCAdmix* employs PCA to assign ancestry to localized windows along the genome for each individual (Brisbin *et al.* 2012). Pugach *et al.* (2011) also use PCA, but do not directly assign ancestry to genomic regions, instead applying a wavelet transform to obtain an indirect measure of the average admixture block length. While this approach has been shown to be powerful for dating old admixture events, there remains considerable scope for (i) the development of more sophisticated wavelet constructions, (ii) examining the resulting wavelet decompositions in greater detail (particularly to identify aspects of non-time-related information in the transformed data), and (iii) to provide a more user-friendly software solution for wavelet analysis.

Wavelet techniques themselves are an active and evolving area, with much potential for novel application in population genetics, as highlighted in the review article by Liò (2003). Wavelets can be thought of as localized, oscillatory functions and are particularly useful for representing data that has local features such as sharp changes and discontinuities. In the context of genome-wide single nucleotide polymorphism (SNP) data, wavelets can be used to represent the mosaic pattern of ancestry blocks. A wavelet decomposition of the data provides information on the size of the ancestry blocks and, importantly, how they are distributed along the chromosomes. Summary measures of the wavelet decomposition allow aspects of the admixture process to be reconstructed, such as the time of admixture and admixture proportions.

Here, we present a substantially revised wavelet-based approach to describe population admixture that builds on the work of Pugach *et al.* (2011). This new method has significantly fewer model assumptions and allows us to identify more complex demographic processes, such as multiple admixture events. As with previous methods, PCA is first employed to describe the population structure. The maximal overlap discrete wavelet transform (MODWT) is then applied directly to the SNP-level data, without the need to compute averages over localized genomic windows as implemented in related procedures (Pugach *et al.* 2011; Brisbin *et al.* 2012). Instead, windowing is performed naturally and objectively as part of the wavelet decomposition procedure. We show that this new method provides robust estimates of admixture time (including improved control of uncertainty estimates), as well as recognizing other aspects of admixture processes that previous wavelet-based methods have not been able to identify with any accuracy.

Methods

General framework

Initially, we consider a simple admixture scenario where two ancestral populations P_A and P_B merged T generations ago to form the admixed population P_C. The ancestral populations contribute to the admixed population with probabilities p and $1-p$. The sizes of the populations, the admixture time, and the admixture proportions are free to vary.

To quantify patterns of genomic block size variation, a three-step analysis procedure was used: (i) PCA was applied to the genomic data to describe population structure; (ii) the wavelet variance was computed to provide a scale-by-scale decomposition of the variance for each population; and (iii) the portion of this measure that is informative for admixture processes was extracted relative to background levels observed in the ancestral populations.

Data simulation

Genome-wide SNP data were simulated using the sequential coalescent simulator *MaCS* (Chen *et al.* 2009). Because our primary interest is in the admixture history of Island Southeast Asia (see *Real genomic data* section below), we chose parameter settings that produce genomic data that broadly fit observed patterns of genetic diversity in this study region (Cox *et al.* 2008). The demographic model, parameters, and information sources are described in more detail in the Supporting Information (Figure S1). We emphasize, however, that the method we describe is general and can be applied to most admixed genomic systems.

Data setup

Given an admixed population P_C derived from two ancestral populations P_A and P_B, the number of individuals in the analysis (*i.e.*, present day samples) is $n = n_A = n_B + n_C$. For each individual i, we observe a collection of T SNPs along a chromosome. Thus the raw data matrix $\boldsymbol{X}$ is a $T \times n$ matrix with T genotype counts in columns and n individuals in rows. The SNPs s are ordered by their physical positions along the chromosome, with the cells of the data matrix $X_{s,i}$ taking the value 0 if heterozygous, and arbitrarily -1 or 1 for the alternative homozygous states. Prior to principal components analysis, the data matrix is centered such that the column mean with respect to the ancestral reference populations is zero, giving

$$X'_{s,i} = X_{s,i} - \frac{1}{n_A + n_B} \sum_{i \in P_E, P_B} X_{s,i}.$$

Principal components analysis

PCA is performed using only individuals from the ancestral populations. Rather than performing PCA on all samples combined, this approach has the advantage that other features of the admixed sample (such as admixture from additional ancestral populations) will not influence the projection (McVean 2009). The first eigenvector v_1 reflects the primary population structure. Projection of individuals onto this axis of variation is given by

$$y_1^i = \sum_{s=1}^{T} X'_{s,i} v_{1,s}. \quad (1)$$

The proportion of ancestry inherited from population P_A can be estimated for each individual (or population) using the distance from the centroids of the ancestral populations; that is, $p_i = (c_B - y_1^i)/(c_B - c_A)$, where $c_A = (1/n_A)\sum_{i \in P_E} y_1^i$ and $c_B = (1/n_B)\sum_{i \in P_B} y_1^i$ are the centroids of the ancestral populations along the first principal axis (Bryc *et al.* 2010). Note that variation between individuals within a population is represented by the smaller eigenvalues and corresponding eigenvectors (Ma and Amos 2010).

This representation of admixed individuals in PCA space, as shown in Figure 1A, provides a genome-wide estimate of average ancestry, but does not indicate how admixture tracts are distributed along the chromosomes. To obtain localized estimates, the projection is performed at the SNP level rather than summing over the length of the genome as in Equation 1. The raw SNP-level admixture signals are given by

$$Y_s^i = \begin{cases} \dfrac{2X'_{s,i} v_{1,s} - (\overline{Y}_s^B + \overline{Y}_s^A)}{(\overline{Y}_s^B - \overline{Y}_s^A)}, & \left|\overline{Y}_s^B - \overline{Y}_s^A\right| \geq \varepsilon \\ 0, & \left|\overline{Y}_s^B - \overline{Y}_s^A\right| < \varepsilon \end{cases}, \quad (2)$$

where $\overline{Y}_s^G = (1/n_G)\sum_{i \in P_G} X'_{s,i} v_{1,s}$ for $G \in A,\ B$. The additional terms in Equation 2 ensure that the signals are normalized such that the mean of the ancestral populations are arbitrarily 1 and −1. This normalization step makes the measure robust to uneven sample sizes, which can affect the structure of the PCA (Novembre and Stephens 2008; Mcvean 2009). Stability of the signals is maintained by specifying a tolerance $\mathcal{E}$ for separating the ancestral populations at a given SNP. This ensures that SNPs with poor discrimination are treated as uninformative in the next step of the analysis.

Wavelet transform

The resulting SNP-level admixture signals indicate how ancestry varies along the genome, but they invariably exhibit a high noise-to-information ratio. To interpret the signal, its frequency content can be described using the wavelet variance (Percival 1995). The wavelet variance S_j for scales $j = 1, \ldots, J$ provides a scale-by-scale decomposition of the variance of the signal. The first scale ($j = 1$) captures the highest frequency patterns, representing very local information. Increasing the scale index provides successively coarser, or lower frequency information, equivalent to "zooming out" on the signal until the level of the entire chromosome is reached. A plot of S_j *vs.* j indicates which scales are important contributors to the process variance and indirectly provides information about the distribution of admixture tracts. For example, recent admixture produces a peak in the wavelet variance at a large wavelet scale, reflecting long admixture tracts, while more ancient admixture events produce peaks at lower wavelet scales, reflecting shorter admixture tracts.

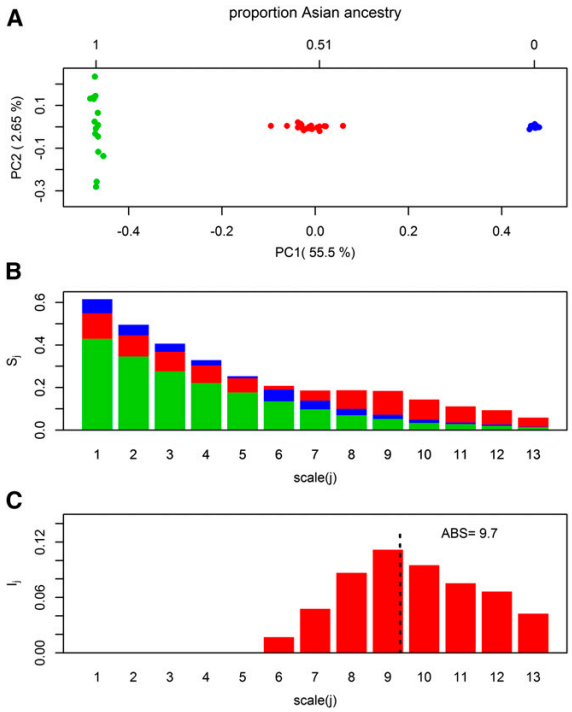

Figure 1 Simulated example with 13,000 SNPs, 15 diploid individuals in ancestral populations (P_A, P_B), and 20 diploid individuals in the admixed population (P_C). Populations are shown in green (P_A), blue (P_B), and red (P_C). (A) PCA is used to describe the primary population structure; (B) raw wavelet variance for each population illustrates high frequency noise; (C) informative variation in the admixed population after standard correction for noise estimated from the ancestral populations. Note that this example uses the default threshold $\mu = 1$.

The wavelet variance for an individual i is given by

$$S_j^i = \frac{1}{T} \sum_{k=1}^{T} |d_{j,k}^i|^2, \quad (3)$$

where $d_{j,k}^i = \sum_s Y_s^i \psi_{j,s-k}$ are the wavelet coefficients for the signal Y^i constructed using the wavelet system ψ. To

appreciate the methodology, it is sufficient to understand that the wavelet variance reflects the frequency content of the signal, but more detailed background material is provided in the Supporting Information (Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, and File S1). Our implementation employs Daubechies' least asymmetric wavelet number 8 in the *waveslim* (Whitcher 2013) package of the statistical software R (R Development Core Team 2014). We emphasize, however, that the methods proposed here are robust to other choices of analyzing wavelet (see Supporting Information, Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, and File S1).

Population averages are computed as

$$S_j^C = \frac{1}{n_C} \sum_{i \in P_C} S_j^i$$

(and similarly for populations P_A and P_B). An example of the average wavelet variance for each population is shown in Figure 1B. The wavelet variance is highest at fine scales, but as the ancestral populations also show this pattern, it should be considered background noise. It is intuitive that the very finest wavelet scales are uninformative because small numbers of SNPs should be insufficient to differentiate between populations. The raw wavelet variance is therefore considered as a combination of informative variation and background noise

$$S_j^C = I_j^C + N_j. \tag{4}$$

To extract the informative variance I_j^C, we subtract the proportion that can be attributed to noise. This is estimated from the variation observed in the ancestral populations; $\hat{N}_j = \mu \cdot \max(\overline{S}_j^A, \overline{S}_j^B)$, where μ is a multiplicative factor that allows the degree of thresholding to be controlled. Under almost all conditions, a default value of $\mu = 1$ may be assumed, and this threshold should be raised only if the admixture signals exhibit high levels of noise (see Supporting Information, Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, and File S1 for details). Population characteristics that influence noise levels in the admixture signals are explored in the next section. The final measure of the informative variance is given by

$$I_J^C = \max\left(S_J^C - \hat{N}_j, 0\right), \tag{5}$$

which describes the frequency content that is unique to the admixed population (in contrast to the ancestral populations).

Real genomic data

To illustrate that our method performs well in real-world situations, it was applied to a SNP genotyping chip data set of 394 individuals from 16 communities spread across the Indonesian archipelago (Table 1). Equivalent SNP data from Southern Han Chinese and Papua New Guinea Highlanders were used as proxies for the ancestral populations. Permission to conduct research in Indonesia was granted by the Indonesian Institute of Sciences. Blood samples or buccal swabs were collected from consenting, closely unrelated, and seemingly healthy individuals by J. Stephen Lansing (University of Arizona) and Herawati Sudoyo (Eijkman Institute for Molecular Biology, Indonesia), with the assistance of Indonesian Public Health clinic staff. All sample collection followed protocols for the protection of human subjects established by both the Eijkman Institute and the University of Arizona institutional review boards. Participant interviews confirmed local residence for at least two generations into the past. Samples were genotyped with the Affymetrix Axiom chip, yielding 548,994 SNPs across the autosomes. (Sex-linked markers were excluded from the analysis.) The SNP data were cleaned using standard protocols in PLINK v. 1.07 (Purcell *et al.* 2007; Purcell 2009) and the wavelet transform performed as described above.

The approximate Bayesian computation analysis employed 1000 data sets with sample sizes and SNP numbers set to those of the real data. These data sets were simulated by drawing from a uniform prior of admixture times between 10 and 300 generations. The admixture proportion for the Bena population (used as our primary test case) was set to 0.6, as estimated previously from the real data. The ABS metric was calculated for each simulation, and the multiple chromosome structure of the data were mimicked by sampling each individual repeatedly with different data densities.

Results

As proof of concept, we first applied our wavelet method to simulated data. A range of admixture scenarios was explored by varying parameters of the demographic model, particularly the time of admixture, admixture proportion, and single *vs.* multiple admixture events. Fifty simulations were performed for each scenario with modest (but therefore realistic) ancestral sample sizes of $n_A = n_B = 15$ and an admixed sample size $n_C = 20$.

Admixture time

Because the ability of wavelet methods to calculate the time of admixture is well known from earlier work (Pugach *et al.* 2011), we explored this feature first. Simulations were performed for admixture times ranging from 10 to 320 generations (*i.e.*, from the recent past to ~10,000 years ago, using a generation interval of 30 years; Fenner 2005). Admixture at 10 generations shows the highest informative wavelet variance at scale 13, reflecting relatively few, long admixture blocks (Figure 2). As the time of admixture occurs further back in the past, the peak in wavelet variance shifts toward successively lower wavelet scales, reflecting ever-smaller admixture blocks driven by cumulative recombination along the chromosome. The average frequency content can be characterized by the average block size metric ABS, termed the "wavelet center" by Pugach *et al.* (2011), which as shown later, can be used to date the admixture event

Table 1 Summary of case study populations describing sample size (*n*), proportion of Asian ancestry as inferred by PCA (*p*), and the average block size metric (ABS, for admixed populations only)

Population	*n*	*p*	Average block size metric (ABS)
Southern Han Chinese	13	1.00	–
Nias	28	0.87	4.26
Mentawai	29	0.87	4.30
Java	21	0.84	4.41
Sumatra	30	0.83	4.49
Bali	19	0.83	4.90
Sulawesi	21	0.80	6.57
Sumba, Wunga	30	0.67	7.90
Sumba, Anakalang	30	0.66	7.83
Flores, Rampasasa	12	0.66	8.05
Flores, Bena	30	0.57	8.14
Flores, Bama	30	0.55	8.34
Timor, Umanen Lawalu	17	0.55	8.44
Timor, Kamanasa	19	0.53	8.42
Lembata	28	0.53	8.39
Pantar	27	0.45	8.47
Alor	23	0.42	8.46
Papua New Guinea Highlands	13	0.00	–

$$\mathrm{ABS} = \frac{\sum_j j \cdot \bar{I}_j^{\mathrm{AB}}}{\sum_j \bar{I}_j^{\mathrm{AB}}}. \qquad (6)$$

Admixture proportion

Admixture proportions were varied between 0.5 (equal ancestry from P_A and P_B) and 0.025 (ancestry predominately from P_A). For this analysis, the time of admixture was fixed at 160 generations. As the proportion of admixture decreases, the raw wavelet variance exhibits increasing levels of noise relative to informative variation. This is shown by the reduced magnitude of the informative wavelet variance (Figure 3) and emphasizes that, as expected, it is increasingly difficult to extract informative variation at low admixture proportions (small p) even where the signal is technically present. In this example, informative estimates were obtained for admixture proportions as low as 2.5%, although in general, the range of p for which this method is applicable will also depend on other characteristics of the data, such as the SNP density and sample size, as considered in the next section.

Sensitivity analyses

The sensitivity of the method to a wide range of data characteristics was considered by repeating the results of the admixture time example with a large number of simulated data sets. Results are summarized in Table 2 and Figure 4.

Condition 1 shows the original results, exactly as described above. New simulations were then performed to mimic realistic linkage disequilibrium (LD) (condition 2). To do so as accurately as possible, we applied the real recombination rates observed along the first 100 Mb of chromosome 1, as recombination rates for chromosome 1 are near the average of all chromosome-level recombination rates (Figure S2). The effect of lower sample size (condition 3) was investigated by reducing the number of individuals sampled from each population by 5, thus yielding sample sizes that would be smaller than almost any published population genetics study (n_A, n_B = 10, n_C = 15). The effect of more recent divergence between the ancestral populations (condition 4) was investigated by decreasing $T_{\mathrm{Ancestral}}$ from 2000 to 1200 generations ago (50,000–30,000 years ago). The effect of using a misrepresentative modern population as a proxy for an ancestral population (condition 5) was investigated by studying ancestral populations with mixed (rather than "pure") ancestry. Rather than using samples from the true ancestral population P_A, an admixed ancestral population P_A^* was employed instead ($p = 0.1$). A wide range of parameters was applied for sensitivity testing, but for clarity, only results for single parameter values are shown on Figure 4. These examples are representative of all the tests that were run.

Variation in summary measures between simulations was compared by computing the relative standard deviation (RSD) at each admixture time. For all of the error conditions above, the computed ABS metrics are consistent with the reference case (condition 1), but with slightly larger relative standard deviations. Only for one case (condition 4; reduced divergence between the ancestral populations and admixture at 320 generations) are the ABS metrics biased, with the mean falling outside the range of values observed for the reference example. We emphasize that this is expected: admixture should be more difficult to detect when it occurs between two ancestral populations that diverged only recently. Stability of the ABS metrics in this particular scenario could be improved by applying a higher level of thresholding. However, the default value of $\mu = 1$ was retained here to provide consistency across scenarios, to demonstrate the deterioration in resolution, and to illustrate that the thresholding parameter can be ignored for all but the most extreme admixture cases.

In all of these examples, including the standard reference case, the localized admixture signals provide a noisy indication of how ancestry varies along the chromosome. Indeed, the inherent stochasticity of the block structure is the primary reason why other sources of variance, such as the cases discussed above, have relatively little additional effect on the overall results. This noise is addressed using wavelets to capture the distribution of block sizes, coupled with a correction based on the ancestral populations to distinguish informative signals from background variation. The cases considered above all slightly increase noise levels relative to informative variation, which, as demonstrated by the admixture proportion example in Figure 3, reduces the magnitude of the extracted informative wavelet variance. As noise increases, it naturally becomes more difficult to extract informative variation. However, this increase in noise levels is minimal for all but the most extreme confounds, thus allowing the technique to be applied robustly to a very wide range of scenarios.

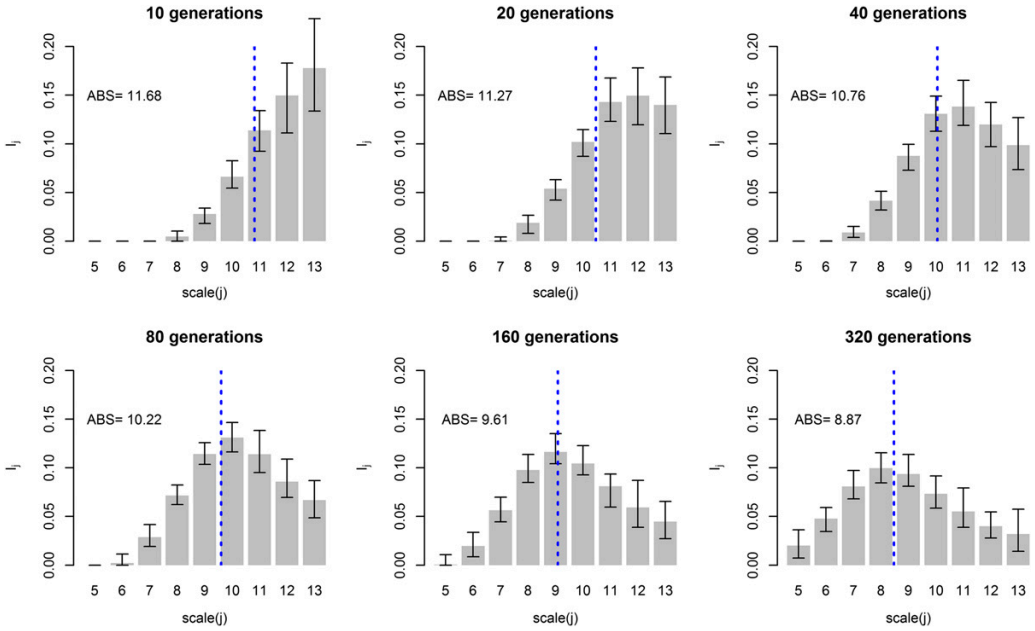

Figure 2 Informative wavelet variance for each time of admixture (10–320 generations using default thresholding $\mu = 1$). Shaded bars represent the average over 50 simulations at each admixture time; black bars represent the range across individual simulations. The average block size metric for each scenario is indicated by a dotted blue line.

The effect of SNP density (which is always a known variable) is demonstrated by down sampling the data (conditions 6–8). The original density of 4306 SNPs (condition 6) was chosen to correspond to the size of our real chromosome 22 data set. Reducing the SNP density of this data set means that the resulting wavelet decomposition is given over 11 wavelet scales rather than the earlier 13, and so as expected, the computed mean ABS metrics are correspondingly much smaller. However, this has no effect on the inference, as the data size is always known and simulations are simply run to match the size of the observed data. Further reductions in SNP density to 3250 SNPs (condition 7) and 1625 SNPs (condition 8) are also shown. Note that although the absolute values of the ABS metrics are shifted, the trend with admixture time remains consistent.

Method comparison

The original *StepPCO* method (Pugach *et al.* 2011) has already been tested extensively against other admixture detection methods, particularly *HAPMIX* (Price *et al.* 2009). We therefore focus here on comparing our improved wavelet method against the *StepPCO* procedure. Figure 5 shows that the summary measure (wavelet center) used in *StepPCO* is comparable to the *adwave* ABS metrics, as both exhibit a strong trend with time of admixture. However, the dispersion is consistently smaller for the *adwave* ABS metrics. For example, the wavelet centers (*StepPCO*) computed for $T = 320$ and $T = 160$ show substantial overlap, while the ABS metrics (*adwave*) for the same populations show only minimal overlap. This illustrates that *adwave* offers increased power to differentiate between older admixture scenarios, with substantially reduced uncertainty in dating.

We also emphasize that *adwave* requires far fewer user specifications with regard to runtime options. The only variable for *adwave* is the thresholding parameter, and as shown above, the default value of $\mu = 1$ should be used for almost all admixture scenarios. In contrast, the *StepPCO* results required a signal length parameter ($K = 1024$), a window size parameter ($\lambda = 5$), and two thresholding parameters (threshold = 0.1, maxlevel = 6) (all notations from Pugach *et al.* 2011). A detailed demonstration of this method comparison, with explanation of the settings chosen for *StepPCO*, is provided in Figure S3.

Admixture in Indonesian populations

Populations across Indonesia show genomic admixture between Asian and Melanesian ancestral sources (Cox *et al.* 2010), which has been dated using other methods to an admixture event ~4000 years ago (~130 generations) (Xu *et al.* 2012). We calculated wavelet summary measures for 16 communities across the Indonesian archipelago using 548,994 autosomal SNPs screened in 394 individuals (Table 1). Equivalent data from Southern Han Chinese and Papua New Guinea Highlanders was used as proxies for ancestral populations, as described in Cox *et al.* (2010).

The PCA for all individuals, where only the ancestral populations were used to define the axes, is shown in Figure 6. Admixed individuals dispersed along the first principal

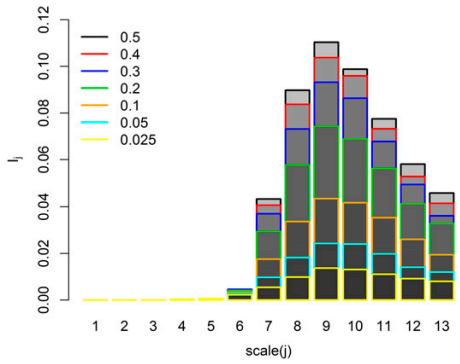

Figure 3 Relationship between proportion of admixture and informative wavelet variance. For this example only, a nondefault value for the threshold $\mu = 1.1$ was used to account for increased noise in the admixture signals due to low proportions of admixture, as described in the text. The magnitude of the wavelet variance decreases with the admixture proportion, shown as colored bars from black ($P = 0.50$) to yellow ($P = 0.025$).

component illustrate the primary genomic signal, a strong gradient in Asian-Melanesian ancestry that has previously been observed across the region (Cox *et al.* 2010). The informative wavelet variance was computed separately for each chromosome and individual and subsequently combined to provide a single measure for each population (Figure S4). To combine information across chromosomes, which vary considerably in size, the raw admixture signals were windowed: all signals were reduced to the size of the smallest chromosome (importantly without discarding any data) by computing averages over a window of SNPs (details of the windowing procedure are provided in Supporting Information, Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, and File S1). The SNP density and window size for each chromosome are shown in Table S1. This windowing procedure is used only to standardize chromosomes to the same length and utilizes very short windows of SNPs (unlike the approach of Pugach *et al.* 2011).

The average block size metrics calculated for each population are shown in Table 1. The first six Indonesian populations (Nias, Mentawai, Java, Sumatra, Bali, and Sulawesi) exhibit predominantly Asian ancestry, with high-frequency noise in the signals causing some bias in the computed ABS metrics (Figure S4). The remaining Indonesian populations exhibit less extreme Asian ancestry proportions (42–67%), with the resulting ABS metrics appearing broadly similar between populations.

Under the assumption of a single admixture time (relaxed in later sections), the average block size metric can be used to date the time of admixture using approximate Bayesian computation (ABC). A general introduction to ABC can be found in Csilléry *et al.* (2010) and Sunnåker *et al.* (2013), while ABC in the context of parameter estimation for population admixture has been considered by Sousa *et al.* (2009) and Robinson *et al.* (2014).

The ABC inference procedure allows us to capture uncertainty in admixture time estimates more robustly than earlier wavelet dating approaches (Pugach *et al.* 2011; Xu *et al.* 2012). To illustrate this process, dating was performed on the Bena population of Flores in eastern Indonesia, resulting in an estimated median admixture time of 147 generations (95% credible region: 122–178 generations), or 4410 years before present (95% CR: 3660–5340 years BP). This almost exactly matches earlier point estimates of the admixture time (Xu *et al.* 2012) and is consistent with our current understanding of Island Southeast Asian prehistory (Bellwood 2007).

The relationship between time of admixture and the ABS metric across all simulations is illustrated in Figure 7A. ABC was implemented using the R package *abc* (Csilléry *et al.* 2012), and the posterior distribution of admixture time was computed using a local linear regression (Beaumont *et al.* 2002) with a tolerance rate of 0.2. Cross validation was used to evaluate the accuracy of this estimate: the prediction error was low (0.038) and insensitive to the exact tolerance value. For future research focusing on parameter inference, this procedure could be modified to use a larger number of simulated data sets and a lower tolerance rate. However, this simple example clearly illustrates that the *adwave* method has good statistical power to date admixture using a relatively small number of simulations.

Multiple admixture events

Another aim of this work is to show that our improved wavelet approach can be used to study other features of the admixture process beyond the well-explored question of admixture time. In the examples covered thus far, it has been assumed that admixture occurred as a single event. However, additional waves of admixture will result in the introduction of new ancestry tracts, replacing a proportion of older, shorter ancestry blocks with newer, longer ones. Pugach *et al.* (2011) briefly considered the effects of continuous admixture within a wavelet setting, showing that this leads to underestimated admixture times in their original methodological framework. In contrast, we instead consider scenarios with two distinct admixture events. We show that this process creates distinctive patterns in the observed informative variation, which can be used to reconstruct more complex demographic processes (as opposed to being treated solely as a potential source of bias).

In the following dual-admixture scenarios, the first admixture event always occurs at 160 generations. To investigate the effect of separation between admixture events, the second admixture event varies between 10 and 80 generations. In the extreme case of admixture at 160 and 10 generations ago, the localized admixture signals contain two dominant frequencies. Single admixture events at 160 and 10 generations lead to peaks in the informative wavelet variance at wavelet scales of 9 and 13, respectively. When two admixture events occur, the informative wavelet variance is instead spread between these scales (Figure 8A). As the admixture events occur closer

Table 2 Sensitivity of the *adwave* method to a range of data limitations

Data limitations		Admixture time (generations)					
Condition	Description	10	20	40	80	160	320
1	Reference	11.55 (0.69)	11.14 (0.58)	10.65 (0.7)	10.13 (0.66)	9.54 (1.12)	8.84 (1.72)
2	Realistic LD	11.71 (0.84)	11.31 (0.88)	10.85 (1.08)	10.32 (1.04)	9.82 (1.23)	9.09 (2.28)
3	Reduced sample size	11.66 (0.90)	11.25 (0.63)	10.74 (0.88)	10.20 (0.85)	9.58 (1.19)	8.85 (1.78)
4	Reduced divergence between ancestral populations	11.64 (0.83)	11.24 (0.69)	10.76 (0.82)	10.21 (1.02)	9.49 (1.35)	8.32 (3.59)
5	Non-representative ancestral populations	11.73 (1.34)	11.23 (1.42)	10.77 (1.51)	10.21 (1.45)	9.58 (1.94)	8.60 (3.06)
6	SNP density $T = 4036$	10.47 (0.91)	10.04 (0.73)	9.55 (0.92)	8.98 (1.22)	8.37 (1.62)	7.60 (3.13)
7	SNP density $T = 3250$	9.96 (0.76)	9.60 (0.60)	9.19 (0.76)	8.68 (1.08)	8.12 (1.41)	7.37 (2.6)
8	SNP density $T = 1625$	8.99 (1.23)	8.63 (1.09)	8.20 (1.50)	7.68 (2.29)	7.10 (3.15)	6.33 (5.86)

Mean average block size values (relative standard deviation in parentheses) are shown for each admixture time. Reference data were simulated with $T = 13{,}000$ SNPs, populations sizes of n_A, $n_B = 15$, $n_C = 15$, and divergence between the ancestral populations at $T_{Ancestral} = 2000$ generations ago.

together, this spread in the observed informative wavelet variance decreases (Figures 8, B–D).

For one admixture event, a single dominant peak is observed in the informative wavelet variance, and the ABS metric therefore provides a convenient summary measure. For multiple admixture events, the ABS metric describes the average admixture time, but provides no information about the duration over which admixture occurred. In contrast, the informative wavelet variance should provide additional information about the peak dispersion. To explore the potential for identifying more complex admixture scenarios, a simple classification rule was implemented. An admixed population P_C is assigned to one of two groups G_1, G_2, which are characterized by the summary measures M_1, M_2. This scheme is described with abstract choice of summary measure, but below, we consider how different summary measures (taking M_1, M_2 to be either the ABS metric or wavelet informative variance) affect the success of classification.

The classification rule is implemented as follows:

1. The "true" summary measures M_1, M_2 are computed for each group using values obtained from the first 25 simulations.
2. For each of the remaining 25 trial data sets ($s = 1,\ldots,25$), estimated summary statistics $\widehat{M_s}$ are calculated. The divergence measures are defined as

$$D_i = \sum_{s=1}^{S} \left| \widehat{M_s} - M_i \right|, \qquad (7)$$

for $i = 1, 2$.

3. If $D_1 < D_2$, classify to G_1; otherwise classify to G_2.

The classification rates are shown in Table 3 for scenario 1 (a single admixture event at 60 generations; mean ABS 10.47, range 10.30–10.64) and scenario 2 (two admixture events at 160 and 10 generations; mean ABS 10.57, range 10.35–10.90). With a sample size of just 10 individuals for the admixed population, perfect classification is achieved using the informative wavelet variance, while the ABS metric correctly classifies only 60% of cases. For real multiple admixture situations, this classification framework could be extended to a more complex inferential setting (such as ABC), but this simple example demonstrates the potential for reconstructing complex admixture scenarios from the full wavelet variance profile.

Discussion

Wavelet techniques provide information on the ancestry block structure of admixed genomes and hence can be used to reconstruct the processes involved in past admixture events. Ancestry blocks are strictly unobservable and can be inferred only from the data. Wavelets provide indirect information on the block structure, thus providing an alternative over methods that assign ancestry directly (Sankararaman *et al.* 2008; Price *et al.* 2009). A growing body of methods now assign ancestry indirectly using various unrelated approaches (Moorjani *et al.* 2011; Baran *et al.* 2012; Gravel 2012; Loh *et al.* 2013; Maples *et al.* 2013; Brown and Pasaniuc 2014), but here we extend the use of wavelet techniques as introduced by Pugach *et al.* (2011). Importantly, our implementation differs markedly from the original *StepPCO* program, with the main differences at each stage of the analysis highlighted below:

- Localized admixture signal formation: StepPCO (Pugach *et al.* 2011) uses large windows of SNPs to produce an averaged admixture signal in localized windows along the genome. Our work demonstrates that wavelet methods are equally applicable to the raw unwindowed signals, with the windowing procedure performed intrinsically as part of the wavelet analysis, and therefore not requiring arbitrary *a priori* decisions on window size.
- Wavelet analysis: The wavelet methods we describe are based on the MODWT, which offers more flexibility in its application since there is no restriction on the length of the signals. Conversely, StepPCO employs the discrete wavelet transform (DWT), which has the strict requirement that

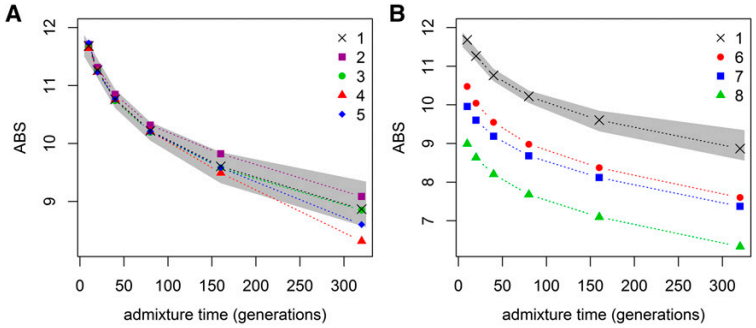

Figure 4 Sensitivity to a range of realistic data limitations. Comparison to reference data (condition 1) simulated with T = 13,000 SNPs, populations sizes $n_A, n_B = 15$, $n_C = 15$, and ancestral population divergence at $T_{Ancestral} = 2000$ generations. The gray area shows the range of ABS metrics observed under the standard reference condition. (A) Potential sources of error (conditions 2–5); (B) varying SNP densities (conditions 6–8). Note that the decline in absolute values of the ABS metrics in B is expected; these are easily accounted for in an inference setting because the SNP density is always a known variable. Condition descriptions and numeric values are presented in Table 2.

signals be of length 2n. Data must therefore be windowed, or discarded, to meet the restrictive length requirements of the DWT framework. Another advantage of the MODWT is that the resulting wavelet coefficients are translation equivariant, meaning that circularly shifting the data results in the same shifting of the coefficients. Said differently, changing the starting point—for instance, to avoid a poor quality SNP—does not affect the resulting wavelet coefficients, whereas this is not true under the DWT framework. This property is particularly important if the results are to be used for specific localized genomic regions (as discussed briefly below) and thus provides a solid statistical foundation for future work.

- Extraction of relevant information: The portion of the resulting wavelet decomposition that is informative about the admixture process is extracted in a simple procedure with reference to the ancestral populations, offering greater simplicity and objectivity than the multistage thresholding procedure described by Pugach *et al.* (2011).
- Software: The adwave software, which implements the method described in this article, is an official package in the R project (http://cran.r-project.org/web/packages/adwave/index.html). This allows extremely easy installation and use, as well as providing a series of simple worked examples as a learning exercise. The adwave package is also faster than the existing StepPCO code and offers more flexibility in the choice of analyzing wavelet (unlike StepPCO, which employs only the simplest "square-shaped" Haar wavelet).

The work presented here also makes several other advances. The average block size metric has previously been shown to capture the time of admixture. Here, we have implemented a more formal dating procedure using ABC under the assumption of a single admixture event. In reality, populations may have experienced multiple admixture events leading to complex patterns of genetic variation. We have shown that the wavelet variance contains additional information to identify these more complex admixture scenarios. This highlights the potential of wavelet-based techniques to be coupled with formal statistical inference procedures to robustly distinguish between the range of scenarios that could have resulted in any observed genetic pattern.

Method performance for the *StepPCO* procedure has already been tested against other admixture detection methods, most extensively with *HAPMIX* (Price *et al.* 2009), with favorable results. This is especially true for older admixture events (Pugach *et al.* 2011). While an in-depth comparison with other local ancestry detection methods would be of great interest (Moorjani *et al.* 2011; Baran *et al.* 2012; Gravel 2012; Loh *et al.* 2013; Maples *et al.* 2013; Brown and Pasaniuc 2014), such an analysis is beyond the scope of this manuscript. We have therefore focused instead on showing how *adwave* markedly improves on the original wavelet method implemented in *StepPCO*. As shown above, *adwave* offers improved statistical power to differentiate between admixture scenarios, offers much reduced uncertainty in model parameter estimates, and importantly, is far easier to use than *StepPCO*, especially by requiring far fewer user-specified runtime parameters.

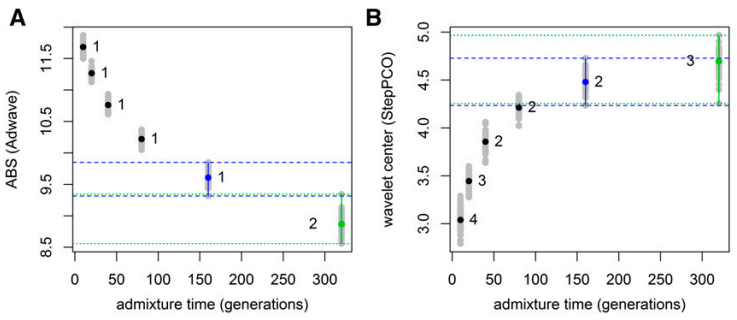

Figure 5 Comparing *StepPCO* and *adwave* showing the relationship between wavelet transform summaries and time of admixture. (A) *Adwave* using $\mu = 1$; (B) *StepPCO* using $K = 1024$, $\lambda = 5$, threshold = 0.1, and maxlevel = 6. Numbers indicate the relative standard deviation (RSD, %) for each admixture time. Note the difference in discrimination power between the two methods for older admixture events (95% confidence intervals as dashed blue and green horizontal lines).

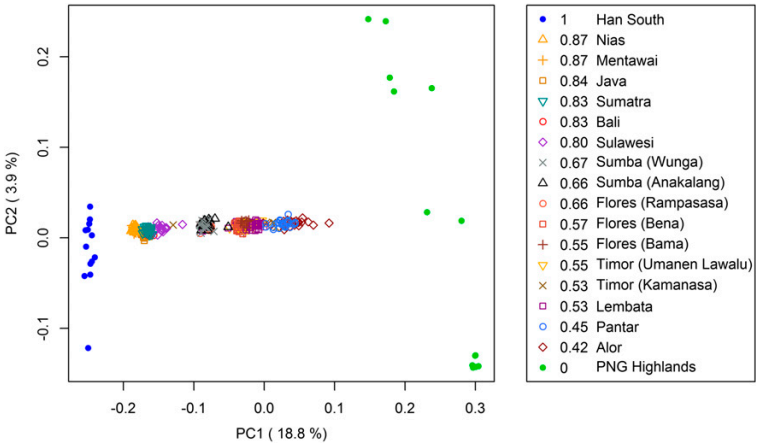

Figure 6 PCA of autosomal SNP data from Indonesian populations, with Southern Han Chinese (blue circles) and Papua New Guinea Highlanders (green circles) employed as proxy ancestral populations. Numbers give calculated admixture proportions.

In the future, considering the full wavelet periodogram, rather than the genome-wide summary measures used in both *adwave* and *StepPCO*, may yield promising results wherever the distribution of ancestry tracts along the genome is substantially nonstationary. Bryc *et al.* (2010) use their formulation of localized admixture signals to address whether regions of the genome show predominant ancestry from a given population. Wavelets are well suited to distinguishing local features in data and could be helpful in this regard, identifying features that may not be easily detected by considering the localized admixture signals in their raw form.

Other prospective areas for further work include the extension of these methods to the more general case of multipopulation admixture. Ma and Amos (2012) describe the use of PCA as a diagnostic in this setting, and PCA has been used to assign multipopulation ancestry in the software *PCAdmix* (Brisbin *et al.* 2012). The wavelet methods described here could be extended in a similar way by considering pairwise combinations of any number of ancestral populations.

In contrast, key restrictions that determine our ability to reconstruct admixture events include the degree of differentiation between the ancestral populations and the representativeness of samples used as surrogate ancestral groups. As the ancestral populations become more similar or the surrogate populations become more different from the true ancestral populations, the localized admixture signals become increasingly noisy. Although this ultimately leads to a loss of identifiability in extreme cases, the method is remarkably robust to moderate deviations from these assumptions. As shown above for low admixture proportions, through judicial choice of the thresholding parameter even extremely noisy data can still provide meaningful estimates (the only situation in which we encourage deviation from the default setting).

Sample size (both in terms of SNPs and individuals) is also important and affects the PCA step of the procedure. The purpose of the PCA step is to summarize the overall variability among individuals, which includes both between-population and within-population variability. In reconstructing population ancestry, we aim to describe between-population variation, while ignoring within-population variation. This is achieved by selecting the first principal component, as long as the sample sizes are sufficiently large. Within-population fluctuations of individual coordinates on the PCA scatterplot can be caused by subtle population substructure. Assuming that no such substructure is present, these fluctuations decrease as the total sample size increases, and an asymptotically stable pattern of the eigenvector plot results (Ma and

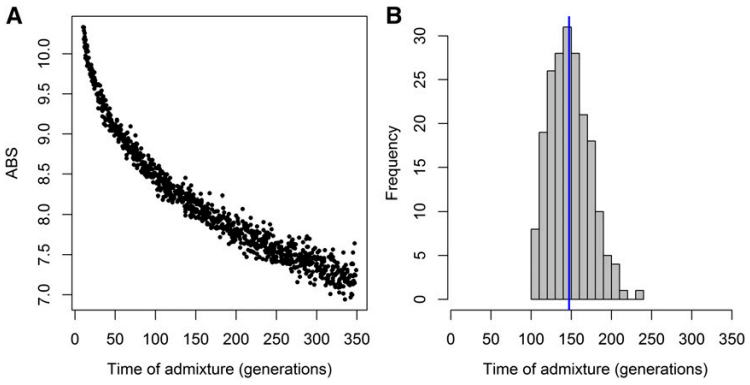

Figure 7 Dating time of admixture for Bena (Flores, eastern Indonesia) using approximate Bayesian computation. (A) Relationship between admixture time and average block size metric for all simulations; (B) weighted posterior distribution of admixture time. Median estimated time of admixture, indicated by the blue line, is 147 generations (95% credible region: 122–178 generations).

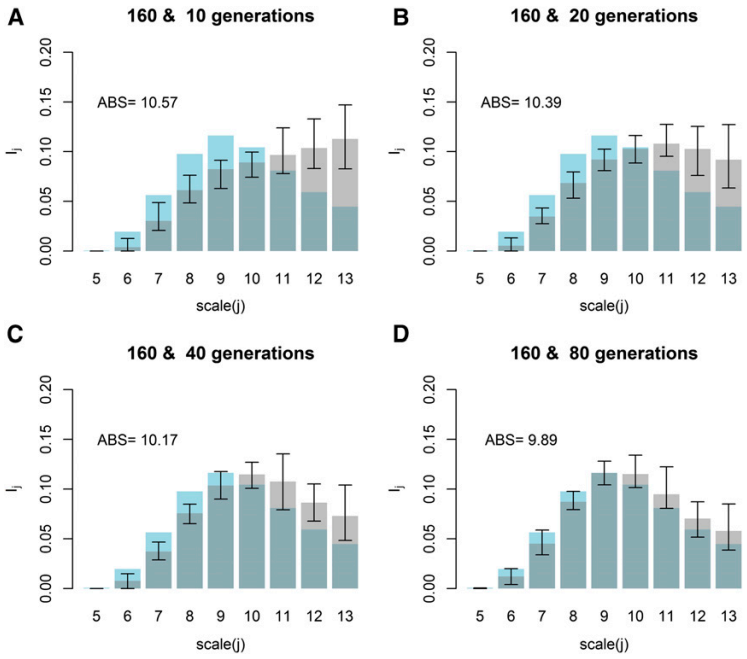

Figure 8 Dual admixture events at 160 and 10–80 generations. Gray bars represent the average over 50 simulations for each scenario; black bars represent the range for individual simulations. Blue bars show the average informative wavelet variance for a single admixture event at 160 generations, providing a reference point for comparison.

Amos 2010). When the number of individuals is large, variation between individuals from the same population is small compared to that of the different populations, so that the first eigenvector describes the primary population structure of the data. However, as the sample size decreases, individual variation carries more weight, which may be addressed in more than the first principal component. Note that the use of methods other than PCA may be helpful in this regard. For example, Jombart *et al.* (2010) introduced discriminant analysis of principal components to achieve separation of individuals into predefined groups. In practice, as long as the sizes of the ancestral population samples is sufficiently large, discriminant analysis provides the same result as PCA (unpublished data). The two methods may, however, perform differently for small sample sizes.

How far back in time admixture processes can be reliably identified is strongly influenced by the number of genotyped SNPs. The relationship between the number of admixture blocks, time of admixture and wavelet scale is summarized in Table 4. The shaded column indicates the findings described in the *Results*, using simulated data sets of 13,000 SNPs (chosen for a region ~100 Mb in length, comparable to the SNP content of our 100 Mb chromosome 15 data set). For admixture at 10 generations, the informative wavelet variance is highest at scale 13, reflecting a small number of large admixture blocks. As the time of admixture increases, the peak shifts toward lower scales, reflecting a larger number of smaller admixture blocks. This pattern is illustrated for admixture up to 320 generations (~10,000 years), but importantly, it is possible to reconstruct even older admixture events. The highest frequency (relating to the smallest admixture blocks) that can be detected, as determined purely by the data density, is termed the Nyquist frequency (Chatfield 2003). However, resolution power is likely to deteriorate well before this point and will be strongly influenced by the degree of differentiation between the ancestral populations. The more closely related the ancestral populations, the less well they can be discriminated using only a small number of SNPs. Increasing the SNP density allows detection of higher frequency information, relating to shorter (more ancient) admixture tracts. To illustrate this, the mapping to wavelet scale is illustrated for a hypothetical twofold and fourfold increase in the number of genotyped SNPs (26,000 and 104,000 SNPs, respectively). As genetic data sets improve (particularly through whole-genome sequencing), wavelet methods will therefore

Table 3 Classification rate for the summary measures average block size and informative wavelet variance with increasing sample size ($1 \le n_C \le 10$)

	Correct classification (%)	
Sample size (individuals)	Wavelet variance	Average block size
1	76	56
2	84	58
3	89	59
4	92	60
5	94	61
6	96	61
7	97	62
8	98	62
9	99	61
10	100	60

Table 4 Relationship between the number of admixture blocks, time of admixture, and wavelet scale

		Wavelet scale (no. of SNPs)		
Admixture blocks	Time of admixture (generations)	13,000	26,000	104,000
8,192–16,384	163,840	—	—	1
4,096–8,192	81,920	—	1	2
2,048–4,096	40,960	1	2	3
1,024–2,048	20,480	2	3	4
512–1,024	10,240	3	4	5
256–512	5,120	4	5	6
128–256	2,560	5	6	7
64–128	1,280	6	7	8
32–64	640	7	8	9
16–32	320	8	9	10
8–16	160	9	10	11
4–8	80	10	11	12
2–4	40	11	12	13
1–2	20	12	13	14
0–1	10	13	14	15

The dominant admixture block size decreases with time since admixture, while conversely, the number of admixture blocks increases. Underlined numbers are from the example presented in the *Results* section (13,000 SNPs from a genomic region ~100 Mb in length, comparable to the data set for chromosome 15). Columns to the right show how mapping to wavelet scale depends heavily on SNP density: increasing the number of SNPs two- and fourfold allows higher frequency information to be detected, which in turn informs about shorter (more ancient) admixture tracts.

gain substantial resolution. It seems entirely feasible that wavelet approaches will have sufficient statistical power to reconstruct admixture events far deeper in time than those currently studied. Advances in wavelet methods therefore offer exciting potential for future research, particularly for ancient and complex human admixture processes.

Software

Software for the analyses described here has been released in the form of an R package, *adwave*, which is freely available from the R project's central package repository: http://cran.r-project.org/web/packages/adwave/index.html

Acknowledgments

We gratefully acknowledge assistance with sample collection by Agustini Leonita and Alida Harahap (Eijkman Institute for Molecular Biology, Jakarta, Indonesia) and J. Stephen Lansing (University of Arizona), as well as data processing by Olga Savina (University of Arizona). We also thank Matthew Nunes (University of Lancaster, United Kingdom) and Martin Hazelton (Massey University, New Zealand) for their valuable comments.This research was supported by the Royal Society of New Zealand through a Rutherford Fellowship (RDF-10-MAU-001) and Marsden Grant (11-MAU-007) to M.P.C., and by funding from the Allan Wilson Center for Molecular Biology and Evolution.

Literature Cited

Baran, Y., B. Pasaniuc, S. Sankararaman, D. G. Torgerson, C. Gignoux *et al.*, 2012 Fast and accurate inference of local ancestry in Latino populations. Bioinformatics 28: 1359–1367.

Beaumont, M. A., W. Zhang, and D. J. Balding, 2002 Approximate Bayesian Computation in population genetics. Genetics 162: 2025–2035.

Bellwood, P., 2007 *Prehistory of the Indo-Malaysian Archipelago*. ANU E Press, Canberra, Australia.

Brisbin, A., K. Bryc, J. Byrnes, F. Zakharia, L. Omberg *et al.*, 2012 PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Hum. Biol. 84: 343–364.

Brown, R., and B. Pasaniuc, 2014 Enhanced methods for local ancestry assignment in sequenced admixed individuals. PLOS Comput. Biol. 10: e1003555.

Bryc, K., A. Auton, M. R. Nelson, J. R. Oksenberg, S. L. Hauser *et al.*, 2010 Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci. USA 107: 786–791.

Chatfield, C., 2003 *The Analysis of Time Series: An Introduction*, 6th Ed. Chapman & Hall/CRC, Boca Raton, FL.

Chen, G. K., P. Marjoram, and J. D. Wall, 2009 Fast and flexible simulation of DNA sequence data. Genome Res. 19: 136–142.

Cox, M. P., A. E. Woerner, J. D. Wall, and M. F. Hammer, 2008 Intergenic DNA sequences from the human X chromosome reveal high rates of global gene flow. BMC Genet. 9: 76.

Csilléry, K., M. G. B. Blum, O. E. Gaggiotti, and O. François, 2010 Approximate Bayesian Computation (ABC) in practice. Trends Ecol. Evol. 25: 410–418.

Csilléry, K., O. François, and M. G. B. Blum, 2012 abc: an R package for approximate Bayesian computation (ABC). Methods Ecol. Evol. 3: 475–479.

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.

Fenner, J. N., 2005 Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am. J. Phys. Anthropol. 128: 415–423.

Gravel, S., 2012 Population genetics models of local ancestry. Genetics 191: 607–619.

Hellenthal, G., G. B. J. Busby, G. Band, J. F. Wilson, C. Capelli *et al.*, 2014 A genetic atlas of human admixture history. Science 343: 747–751.

Jombart, T., S. Devillard, and F. Balloux, 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11: 94.
Liò, P., 2003 Wavelets in bioinformatics and computational biology: state of art and perspectives. Bioinformatics 19: 2–9.
Loh, P.-R., M. Lipson, N. Patterson, P. Moorjani, J. K. Pickrell *et al.*, 2013 Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193: 1233–1254.
Ma, J., and C. I. Amos, 2010 Theoretical formulation of principal components analysis to detect and correct for population stratification. PLoS ONE 5: e12510.
Ma, J., and C. I. Amos, 2012 Principal components analysis of population admixture. PLoS ONE 7: e40115.
Maples, B. K., S. Gravel, E. E. Kenny, and C. D. Bustamante, 2013 RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. 93: 278–288.
McVean, G., 2009 A genealogical interpretation of principal components analysis. PLoS Genet. 5: e1000686.
Moorjani, P., N. Patterson, J. N. Hirschhorn, A. Keinan, L. Hao *et al.*, 2011 The history of African gene flow into southern Europeans, Levantines, and Jews. PLoS Genet. 7: e1001373.
Novembre, J., and M. Stephens, 2008 Interpreting principal component analyses of spatial population genetic variation. Nat. Genet. 40: 646–649.
Patterson, N., A. L. Price, and D. Reich, 2006 Population structure and eigenanalysis. PLoS Genet. 2: e190.
Percival, D. P., 1995 On estimation of the wavelet variance. Biometrika 82: 619–631.
Pool, J. E., and R. Nielsen, 2009 Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181: 711–719.
Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels *et al.*, 2009 Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 5: e1000519.
Pugach, I., R. Matveyev, A. Wollstein, M. Kayser, and M. Stoneking, 2011 Dating the age of admixture via wavelet transform analysis of genome-wide data. Genome Biol. 12: R19.
Purcell, S., 2009 PLINK, http://pngu.mgh.harvard.edu/purcell/plink/.
Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575.
R Development Core Team, 2014 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Robinson, J. D., L. Bunnefeld, J. Hearn, G. N. Stone, and M. J. Hickerson, 2014 ABC inference of multi-population divergence with admixture from unphased population genomic data. Mol. Ecol. 23: 4458–4471.
Sankararaman, S., G. Kimmel, E. Halperin, and M. I. Jordan, 2008 On the inference of ancestries in admixed populations. Genome Res. 18: 668–675.
Sousa, V. C., M. Fritz, M. A. Beaumont, and L. Chikhi, 2009 Approximate Bayesian Computation without summary statistics: the case of admixture. Genetics 181: 1507–1519.
Sunnåker, M., A. G. Busetto, E. Numminen, J. Corander, M. Foll *et al.*, 2013 Approximate Bayesian computation. PLOS Comput. Biol. 9: e1002803.
Whitcher, B., 2013 waveslim: Basic wavelet routines for one-, two- and three-dimensional signal processing. http://cran.r-project.org/web/packages/waveslim/index.html.
Xu, S., I. Pugach, M. Stoneking, M. Kayser, L. Jin *et al.*, 2012 Genetic dating indicates that the Asian-Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. Proc. Natl. Acad. Sci. USA 109: 4574–4579.

Communicating editor: K. M. Roeder

5

THE INDONESIAN ARCHIPELAGO: AN ANCIENT GENETIC HIGHWAY LINKING ASIA AND THE PACIFIC

Tumonggor, M.K., T.M. Karafet, B. Hallmark, J.S. Lansing, H. Sudoyo, M.F. Hammer, and **M.P. Cox**. 2013. The Indonesian archipelago: An ancient genetic highway linking Asia and the Pacific. *Journal of Human Genetics* 58:165-173.
http://doi.org/10.1038/jhg.2012.154

ABSTRACT Indonesia, an island nation linking mainland Asia with the Pacific world, hosts a wide range of linguistic, ethnic and genetic diversity. Despite the complexity of this cultural environment, genetic studies in Indonesia remain surprisingly sparse. Here, we report mitochondrial DNA (mtDNA) and associated Y-chromosome diversity for the largest cohort of Indonesians examined to date – 2740 individuals from 70 communities spanning 12 islands across the breadth of the Indonesian archipelago. We reconstruct 50 000 years of population movements, from mitochondrial lineages reflecting the very earliest settlers in island southeast Asia, to Neolithic population dispersals. Historic contacts from Chinese, Indians, Arabs and Europeans comprise a noticeable fraction of Y-chromosome variation, but are not reflected in the maternally inherited mtDNA. While this historic immigration favored men, patterns of genetic diversity show that women moved more widely in earlier times. However, measures of population differentiation signal that Indonesian communities are trending away from the matri- or ambilocality of early Austronesian societies toward the more common practice of patrilocal residence today. Such sex-specific dispersal patterns remain even after correcting for the different mutation rates of mtDNA and the Y chromosome. This detailed palimpsest of Indonesian genetic diversity is a direct outcome of the region's complex history of immigration, transitory migrants and populations that have endured *in situ* since the region's first settlement.

SUPPLEMENTARY MATERIALS
http://www.nature.com/jhg/journal/v58/n3/suppinfo/jhg2012154s1.html

Journal of Human Genetics (2013) 58, 165–173

www.nature.com/jhg

ORIGINAL ARTICLE

The Indonesian archipelago: an ancient genetic highway linking Asia and the Pacific

Meryanne K Tumonggor[1,2], Tatiana M Karafet[3], Brian Hallmark[3], J Stephen Lansing[1,4], Herawati Sudoyo[2], Michael F Hammer[3] and Murray P Cox[5]

Indonesia, an island nation linking mainland Asia with the Pacific world, hosts a wide range of linguistic, ethnic and genetic diversity. Despite the complexity of this cultural environment, genetic studies in Indonesia remain surprisingly sparse. Here, we report mitochondrial DNA (mtDNA) and associated Y-chromosome diversity for the largest cohort of Indonesians examined to date—2740 individuals from 70 communities spanning 12 islands across the breadth of the Indonesian archipelago. We reconstruct 50 000 years of population movements, from mitochondrial lineages reflecting the very earliest settlers in island southeast Asia, to Neolithic population dispersals. Historic contacts from Chinese, Indians, Arabs and Europeans comprise a noticeable fraction of Y-chromosome variation, but are not reflected in the maternally inherited mtDNA. While this historic immigration favored men, patterns of genetic diversity show that women moved more widely in earlier times. However, measures of population differentiation signal that Indonesian communities are trending away from the matri- or ambilocality of early Austronesian societies toward the more common practice of patrilocal residence today. Such sex-specific dispersal patterns remain even after correcting for the different mutation rates of mtDNA and the Y chromosome. This detailed palimpsest of Indonesian genetic diversity is a direct outcome of the region's complex history of immigration, transitory migrants and populations that have endured *in situ* since the region's first settlement.
Journal of Human Genetics (2013) **58,** 165–173; doi:10.1038/jhg.2012.154; published online 24 January 2013

Keywords: Indonesia; mitochondrial DNA; molecular anthropology; Y chromosome

INTRODUCTION

Indonesia, a maritime nation comprising over 17 000 islands straddling the Pacific and Indian Oceans, links mainland Asia with the Pacific world. Although a single lingua franca is spoken widely across the archipelago today (Bahasa Indonesia), Indonesia hosts over 730 indigenous languages and associated ethnic groups.[1] Most of these languages belong to the geographically dispersed Austronesian language family, but Papuan languages are spoken by some groups in the far east of the archipelago.[2] Relative to its land area, Indonesia is one of the most varied regions on earth in terms of ethnic, linguistic and genetic diversity.

As an island nation, past changes in global climate have had an especially strong influence on Indonesia.[3] Lands in the west, now largely submerged, once formed a vast continental shelf jutting out from Asia (Sundaland). In the east, Australia and New Guinea were linked into a single continent (Sahul). During most of the late Pleistocene, it was possible to walk—with only minor water crossings—from Bangkok to Sydney. However, following the end of the last glacial period, ~18 kya, ice melting in the arctic fueled a rapid rise in global sea levels. Continental Indonesia swiftly fragmented into the long chain of islands that characterizes the nation today.

The human history of Indonesia played out against the backdrop of this dynamically changing geography. Settled by anatomically modern humans at least 47 kya[4,5] and perhaps much earlier,[6] our species has inhabited Indonesia longer than Europe.[7,8] Archeology provides sporadic, but widespread, evidence of early hunter-gatherer groups throughout the Pleistocene.[9] However, the biggest cultural change occurred within the last 10 kya when the archeological record abruptly chronicles the appearance of agricultural communities together with pottery, plant cultivation and animal domestication. The vast spread of Austronesian languages likely also occurred during this time.[10–12] Whether the Neolithic era was ushered in by population movements from Taiwan[13] or was instead dominated by regional developments[14,15] remains a highly contentious topic of discussion—a question that genetics is increasingly striving to answer.

For such a large and ethnically diverse nation at the pivot point of Asia and the Pacific, studies of Indonesian genetic diversity are surprisingly sparse.[16–24] The most complete study of mitochondrial DNA (mtDNA) posited that Indonesian diversity has largely been shaped by two forces: population movements driven by sea level changes, and by farming populations expanding from the Asian mainland into the islands of southeast Asia.[25,26] Y-chromosome

[1]Department of Anthropology, University of Arizona, Tucson, AZ, USA; [2]Eijkman Institute for Molecular Biology, Jakarta, Indonesia; [3]Arizona Research Laboratories, Division of Biotechnology, University of Arizona, Tucson, AZ, USA; [4]Santa Fe Institute, Santa Fe, NM, USA and [5]Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand
Correspondence: Professor MF Hammer, Arizona Research Laboratories, Division of Biotechnology, University of Arizona, Tucson, AZ 85721, USA.
E-mail: mfh@email.arizona.edu
Received 23 September 2012; revised 28 November 2012; accepted 14 December 2012; published online 24 January 2013

studies, reflecting the history of Indonesian men, mirror these themes, emphasizing a complex multifaceted history of the region's islands and communities.[20] Although autosomal markers are rapidly changing our understanding of regional prehistory,[27–29] the geographical resolution and widespread availability of comparative data ensure that haploid markers remain extremely powerful tools for studying the human past. Here, we present the largest sample of Indonesian mtDNA diversity assembled to date—2740 individuals from 70 populations on 12 islands spanning the full range of Indonesia's geographical, ethnic and linguistic diversity. We reconstruct a partial history of Indonesia's women, piecing together 50 ky of population movements that have shaped the diversity of Indonesians living today. Finally, by comparison with Y-chromosome data for the same set of individuals,[20] we contrast the histories of Indonesian men and women, presenting common patterns of shared inheritance with key points of demographic and social difference.

MATERIALS AND METHODS

Samples

Genetic diversity was screened in 2740 consenting, closely unrelated and seemingly healthy individuals drawn from across the Indonesian archipelago. Permission to conduct research in Indonesia was granted by the Indonesian Institute of Sciences. Biological samples (peripheral blood and buccal swabs) were obtained by MKT, JSL, HS, Golfiani Malik, Wuryantari Setiadi and Loa Helena Suryadi of the Eijkman Institute for Molecular Biology, Jakarta, Indonesia, with the assistance of Indonesian Public Health clinic staff, and followed protocols for the protection of human subjects established by both the Eijkman Institute and the University of Arizona Institutional Review Boards. Participant interviews confirmed ethnic, linguistic and geographic classifications for at least two generations into the past.

Seventy populations from 12 island groups were sampled across the Indonesian archipelago, including (from west to east) Sumatra ($n=42$), Mentawai ($n=128$), Nias ($n=59$), Java ($n=51$), Bali ($n=487$), Sulawesi ($n=200$), Sumba ($n=634$), Flores ($n=469$), Lembata ($n=92$), Pantar ($n=29$), Timor ($n=526$) and Alor ($n=23$) (Supplementary Table S1). For comparative purposes, we also explored previously published genetic data from mainland Asian populations: northwestern Chinese (Yili, Xinjiang; $n=47$), northeastern Chinese (Fencheng, Liaoning; $n=51$), southwestern Chinese (Kunming, Yunnan; $n=43$) and southeastern Chinese (Zhanjiang, Guangdong; $n=30$),[30] Thai ($n=52$) and Vietnamese ($n=41$),[31,32] and indigenous Malaysians ($n=260$).[33] We also compared our Indonesian data with neighboring island southeast Asian populations, including indigenous Taiwanese ($n=640$),[34] Filipinos ($n=423$), Papua New Guineans ($n=231$), Island Melanesians ($n=1366$) and Micronesians ($n=47$).[35,36] The geographical locations of these populations are illustrated in Figure 1.

DNA extraction and genetic screening

DNA was extracted from peripheral blood samples using the salting-out procedure of Miller, Dykes and Polesky.[37] DNA from buccal swabs was extracted using standard phenol–chloroform protocols.

The first hypervariable segment (HVS I) of mtDNA was amplified using primers L15926 (5′-TCAAAGCTTACACCAGTCTTGTAAACC-3′) and H639 (5′-GGGTGATGTGAGCCCGTCA-3′). PCR amplicons were sequenced in both forward and reverse directions using primers L15965 (5′-CAAGGACAAAT CAGAGAA-3′) and H11 (5′-GTGGTTAATAGGGTGATAG-3′). Traditional Sanger sequences were aligned and edited with Sequencher v. 5.0 (Gene Codes Corporation, Ann Arbor, MI, USA; http://www.genecodes.com). Polymorphisms were scored relative to the revised Cambridge reference sequence.[38]

Haplogroups were initially predicted from HVS I sequences using known reference genomes.[23,25,36,39–42] Assignments were subsequently confirmed using Taqman and restriction fragment length polymorphism assays (Supplementary Table S2).

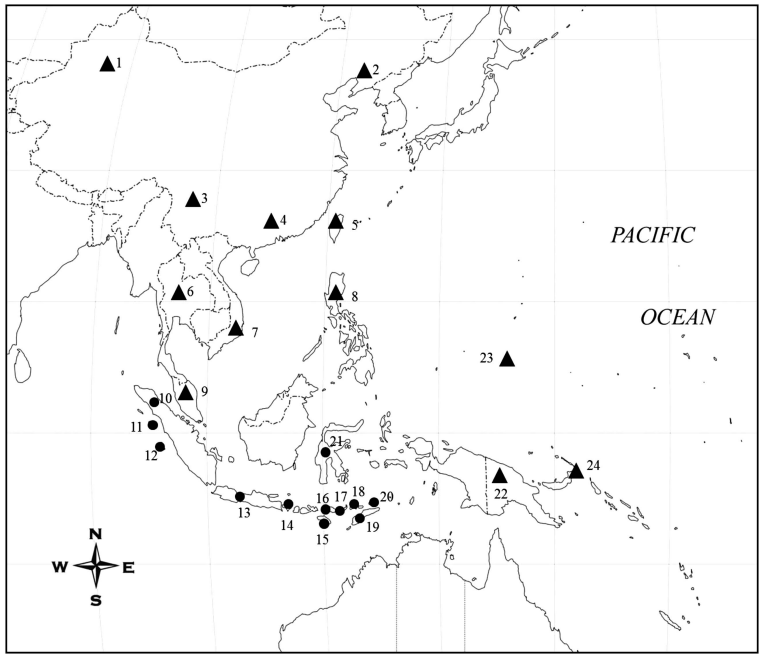

Figure 1 Locations of Studied Populations. (1) Northwest China,[30] (2) northeast China,[30] (3) southwest China,[30] (4) southeast China,[30] (5) Indigenous Taiwanese,[34] (6) Thailand,[31] (7) Vietnam,[32] (8) Philippines,[36] (9) Indigenous Malaysians,[33] (10) Sumatra (present study), (11) Nias (present study), (12) Mentawai (present study), (13) Java (present study), (14) Bali (present study), (15) Sumba (present study), (16) Flores (present study), (17) Lembata (present study), (18) Pantar (present study), (19) Timor (present study), (20) Alor (present study), (21) Sulawesi (present study), (22) Papua New Guinea,[35] (23) Micronesia[35] and (24) Melanesia[35].

Table 1 Molecular diversity indices and growth summary statistics for Indonesian island groups

				Diversity indices						*Growth Statistics*			
Island	N	h	S	*Haplotype diversity*	*s.d.*	*MNPD*	*s.d.*	*Nucleotide diversity*	*s.d.*	*Tajima's D*	P	*Fu's Fs*	P
Sumatra	42	29	48	0.973	0.013	7.90	3.75	0.015	0.008	−1.03	0.146	***−13.2***	***<0.001***
Mentawai	128	20	38	0.890	0.012	6.77	3.21	0.013	0.007	−0.10	0.504	0.500	0.632
Nias	59	24	44	0.862	0.040	7.02	3.34	0.013	0.007	−0.87	0.196	−4.54	0.104
Java	51	33	52	0.968	0.013	7.68	3.64	0.014	0.008	−1.16	0.084	***−16.1***	***<0.001***
Bali	487	129	99	0.975	0.002	8.16	3.79	0.015	0.008	−1.29	0.071	***−24.1***	***0.002***
Sulawesi	200	97	87	0.976	0.005	7.53	3.53	0.014	0.007	***−1.51***	***0.040***	***−24.5***	***0.001***
Sumba	634	159	108	0.981	0.001	7.89	3.68	0.015	0.008	***−1.40***	***0.038***	***−24.0***	***0.001***
Flores	469	149	109	0.987	0.001	8.22	3.82	0.016	0.008	***−1.44***	***0.038***	***−24.1***	***0.002***
Lembata	92	47	68	0.968	0.008	9.12	4.23	0.017	0.009	−1.04	0.154	***−20.7***	***<0.001***
Timor	526	117	96	0.955	0.005	8.72	4.03	0.016	0.008	−1.08	0.115	***−24.0***	***0.002***
Alor	23	22	45	0.996	0.014	9.12	4.36	0.017	0.009	−0.98	0.176	***−13.4***	***<0.001***
Pantar	29	22	52	0.978	0.015	10.0	4.73	0.019	0.010	−0.89	0.192	***−6.78***	***0.016***

Abbreviations: *h*, number of haplotypes; MNPD, mean number of pairwise differences; *N*, number of sequences; *P*, probability value; *S*, number of polymorphic sites.
Significant growth summary statistics are bold and italicized.

Statistical analyses

Molecular diversity, population structure estimates and genetic distances between populations were calculated using Arlequin v. 3.11[43] (http://cmpg.unibe.ch/software/arlequin3). The statistical significance of F_{ST} values was evaluated using 10^4 permutations of the data. Pairwise genetic distances between populations were computed as the linearized value, $F_{ST}/(1-F_{ST})$.[44,45] Differences in haplogroup frequencies between eastern and western Indonesia were determined via parametric bootstrapping using code implemented in R (available from the authors on request).[46]

Median-joining networks were built using Network v. 4.5.1.6 (Fluxus Engineering; http://www.fluxus-engineering.com).[47] Haplogroups were tentatively dated with the ρ statistic method[48] using a rate of one mutation every 19 171 years.[49] Dates are only intended as a rough guide for relative haplogroup ages.[50]

MtDNA and Y-chromosome comparisons

Y-chromosome diversity has also been screened for the same panel of individuals described above.[20] Differences in mtDNA and Y-chromosome diversity between populations were analyzed using an analysis of molecular variance implemented in Arlequin. A measure of interlocus differentiation $\acute{G}_{ST}$,[51] standardized for different mutation rates, was calculated using code implemented in R (available from the authors on request).[46]

MtDNA HVS I sequences have been deposited in GenBank (accession numbers: KC113641–KC115854). Y-chromosome STR data are provided as Supplementary Data Set S1.

RESULTS

We screened 2740 individuals from 70 communities on 12 Indonesian islands. MtDNA HVS I sequences showed high levels of diversity, as measured by the number of polymorphic sites, number of haplotypes, the mean number of pairwise differences and Nei's haplotype diversity (Table 1). The haplotype diversity of communities ranged from 0.862–0.996, which indicates that most individuals within these groups carry unique mtDNA lineages. The most diverse communities are found in eastern parts of the archipelago (Sumba, Flores, Pantar and Alor), where both Asian and Papuan lineages occur side by side. The western barrier islands of Nias and Mentawai are least diverse, even when compared with other Asian populations. For instance, Nias and Mentawai are the only Indonesian populations with diversity as low as that of indigenous Taiwanese groups (0.838–0.924).[34,52]

Summary statistics such as Fu's Fs and Tajima's D can be informative about the roles of selection and demography. For mtDNA control region sequences, which seem little affected by natural selection, observed values are suggestive of low levels of growth across the archipelago, with the exception of the barrier islands, Nias and Mentawai (Table 1). Growth seems strongest in the central cluster of islands (Bali, Sulawesi, Sumba and Flores), but is less pronounced in the extreme west and east of the archipelago.

Individuals were assigned to mtDNA haplogroups using a combination of HVS I sequence motifs and single-nucleotide polymorphisms (SNPs) distributed around the coding region of the mtDNA genome. Fifty-one haplogroups were identified, with all lineages falling into macrohaplogroups M (47.05%) and N (52.95%). The 51 Indonesian haplogroups are plotted on a tree of mtDNA diversity constructed using previously published HVS I sequences and coding-region SNPs[23,25,36,39–42] (Figure 2).

Haplogroup frequencies differ between western and eastern Indonesia (Supplementary Table S3). In the west, haplogroups B5a (12%), B4c1b3 (9%) and Y2 (10.5%) are carried by a third of individuals. These haplogroups, frequent in western Indonesia, are notable by their near absence in eastern Indonesia. In the east, haplogroups F1a4 (8.7%), Q including Q1 and Q2 (7.7%), P (2.8%) and B4a1a1a (2.3%) represent nearly a quarter of individuals. Correspondingly, these haplogroups are rare or absent in western Indonesia, which is expected for lineages with strong Papuan connections (P and Q), but more surprising for lineages like the Polynesian motif (B4a1a1a). The Polynesian motif is found as far west as Bali, albeit in just two individuals (0.4%). However, it was not detected in samples from the western Indonesian islands of Java, Sumatra, Nias and Mentawai, even though this region is thought to have contributed to the settlement of Madagascar where the Polynesian motif is carried by nearly a third of individuals.[53,54] The prevalence of the Polynesian motif in Madagascar, and its absence from the island region where the inhabitants of Madagascar originated, has yet to be satisfactorily explained.

We compared the distribution of Indonesian mtDNA haplogroups with those of surrounding populations (Supplementary Table S4). Most haplogroups are shared. The deep maternal lineages M17a, M73, M47, N21, N22, R21, R22 and R23 have patchy distributions across

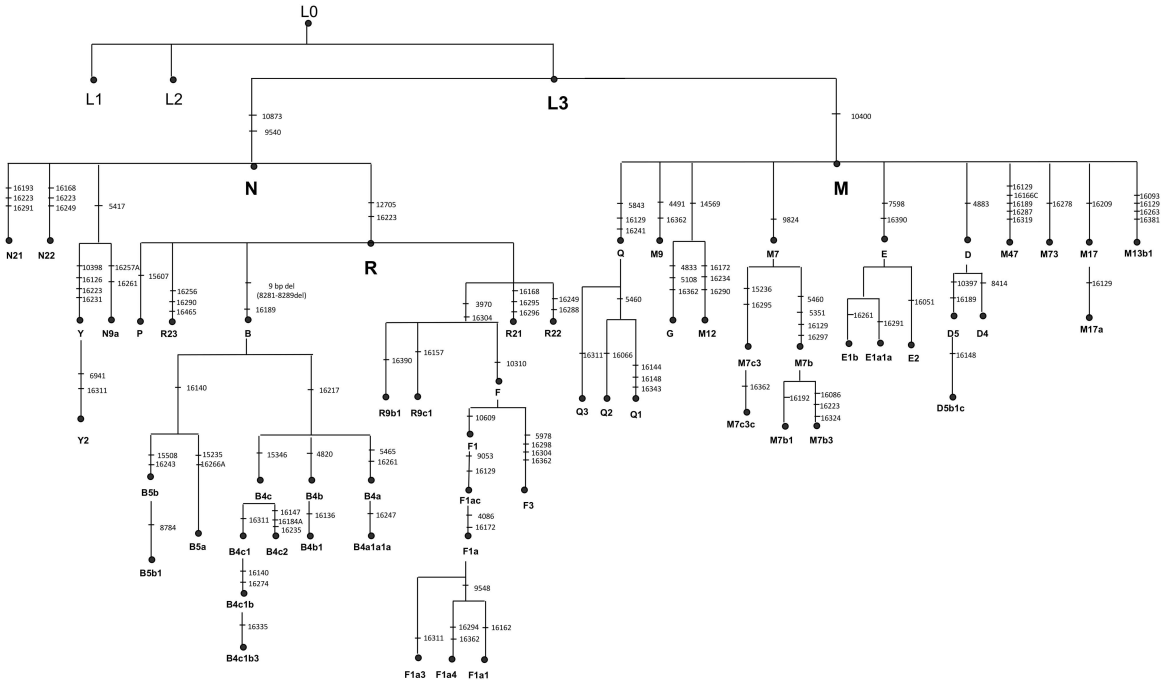

Figure 2 Phylogeny of mitochondrial DNA haplogroups observed in Indonesian populations in the present study.

mainland and island southeast Asia, likely reflecting ancient maternal lineages tracing back to the first settlers in this region.[25,33,36,55] Four of these lineages (M17a, N21, R22 and R23) reach higher frequencies in western compared with eastern Indonesia (parametric bootstrap, all $P \ll 0.01$) (Supplementary Table S3).

Owing to proposed population origins of Austronesian language speakers in Taiwan, Indonesian links to the Philippines and indigenous Taiwanese are of especial interest. These three locations share four haplogroups (E1a1a, M7b3, M7c3c and Y2), which have previously been suggested as candidates for a mid-Holocene dispersal out of Taiwan.[25,36] Figure 3 illustrates that lineages shared with Filipinos and indigenous Taiwanese are generally more common in the east than in the west of Indonesia. The exceptions are Y2 and M7c3c, where high frequencies in Nias and Mentawai may perhaps be caused by genetic drift or strong founder events in these extremely small and geographically isolated populations on the barrier islands of Sumatra.[52]

Median-joining networks were constructed for haplogroups shared between Indonesia, the Philippines and Taiwan (Figure 3). We note that at least one haplotype was shared between Indonesia and Taiwan for each of these four lineages (E1a1a, M7b3, M7c3c and Y2). In all cases, this shared lineage was the ancestral haplotype, and descendent lineages depict a star-like expansion indicative of population growth and/or geographical expansion. Unfortunately, the networks are not informative about the direction of migration: the data would fit a model of rapid expansion from Taiwan to the Philippines and Indonesia, but are equally consistent with population movements in the opposite direction.

To explore population relationships further, multidimensional scaling was performed on all Indonesian HVS I sequences using Slatkin's linearized F_{ST} as the genetic distance between groups (Supplementary Figure S1). Although there is no simple population division, western Indonesian groups cluster away from eastern Indonesian populations. Multidimensional scaling analysis with regional neighbors shows that Indonesians fall together with Asian groups, but away from Oceanian populations (Supplementary Figure S2). Among Asian populations, Indonesians cluster most closely with the Philippines and Vietnam, and more distantly with Taiwan (Supplementary Figure S3).

An analysis of molecular variance illustrates that the Y-chromosome STRs ($F_{ST} = 0.202$) have markedly higher variation among populations than mtDNA HVS I ($F_{ST} = 0.073$) (Table 2), suggesting that women have dispersed more widely in the past than men. This trend is maintained when populations are collapsed to their 12 island groups or an even broader east–west division. The trend holds, albeit more weakly, when distances are standardized for the ~400-fold higher mutation rate of the Y chromosome (on the order of 10^{-5} mutation events per STR per year)[56–59] relative to mtDNA (on the order of 10^{-7} mutation events per base pair per year).[49] The $\acute{G}_{ST}$ of Y-chromosome STRs ($\acute{G}_{ST} = 0.972$) is still notably higher than that of mtDNA HVS I ($\acute{G}_{ST} = 0.862$) (Supplementary Table S5). Curiously, the trend even holds on very small geographical scales, such as among communities situated along the highland river systems of Bali.[60]

DISCUSSION

Austronesian languages, spoken from Madagascar in the west to Rapanui/Easter Island in the east, form one of the world's largest and most geographically dispersed language families. Despite intense archeological, linguistic and genetic research, where people speaking these languages came from, how they dispersed and what their ancestral communities looked like remain major open questions. One hypothesis states that Austronesian speakers originated in Taiwan, which is supported by the fact that the most basal languages of the

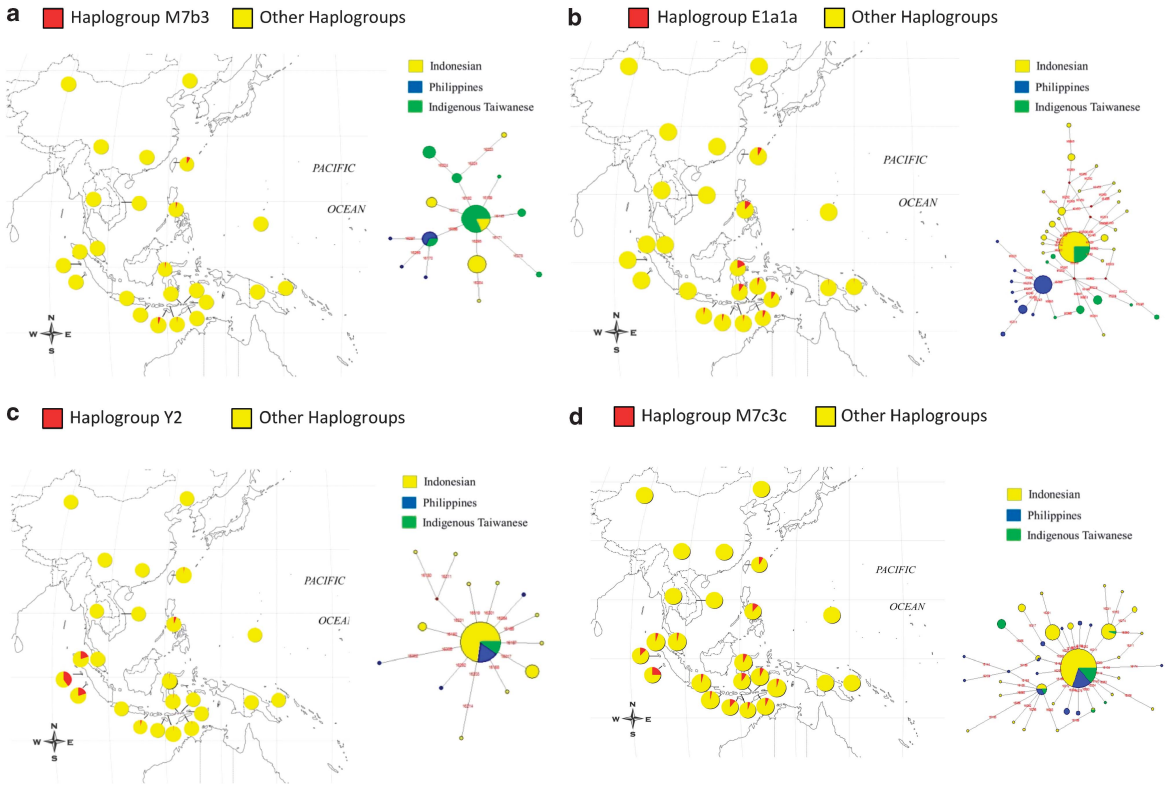

Figure 3 Frequency distributions and median-joining networks of mitochondrial DNA haplogroups (**a**) M7b3, (**b**) E1a1a, (**c**) Y2 and (**d**) M7c3c shared by indigenous Taiwanese (green), Filipino (blue) and Indonesian populations (yellow). Note that haplogroup networks are largely agnostic about the direction of population movements between Taiwan, the Philippines and Indonesia.

Table 2 Analysis of molecular variance (AMOVA) for subsets of Indonesian populations

		Within populations		*Among populations within groups*		*Among groups*	
		F_{ST}	*Var (%)*	F_{SC}	*Var (%)*	F_{CT}	*Var (%)*
Y-SNPs							
All Indonesian populations	70 populations	0.361	63.9				
Islands	12 islands	0.384	61.6	0.195	14.9	0.236	23.5
East vs West	2 main groups	0.475	52.5	0.247	17.2	0.304	30.3
Y-STRs							
All Indonesian populations	70 populations	0.202	79.8				
Islands	12 islands	0.210	78.9	0.152	14.2	0.069	6.90
East vs West	2 main groups	0.222	77.8	0.188	18.0	0.042	4.20
mtDNA SNPs							
All Indonesian populations	70 populations	0.092	90.8				
Islands	12 islands	0.096	90.4	0.067	6.45	0.032	3.15
East vs West	2 main groups	0.109	89.1	0.080	7.76	0.031	3.14
mtDNA HVS I							
All Indonesian populations	70 populations	0.073	92.7				
Islands	12 islands	0.077	92.3	0.048	4.61	0.031	3.09
East vs West	2 main groups	0.088	91.2	0.062	6.05	0.028	2.75

Abbreviations: HVS I, first hypervariable segment; mtDNA, mitochondrial DNA; SNP, single-nucleotide polymorphism; Var, variance.
All values are statistically significant ($P \ll 0.0001$).

Austronesian family are spoken there.[61] From Taiwan, nascent farming groups are believed to have expanded 3–4000 years ago through the Philippines, Indonesia and out into the Pacific.[3,13] Alternative hypotheses, such as Austronesian groups originating in and dispersing from Indonesia, remain possible,[14,62] with many genetic lineages in Indonesia showing old and local connections. Models combining these two extremes may ultimately be the best predictors. We envisage some genetic contributions from Taiwan, possibly including speakers of early Austronesian languages, with a substantial biological heritage from waves of ancestral populations arriving in island southeast Asia following its first settlement 50 kya.

We consider how the mitochondrial data aligns with this spectrum of origin models. Some older lineages appear to trace back to the very earliest settlers in southeast Asia. R21, which is found only in Mentawai (0.8%), diverged from the common haplogroup R ancestor ~60 kya.[55,63] Other basal mtDNA lineages (for example, N21 and N22) are shared at extremely low levels by Indonesians and mainland southeast Asian groups.

Candidate mtDNA lineages of a Taiwanese dispersal (E1a1a, M7b3, M7c3c and Y2) have also been proposed.[25,36] These four haplogroups have similar distributions, with basal haplotypes shared between indigenous Taiwanese, Filipinos and Indonesians. However, the direction of dispersal is inconclusive: ancestral-derived haplotype orders are consistent with a rapid expansion from Taiwan to the Philippines and Indonesia, but population dispersals in the opposite direction are equally likely. Although we provide haplogroup dates with some reluctance,[50] we note that relative ages are inconsistent with a simple dispersal from Taiwan to the Philippines, and thence to Indonesia. Instead, they seem a better fit to widespread population movements within island southeast Asia during the Holocene. However, we note that sample sizes differ substantially between these three locations (2740, 423 and 640 for Indonesians, Filipinos and indigenous Taiwanese, respectively), which adds variance—and perhaps bias—to diversity and dating estimates. Considered together with their large confidence intervals, we are reluctant to draw strong conclusions from molecular dates alone.

The Polynesian motif is also generally associated with a Taiwanese dispersal, but actually possesses an unusual geographical distribution. The ancestral form occurs widely throughout mainland and island southeast Asia. However, the Polynesian motif itself is found only at low frequency in the Philippines (0.5%)[36] and eastern Indonesia (2.3%).[16] Although frequencies reach as high as 7.4% on Timor, the lineage is found no further west than Bali (0.4%, or just 2 of 457 individuals). This is consistent with a proposed origin in island Melanesia,[64] but notably conflicts with the high frequency of the Polynesian motif in Madagascar, which was settled ~1200 years ago from western Indonesian sources.[53] We suggest that an inclusive framework that describes the full distribution of this unusual mtDNA lineage is still lacking. Nevertheless, an unambiguous connection with population dispersals from Taiwan during the Neolithic seems increasingly unlikely.

MtDNA evidence does suggest that most Indonesian groups may have increased in size. Negative values of Fu's Fs and Tajima's D point toward population growth and/or geographical expansion (Table 1), and reinforce similar conclusions that might be drawn from the star-like phylogenies of shared island southeast Asian mtDNA lineages (Figure 3). Signals of growth are greatest in the center of the archipelago (Bali, Sulawesi, Sumba and Flores), but weaker towards the eastern and western peripheries. Key exceptions are the barrier islands, Nias and Mentawai, whose mtDNA profiles are statistically more consistent with constant population size. These two groups also show some of the lowest haplotype diversities of any island southeast Asian populations (for instance, 43.5% of the Nias population carry haplogroup Y2). These outlier patterns may be due to genetic drift, as settlements on Nias and Mentawai are small even by indigenous Indonesian standards, or they may be caused by founder events and therefore reflect the unusual genetic profile of the islands' first settlers. The strong retained cultural heritage of these barrier islands may speak to this point. The inhabitants of Nias still practice old traditions, including the construction and maintenance of megaliths, and stone jumping (*hombo batu*), whereby young men show prowess by repeatedly jumping up onto a tall stone. Megalithic traditions associated with Austronesian culture were once common across Indonesia, but are now largely restricted to peripheral communities including those on Nias, Mentawai and Sumba.

In comparisons with neighboring populations, Indonesia's closest genetic connections lie toward mainland and island southeast Asia rather than Oceania (Supplementary Figures S2 and S3). Western Indonesian groups are notably distinct from Papuan groups (Figure 4), largely owing to low levels of haplogroups P and Q.[65] Whether these lineages are an enduring local presence from the original inhabitants of the region, or instead reflect recent westward movements from New Guinea is unclear. Soares *et al.*[64] suggest that the Polynesian motif spread westward from the Bismarck Archipelago, and similar movements have been proposed to explain the distribution of Papuan languages in eastern Indonesia.[2,66] SNPs from across the genome have been used to argue for an indigenous presence of Papuan genotypes rather than back migration from New Guinea, but strong statistical support for this is currently lacking.[67] Unfortunately, the data presented here, while powerful in helping to distinguish shared connections, are less informative about directions of movement.

The individuals screened for mtDNA were also assayed for Y-chromosome diversity.[20] Analysis of Austronesian languages and cultural systems,[10,11,68] as well as autosomal markers,[27,29] suggests that the men and women of island southeast Asia have followed quite different social histories. We compared mtDNA and Y-chromosome diversity to explore this further. Genetic divisions between populations are far weaker for mtDNA HVS I ($F_{ST} = 0.073$) than for Y-chromosome STRs ($F_{ST} = 0.202$), and this effect is even more pronounced at the haplogroup level when Indonesia is separated into its eastern and western parts (mtDNA SNPs $F_{ST} = 0.109$; Y-chromosome SNPs $F_{ST} = 0.475$) (Table 2). This discrepancy suggests that men and women have had different patterns of dispersal, with women moving widely between communities, while men have historically stayed local. One possible social explanation is patrilocality, where men remain in their natal community, but women move to the home village of their husband. Indeed, most of the populations presented in this study are patrilineal today. Interestingly, matrilineal systems are thought to have dominated ancestral Austronesian societies.[11,27,69] These mtDNA/Y-chromosome patterns may have been laid down after that cultural shift, and therefore reflect only the last few thousand years. Alternately, patrilocality could instead be the long-term standard with a transient switch to matrilocality during the Austronesian era. Differences in mutation rates could potentially confound this analysis; Y-STRs have high mutation rates, even when compared with the fast evolving mtDNA. However, these general patterns are still supported even when using standardized measures of genetic variation (Supplementary Table S5).

However, while the genetic evidence indicates that men and women experienced divergent histories, shared characteristics are more pronounced than differences. Karafet *et al.*[20] proposed a four-stage

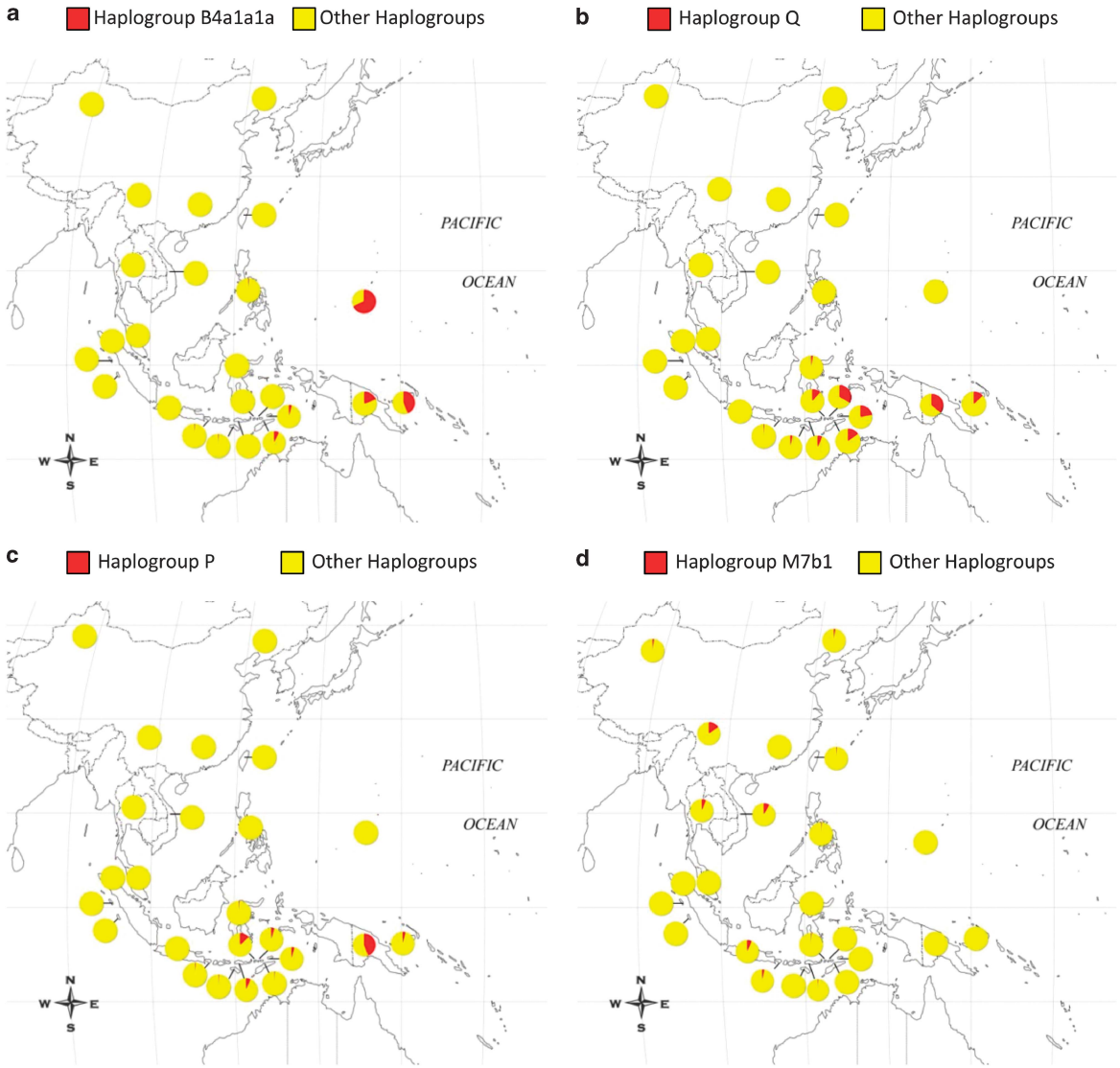

Figure 4 Frequency distribution of mitochondrial DNA haplogroups (**a**) B4a1a1a, (**b**) Q, (**c**) P and (**d**) M7b1.

colonization model built around variation on the Y chromosome. Here, we integrate the mtDNA evidence to provide a broader framework for reconstructing the history of Indonesia.

The first stage of Indonesian prehistory represents the archipelago's initial settlement as part of the African dispersal ~50 kya. The geography of the region was then markedly different from today. Sea levels were much lower, most modern islands had merged into larger landmasses and the westernmost parts of Indonesia were physically contiguous with mainland Asia. This first stage is recorded by deep mtDNA lineages (M17a, M73, M47, N21, N22, R21, R22 and R23), which trace back to the main branching of macrohaplogroups M and N, and have a spotty distribution across both mainland and island southeast Asia today.[25,33,36,55]

The second stage reflects recurrent colonization events from mainland Asia throughout the later Paleolithic. Many haplogroups (B4a, B4b, B4c, B4c1b3, B5a, B5b, B5b1, D and E) show origin dates of 10–40 kya (Supplementary Table S6)[25,63] and are distributed across a wide range of mainland and island southeast Asian populations. As these lineages vary considerably in diversity (and hence, probable age) and show quite different geographical distributions, it is unlikely that any single demic event brought them to Indonesia. Indeed, many of these mtDNA haplogroups have been identified as key components of populations in peninsular Malaysia.[63] Therefore these lineages likely reflect multiple population movements from mainland Asia, possibly hunter-gatherers who followed the now-submerged river systems that once ran from mainland Asia between the modern islands of Sumatra, Java and Borneo.

The third stage represents Neolithic movements into and around island southeast Asia. Some of these may involve population dispersals from (and perhaps to) Taiwan, while others reflect

movements between Indonesian island groups. Representative haplogroups include M7b3, E1a1a, M7c3c and Y2. Autosomal data strongly support large demic movements of Asian populations into eastern Indonesia from around 4 kya.[67] Putatively associated with the spread of Austronesian languages, the direction of these movements remains unknown. If population movements during the Holocene were largely restricted to eastern Indonesia, this may explain why basal mtDNA lineages (such as M17a, N21, R22 and R23) are much rarer in the east. Although the Neolithic seems to have been a period of high population movement, it is curious that distinctions between eastern and western parts of the archipelago persist to the present. Within these regions, genetic similarity between island groups suggests that population movements were relatively frequent, consistent with archeological and linguistic evidence of a strong maritime culture. Borders such as Wallace's biogeographical line are poor boundaries for communities with significant ocean-going capabilities,[17] and the reason for an east–west distinction so clearly persisting in the genetic data is still imperfectly understood.[17,27,70] The most inclusive explanations invoke changes in social behavior, such as the emergence of matrilocal 'house societies' (*sociétés à maison*) during the early Austronesian period, and their subsequent disappearance as communities increasingly turned to patrilocality when the expansion period drew to a close.[70]

The fourth stage reflects historic movements into Indonesia, largely involving trade and the associated spread of major religions from India, Arabia and China.[71] Although found at relatively low frequency today, Y-chromosome lineages representing these movements occur across Indonesia,[20] notably in the west, such as the Hindu dominated island of Bali.[21] It is therefore interesting that no mtDNA lineages representing this period of invasion and migration were observed in our substantial data set. Unlike previous stages of Indonesian settlement, trade and religious connections during the historic era must have involved only male travelers, who subsequently took local wives. This difference between mtDNA and the Y chromosome emphasizes the complex, plural nature of most social processes. While most indigenous Indonesian communities practice patrilocality today (and hence exhibit preferential movement of women), long-distance genetic contributions are still effectively driven by men. As is often the case with complex biological systems, seemingly diametrically opposed processes can even act simultaneously in the same community at the same time.

The enormous diversity of language, culture and genetics in Indonesia is a direct outcome of the region's complex history of migration and settlement. Demic and cultural processes are both apparent: movements of people are indicated by related mtDNA lineages, but many aspects of culture—notably the widespread dispersal of Austronesian languages—are not obviously associated with genetics. Indeed, the only mtDNA lineage found across all Indonesian island groups is M7c3c, but this haplogroup, while also present in Taiwan and the Philippines, appears to be absent from other Austronesian-speaking populations in Oceania. Therefore, no single shared mtDNA lineage links all speakers of Austronesian languages, even if only at low frequency. Instead, Austronesian populations are characterized more by their diversity than by any shared genetic inheritance. However, there are limits to the power of uniparentally inherited markers, and with nearly 3000 individuals now screened, we wonder whether further sampling will substantially change the picture of mtDNA diversity portrayed here. Admixture analysis on genome-wide data sets indicates the extent of Asian immigrants in Indonesian populations and assigns the time of admixture to the mid to late Holocene.[67] Therefore, we are hopeful that new studies of autosomal data may help to answer many of the questions that remain outstanding, and we look forward to the clarification that this new wave of genetic evidence promises to bring.

ACKNOWLEDGEMENTS

This research was supported by a US National Science Foundation grant (SES 0725470) to JSL, MFH, TMK, and Joe C. Watkins, which funded the doctoral research of MKT. The Royal Society of New Zealand provided support through a Rutherford Fellowship (RDF-10-MAU-001) and Marsden Grant (11-MAU-007) to MPC.

1 Lewis, M. P. *Ethnologue: Languages of the World* (SIL International, Dallas, TX, USA, 2009).
2 Pawley, A. The chequered career of the Trans New Guinea hypothesis: Recent research and its implications. In *Papuan Pasts: Cultural, Linguistic and Biological Histories of Papuan-Speaking Peoples* (eds Pawley, A., Attenborough, R., Golson, J. & Hide, R.) 67–107 (Pacific Linguistics, Canberra, 2005).
3 Bellwood, P. *Prehistory of the Indo-Malaysian Archipelago* (Australian National University, Canberra, 2007).
4 Barker, G. The archaeology of foraging and farming at Niah Cave, Sarawak. *Asian Perspect.* **44,** 90–106 (2005).
5 Barker, G., Barton, H., Bird, M., Daly, P., Datan, I., Dykes, A. *et al.* The 'human revolution' in lowland tropical Southeast Asia: the antiquity and behavior of anatomically modern humans at Niah Cave (Sarawak, Borneo). *J. Hum. Evol.* **52,** 243–261 (2007).
6 Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K. E., Rasmussen, S., Albrechtsen, A. *et al.* An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334,** 94–98 (2011).
7 Mellars, P. A new radiocarbon revolution and the dispersal of modern humans in Eurasia. *Nature* **439,** 931–935 (2006).
8 Mellars, P. Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science* **313,** 796–800 (2006).
9 O'Connell, J. F. & Allen, J. Dating the colonization of Sahul (Pleistocene Australia–New Guinea): a review of recent research. *J. Archaeol. Sci.* **31,** 835–853 (2004).
10 Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323,** 479–483 (2009).
11 Jordan, F. M., Gray, R. D., Greenhill, S. J. & Mace, R. Matrilocal residence is ancestral in Austronesian societies. *Proc. R. Soc. B* **276,** 1957–1964 (2009).
12 Currie, T. E., Greenhill, S. J., Gray, R. D., Hasegawa, T. & Mace, R. Rise and fall of political complexity in island south-east Asia and the Pacific. *Nature* **467,** 801–804 (2010).
13 Bellwood, P. *The First Farmers: The Origins of Agricultural Societies* (Blackwell Publishing, Oxford, 2005).
14 Oppenheimer, S. *Eden in the East: The Drowned Continent of Southeast Asia* (Phoenix, London, 1999).
15 Oppenheimer, S. J. & Richards, M. Polynesian origins. Slow boat to Melanesia? *Nature* **410,** 166–167 (2001).
16 Cox, M. P. Indonesian mitochondrial DNA and its opposition to a Pleistocene era origin of proto-Polynesians in Island Southeast Asia. *Hum. Biol.* **77,** 179–188 (2005).
17 Cox, M. P. The genetic environment of Melanesia: Clines, clusters and contact. In *Population Genetics Research Progress* (ed. Koven, V. T.) 45–83 (Nova Science Publishers, New York, 2008).
18 Cox, M. P. & Lahr, M. M. Y-Chromosome diversity is inversely associated with language affiliation in paired Austronesian- and Papuan-speaking communities from Solomon Islands. *Am. J. Hum. Biol.* **18,** 35–50 (2006).
19 Cox, M. P., Redd, A. J., Karafet, T. M., Ponder, C. A., Lansing, J. S., Sudoyo, H. *et al.* A Polynesian motif on the Y chromosome: population structure in Remote Oceania. *Hum. Biol.* **79,** 525–535 (2007).
20 Karafet, T. M., Hallmark, B., Cox, M. P., Sudoyo, H., Downey, S. S., Lansing, J. S. *et al.* Major east-west division underlies Y chromosome stratification across Indonesia. *Mol. Biol. Evol.* **27,** 1833–1844 (2010).
21 Karafet, T. M., Lansing, J. S., Redd, A. J., Reznikova, S., Watkins, J. C., Surata, S. P. *et al.* Balinese Y-chromosome perspective on the peopling of Indonesia: genetic contributions from pre-Neolithic hunter-gatherers, Austronesian farmers, and Indian traders. *Hum. Biol.* **77,** 93–114 (2005).
22 Kayser, M. The human genetic history of Oceania: near and remote views of dispersal. *Curr. Biol.* **20,** R194–R201 (2010).
23 Kayser, M., Brauer, S., Cordaux, R., Casto, A., Lao, O., Zhivotovsky, L. A. *et al.* Melanesian and Asian origins of Polynesians: mtDNA and Y-chromosome gradients across the Pacific. *Mol. Biol. Evol.* **23,** 2234–2244 (2006).
24 Kayser, M., Brauer, S., Weiss, G., Schiefenhövel, W., Underhill, P., Shen, P. *et al.* Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *Am. J. Hum. Genet.* **72,** 281–302 (2003).

25 Hill, C., Soares, P., Mormina, M., Macaulay, V., Clarke, D., Blumbach, P. B. *et al.* A mitochondrial stratigraphy for island southeast Asia. *Am. J. Hum. Genet.* **80,** 29–43 (2007).
26 Soares, P., Trejaut, J. A., Loo, J.-H., Hill, C., Mormina, M., Lee, C.-L. *et al.* Climate change and postglacial human dispersals in Southeast Asia. *Mol. Biol. Evol.* **25,** 1209–1218 (2008).
27 Cox, M. P., Karafet, T. M., Lansing, J. S., Sudoyo, H. & Hammer, M. F. Autosomal and X-linked single nucleotide polymorphisms reveal a steep Asian-Melanesian ancestry cline in eastern Indonesia and a sex bias in admixture rates. *Proc. R. Soc. B* **277,** 1589–1596 (2010).
28 HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. *Science* **326,** 1541–1545 (2009).
29 Wollstein, A., Lao, O., Becker, C., Brauer, S., Trent, R. J., Nürnberg, P. *et al.* Demographic history of Oceania inferred from genome-wide data. *Curr. Biol.* **20,** 1983–1992 (2010).
30 Yao, Y. G., Kong, Q. P., Bandelt, H. J., Kivisild, T. & Zhang, Y. P. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am. J. Hum. Genet.* **70,** 635–651 (2002).
31 Allard, M. W., Wilson, M. R., Monson, K. L. & Budowle, B. Control region sequences for East Asian individuals in the Scientific Working Group on DNA Analysis Methods forensic mtDNA data set. *Leg. Med.* **6,** 11–24 (2004).
32 Li, H., Cai, X., Winograd-Cort, E. R., Wen, B., Cheng, X., Qin, Z. *et al.* Mitochondrial DNA diversity and population differentiation in southern East Asia. *Am. J. Phys. Anthropol.* **134,** 481–488 (2007).
33 Hill, C., Soares, P., Mormina, M., Macaulay, V., Meehan, W., Blackburn, J. *et al.* Phylogeography and ethnogenesis of aboriginal southeast Asians. *Mol. Biol. Evol.* **23,** 2480–2491 (2006).
34 Trejaut, J. A., Kivisild, T., Loo, J. H., Lee, C. L., He, C. L., Hsu, C. J. *et al.* Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. *PLoS Biol.* **3,** e247 (2005).
35 Friedlaender, J. S., Friedlaender, F. R., Hodgson, J. A., Stoltz, M., Koki, G., Horvat, G. *et al.* Melanesian mtDNA complexity. *PLoS ONE* **2,** e248 (2007).
36 Tabbada, K. A., Trejaut, J., Loo, J.-H., Chen, Y.-M., Lin, M., Mirazón-Lahr, M. *et al.* Philippine mitochondrial DNA diversity: a populated viaduct between Taiwan and Indonesia? *Mol. Biol. Evol.* **27,** 21–31 (2010).
37 Miller, S. A., Dykes, D. D. & Polesky, H. F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* **16,** 1215 (1988).
38 Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M. & Howell, N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23,** 147 (1999).
39 Tanaka, M., Cabrera, V. M., Gonzalez, A. M., Larruga, J. M., Takeyasu, T., Fuku, N. *et al.* Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res.* **14,** 1832–1850 (2004).
40 Kong, Q. P., Yao, Y. G., Sun, C., Bandelt, H. J., Zhu, C. L. & Zhang, Y. P. Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. *Am. J. Hum. Genet.* **73,** 671–676 (2003).
41 Friedlaender, J., Gentz, F., Friedlaender, F., Kaestle, F., Schurr, T., Koki, G. *et al.* Mitochondrial genetic diversity and its determinants in Island Melanesia. In *Papuan Pasts: Cultural, Linguistic and Biological Histories of Papuan-Speaking Peoples* (eds Pawley, A., Attenborough, R., Golson, J. & Hide, R.) 693–716 (Pacific Linguistics, Canberra, 2005).
42 van Oven, M. Revision of the mtDNA tree and corresponding haplogroup nomenclature. *PNAS* **107,** E38–E39 (2010).
43 Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Res.* **10,** 564–567 (2010).
44 Slatkin, M. Inbreeding coefficients and coalescence times. *Genet. Res. Camb.* **58,** 167–175 (1991).
45 Slatkin, M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139,** 457–462 (1995).
46 R Development Core Team. R: A Language and Environment for Statistical Computing (2012). http://www.r-project.org/.
47 Bandelt, H. J., Forster, P. & Röhl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16,** 37–48 (1999).
48 Forster, P., Harding, R., Torroni, A. & Bandelt, H.-J. Origin and evolution of native American mtDNA variation: a reappraisal. *Am. J. Hum. Genet.* **59,** 935–945 (1996).
49 Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A. *et al.* Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am. J. Hum. Genet.* **84,** 740–759 (2009).
50 Cox, M. P. Accuracy of molecular dating with the rho statistic: deviations from coalescent expectations under a range of demographic models. *Hum. Biol.* **80,** 335–357 (2008).
51 Hedrick, P. W. A standardized genetic differentiation measure. *Evolution* **59,** 1633–1638 (2005).
52 van Oven, M., Hämmerle, J. M., van Schoor, M., Kushnick, G., Pennekamp, P., Zega, I. *et al.* Unexpected island effects at an extreme: reduced Y-chromosome and mitochondrial DNA diversity in Nias. *Mol. Biol. Evol.* **28,** 1349–1361 (2011).
53 Cox, M. P., Nelson, M. G., Tumonggor, M. K., Ricaut, F. X. & Sudoyo, H. A small cohort of Island Southeast Asian women founded Madagascar. *Proc. R. Soc. B* **279,** 2761–2768 (2012).
54 Razafindrazaka, H., Ricaut, F.-X., Cox, M. P., Mormina, M., Dugoujon, J.-M., Randriamarolaza, L. P. *et al.* Complete mitochondrial DNA sequences provide new insights into the Polynesian motif and the peopling of Madagascar. *Eur. J. Hum. Genet.* **18,** 575–581 (2010).
55 Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W. *et al.* Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* **308,** 1034–1036 (2005).
56 Dupuy, B. M., Stenersen, M., Egeland, T. & Olaisen, B. Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. *Hum. Mutat.* **23,** 117–124 (2004).
57 Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128,** 415–423 (2005).
58 Gusmão, L., Sánchez-Diz, P., Calafell, F., Martín, P., Alonso, C. A., Álvarez-Fernández, F. *et al.* Mutation rates at Y chromosome specific microsatellites. *Hum. Mutat.* **26,** 520–528 (2005).
59 Zhivotovsky, L. A., Underhill, P. A., Cinnioğlu, C., Kayser, M., Morar, B., Kivisild, T. *et al.* The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am. J. Hum. Genet.* **74,** 50–61 (2004).
60 Lansing, J. S., Karafet, T. M., Schoenfelder, J. & Hammer, M. F. A DNA signature for the expansion of irrigation in Bali? In *Past Human Migrations in East Asia: Matching Archaeology, Linguistics and Genetics* (eds Sanchez-Mazas, A., Blench, R., Ross, M. D., Peiros, I. & Lin, M.) 377–395 (Routledge, London, 2008).
61 Blust, R. *The Austronesian Languages.* Research School of Pacific and Asian Studies. (Australian National University, Canberra, Australia, 2009).
62 Oppenheimer, S. & Richards, M. Fast trains, slow boats, and the ancestry of the Polynesian Islanders. *Sci. Prog.* **84,** 157–181 (2001).
63 Jinam, T. A., Hong, L.-C., Phipps, M. E., Stoneking, M., Ameen, M., Edo, J. *et al.* Evolutionary history of continental southeast Asians: 'Early train' hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. *Mol. Biol. Evol* **29,** 3513–3527 (2012).
64 Soares, P., Rito, T., Trejaut, J., Mormina, M., Hill, C., Tinkler-Hundal, E. *et al.* Ancient voyaging and Polynesian origins. *Am. J. Hum. Genet.* **88,** 239–247 (2011).
65 Forster, P., Torroni, A., Renfrew, C. & Röhl, A. Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. *Mol. Biol. Evol.* **18,** 1864–1881 (2001).
66 Ross, M. Pronouns as a preliminary diagnostic of grouping Papuan languages. In *Papuan Pasts: Cultural, Linguistic and Biological Histories of Papuan-Speaking Peoples* (eds Pawley, A., Attenborough, R., Golson, J. & Hide, R.) 15–65 (Pacific Linguistics, Canberra, 2005).
67 Xu, S., Pugach, I., Stoneking, M., Kayser, M., Jin, L. & the HUGO Pan-Asian SNP Consortium. Genetic dating indicates that the Asian-Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. *Proc. Natl Acad. Sci. USA* **109,** 4574–4579 (2012).
68 Gray, R. D. & Jordan, F. M. Language trees support the express-train sequence of Austronesian expansion. *Nature* **405,** 1052–1054 (2000).
69 Hage, P. & Marck, J. Matrilineality and the Melanesian origin of Polynesian Y chromosomes. *Curr. Anthropol.* **44,** S121–S127 (2003).
70 Lansing, S. J., Cox, M. P., de Vet, T. A., Downey, S. S., Hallmark, B. & Sudoyo, H. An ongoing Austronesian expansion in Island Southeast Asia. *J. Anthropol. Archaeol.* **30,** 262–272 (2011).
71 Tibbetts, G. *A Study of the Arabic Texts Containing Material on South-East Asia* (E. J. Brill, Leiden, 1979).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)

6

A SMALL COHORT OF ISLAND SOUTHEAST ASIAN WOMEN FOUNDED MADAGASCAR

Cox, M.P., M.G. Nelson, M.K. Tumonggor, F.-X. Ricaut, and H. Sudoyo. 2012. A small cohort of Island Southeast Asian women founded Madagascar. *Proceedings of the Royal Society B* 279:2761-2768.
http://doi.org/10.1098/rspb.2012.0012

ABSTRACT The settlement of Madagascar is one of the most unusual, and least understood, episodes in human prehistory. Madagascar was one of the last landmasses to be reached by people, and despite the island's location just off the east coast of Africa, evidence from genetics, language and culture all attests that it was settled jointly by Africans, and more surprisingly, Indonesians. Nevertheless, extremely little is known about the settlement process itself. Here, we report broad geographical screening of Malagasy and Indonesian genetic variation, from which we infer a statistically robust coalescent model of the island's initial settlement. Maximum-likelihood estimates favour 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.

PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2012) **279**, 2761–2768
doi:10.1098/rspb.2012.0012
Published online 21 March 2012

A small cohort of Island Southeast Asian women founded Madagascar

Murray P. Cox[1,*], Michael G. Nelson[1], Meryanne K. Tumonggor[2], François-X. Ricaut[3] and Herawati Sudoyo[4]

[1]*Institute of Molecular BioSciences, Massey University, Private Bag 11 222, Palmerston North, New Zealand*
[2]*Department of Anthropology, University of Arizona, Tucson, AZ 85721, USA*
[3]*Laboratoire d'Anthropologie Moléculaire et Imagerie de Synthèse, Centre National de la Recherche Scientifique, Université de Toulouse (Paul Sabatier), Toulouse, France*
[4]*Eijkman Institute for Molecular Biology, Jakarta, Indonesia*

The settlement of Madagascar is one of the most unusual, and least understood, episodes in human prehistory. Madagascar was one of the last landmasses to be reached by people, and despite the island's location just off the east coast of Africa, evidence from genetics, language and culture all attests that it was settled jointly by Africans, and more surprisingly, Indonesians. Nevertheless, extremely little is known about the settlement process itself. Here, we report broad geographical screening of Malagasy and Indonesian genetic variation, from which we infer a statistically robust coalescent model of the island's initial settlement. Maximum-likelihood estimates favour 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.

Keywords: Madagascar; Indonesia; settlement; coalescent; mitochondrial DNA

1. INTRODUCTION

The colonization of Madagascar, an island nation located off the east coast of Africa, is one of the more unusual events in human prehistory. Malagasy, a term that denotes the island's people and their language, shows both biological [1] and linguistic [2] connections to east African populations. More surprisingly, there is unequivocal evidence of biological [3] and linguistic [4] links to Island Southeast Asia, specifically Indonesia. Although these dual African–Indonesian associations have been confirmed by numerous genetic, linguistic, ethnographic and archaeological studies, the process by which Madagascar was initially settled remains poorly understood [5]. Here, we tackle this question by simulating a genetic model of the initial settlement process, with a specific focus on the long-range Indonesian female contribution to modern Malagasy.

Archaeological and paleoenvironmental evidence shows that people settled Madagascar only recently. There may have been sporadic early arrivals from Africa with unknown connections to today's Malagasy [6,7], but a burst of continuous settlement activity is clearly in evidence around the middle of the first millennium AD [8–11]. Despite detailed historical documents for other Indian Ocean regions from at least the Roman era onwards [12], the first several centuries of Malagasy history passed by completely unrecorded. Madagascar and its people are not mentioned in any written source until the Arab geographer al-Idrīsī in AD 1165 [13]. Madagascar's extreme peripheral location on Indian Ocean trade routes may partly explain this historical silence; as late as the fifteenth century, the Arab cartographer ibn Mājid described Madagascar as 'bordering on the inhabited regions of the world' [13]. Just decades later, the Portuguese explorer Vasco da Gama sailed past Madagascar in 1497 on the first European voyage to India [14], thereby ushering in an era of detailed European record keeping on later Malagasy history.

In the absence of early written records, attempts to reconstruct the history of Madagascar necessarily rely on indirect evidence. All Malagasy today speak dialects of an Austronesian language that traces its origin back to Indonesia [4], a fact traditionally attributed to the Jesuit priest Luis Mariano in 1614 [15], but noted as early as the twelfth century by al-Idrīsī [13]. Most of the Malagasy lexicon is ultimately sourced from Ma'anyan, a language spoken along the Barito River valley of southeast Borneo [16–18]. However, only a small landlocked community of ethnic Dayak currently speaks Ma'anyan, and this inland region of southeast Borneo offers little ethnographic or historical evidence for any sort of maritime tradition. It is perhaps unsurprising then that Malagasy harbours far broader linguistic contributions drawn from across the Indonesian archipelago [2]. Loan words from Sanskrit, all with local linguistic modifications via Javanese or Malay [2,19,20], hint that Madagascar may have been colonized by settlers from the Srivijaya Empire, a major regional power in western Indonesia (modern Java, Sumatra and Malaysia) from the sixth to thirteenth centuries AD. Consistent with this hypothesis, Malagasy borrows large numbers of words from Javanese; the regional lingua franca, Malay; and even languages from southern Sulawesi, an island near the centre of the Indonesian archipelago. Cultural evidence—including iron working techniques, outrigger boats, musical instruments such as

* Author for correspondence (m.p.cox@massey.ac.nz).

Received 3 January 2012
Accepted 28 February 2012

the xylophone, and the cultivation of rice, bananas, yams and taro (i.e. a 'tropical food kit')—all supports a strong Southeast Asian connection from at least the eighth century onwards [21–23].

Genetic evidence paints much the same picture. Early blood protein studies identified dual African and Asian contributions [24,25]. Later studies of mitochondrial DNA (mtDNA) and Y chromosome variation improved this geographical resolution. Soodyall *et al.* [26,27] firmly established an Island Southeast Asian connection when they located the Polynesian motif in Madagascar. Apart from Madagascar, this mtDNA haplotype is restricted to remote Oceania, with low frequencies in Melanesia and eastern Indonesia, and only sporadic occurrences as far west as Bali and Borneo [28]. All sampled Malagasy are now known to carry a variant of the Polynesian motif, termed the Malagasy motif (characterized by polymorphisms 1473 and 3423A), whose global geographical range remains unclear [29]. In terms of male history, Y chromosome haplogroup O—otherwise restricted to Southeast Asia and Oceania—has been identified in Madagascar [3,5] and the Comoros Islands, a small archipelago just off Madagascar's northern tip [30]. Where genetic dating has been performed, these genetic connections to Indonesia are consistent with the temporal bounds suggested by archaeology and linguistics [5], although invariably with much larger CI.

In combination, a broad outline of Malagasy history—especially later periods—is beginning to emerge. However, we still lack any real understanding of the island's earliest history, particularly its initial colonization. What was the settlement process like? Was Madagascar colonized via large-scale population movements, or through a smaller translocation? How many people originally founded the Malagasy? What proportion of these settlers was Indonesian? And when did they first arrive? Here, we use new genetic data from Indonesian source populations, and a coalescent model specifically designed for the settlement of Madagascar to address these unresolved questions.

2. METHODS

(a) *Samples*

The Indonesian samples analysed in this study have been described extensively elsewhere [31–36]. In brief, 2745 individuals were screened from 12 island groups spanning the Indonesian archipelago: Sumatra ($n = 40$), Nias ($n = 62$), Mentawai ($n = 126$), Java ($n = 49$), Bali ($n = 517$), Sulawesi ($n = 188$), Sumba ($n = 639$), Flores ($n = 453$), Lembata ($n = 92$), Alor ($n = 23$), Pantar ($n = 27$) and Timor ($n = 529$). Samples were obtained with written informed consent by H.S., M.T., Golfiani Malik, Wuryantari Setiadi and Loa Helena Suryadi of the Eijkman Institute for Molecular Biology (Jakarta, Indonesia), and J. Stephen Lansing of the University of Arizona (Tucson, AZ, USA), with the assistance of Indonesian Public Health clinic staff. Sample collection followed protocols for the protection of human subjects established by both the Eijkman Institute and the University of Arizona institutional review boards. Permission to conduct research in Indonesia was granted by the Indonesian Institute of Sciences. All genetic screening was performed by M.T. at the University of Arizona.

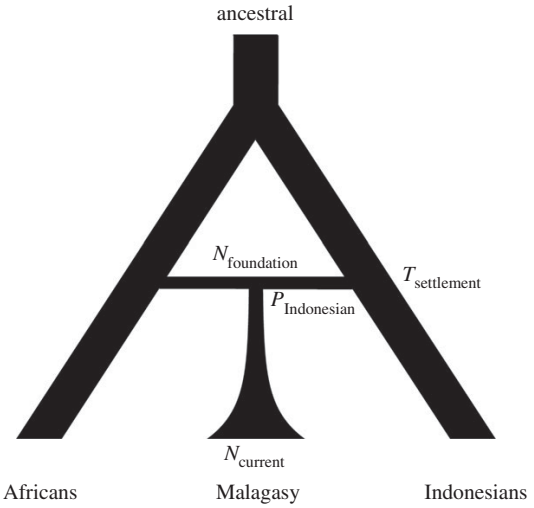

Figure 1. A demographic model of the settlement of Madagascar. An ancestral population splits into Africans and Indonesians. At some time, $T_{\text{settlement}}$, a proportion of the Indonesian population, $P_{\text{Indonesian}}$, and a proportion of the African population, $1 - P_{\text{Indonesian}}$, merge to form the founding population of Madagascar, $N_{\text{foundation}}$. This initial Malagasy population grows exponentially until it reaches its current effective population size, N_{current}.

(b) *Genetic markers*

Individuals carrying the Polynesian motif (haplogroup B4a1a1a) were identified by sequencing a 530 bp fragment of the mtDNA first hypervariable segment in all 2745 samples. All Polynesian motif carriers were screened for polymorphisms specific to the Malagasy motif, that is, restriction fragment length polymorphism (RFLP) losses at nucleotides 1473 (−1473 *Hha*I) and 3423A (−3423 *Aci*I). Single nucleotide polymorphism (SNP) genotyping was performed using the method of Razafindrazaka *et al.* [29].

(c) *Published data*

Genetic diversity among Malagasy populations was drawn from previously published research [29]. In brief, 266 individuals were screened from three Malagasy ethnic groups: Mikea hunter–gatherers ($n = 127$), semi-nomadic Vezo fishermen ($n = 101$) and the dominant Andriana Merina ethnic group ($n = 38$). A detailed description of this dataset and associated research findings is available in the study of Razafindrazaka *et al.* [29].

(d) *Demographic models*

A hierarchical modelling approach was applied to determine the demographic parameters of Malagasy settlement. First, the historical relationships between African and Indonesian populations were determined. We adopted the simulation framework of Wollstein *et al.* [37], in which an ancestral African population ($N_{\text{ancestral}}$) split at some time (T_{split}) to form modern Africans (N_{African}) and Indonesians ($N_{\text{Indonesian}}$). We adjusted the Wollstein model to accommodate an Indonesian sample drawn from across the Indonesian archipelago, rather than just the island of Borneo, and accepted the African component of their model as it stands. Second, this African–Indonesian framework was extended to incorporate the Malagasy (figure 1). At some time ($T_{\text{settlement}}$), the Malagasy population ($N_{\text{foundation}}$) is formed from some proportion of Indonesian ($P_{\text{Indonesian}}$) and African ($1 - P_{\text{Indonesian}}$) settlers.

This Malagasy deme subsequently grows exponentially until it reaches its current effective population size ($N_{current}$).

(e) *Coalescent simulation*

Coalescent datasets were generated using the software MS [38]. The entire mtDNA genome was simulated using two-phase mutational scaling. First, the 530 bp fragment of the control region was simulated using a mutation rate of 8.0×10^{-6} mutations per site per generation [39] corrected for a generation interval of 25 years [40]. Second, the coding region was simulated by applying a mutation rate scalar of 1.57 coding region mutations to every control region mutation [41]. In the coalescent framework, the mutation rate and physical length of a locus are interrelated. We assume that the mitochondrial genome is not subject to recombination—the dominant view held by the research community today [42]. In this situation, the same coalescent outcome can be obtained by linearly scaling the mutation rate and physical length (e.g. doubling the mutation rate, while halving the length). We employ this scaling rule to simulate the entire mitochondrial genome because existing coalescent software cannot simulate multiple fully linked genetic regions with different mutation rates, but are instead flexible with regard to locus length. The control region rate was applied to the entire mitochondrial genome, and the coding region scaled by 1.57 times the control region length to reconstruct the correct overall ratio of coding to control region mutations. Soares *et al.* [41] provide a detailed explanation of this scaling factor.

(f) *Data summaries*

Summary statistics was developed based on unusual aspects of the Indonesian and Malagasy mtDNA datasets. First, the number of segregating sites—S summarizes the total length of the genealogy [43] and is an unbiased estimator of the population mutation rate ($\theta = N_e\mu$). The number of segregating sites in the Indonesian population was set as an acceptance criterion to control for the population mutation rate (i.e. it prevents acceptances of unreasonably large or small effective sizes) [44]. Simulations were accepted if the control region portion of the simulated dataset contained (i) the same number of segregating sites as in the observed Indonesian sample ($S = 186$) or (ii) segregating sites within 5 per cent of this value.

Second, the Malagasy data exhibit an unusual and distinctive pattern of mtDNA polymorphism. The Malagasy motif is a close relative of the Polynesian motif, but differs from it by two polymorphisms. Because mtDNA mutation rates only predict two mutation substitutions every 5216 years (using the corrected mutation rate of Soares *et al.* [41]) and archaeology constrains the settlement history of Madagascar to less than 2000 years, finding the derived Malagasy motif with two substitutions relative to the ancestral Polynesian motif is statistically unusual. Simulations were therefore accepted only if they contained a haplotype in the Malagasy sample that differed from an ancestral Indonesian haplotype by two changes in the coding region, but no changes in the control region.

Third, simulations were accepted only if simulated allele frequencies mirrored observed allele frequencies. The 'ancestral' haplotype in the simulated Indonesian sample was required to have a frequency less than or equal to the Polynesian motif frequency in the observed Indonesian data (less than or equal to 0.02). Similarly, the 'derived' haplotype in the simulated Malagasy sample was required to have a frequency greater than or equal to the Malagasy motif frequency in the observed Malagasy data (greater than or equal to 0.22). Averaged across the Mikea, Vezo and Merina ethnic groups, this value is slightly lower than the B4a1a1a frequency observed for the Antandroy, Antanosy, Antaisaka and Merina (0.34) [5]. However, as Tofanelli *et al.* did not explicitly screen for the Malagasy motif, we chose to perform simulations using our more conservative estimate of the actual Malagasy motif frequency.

C++ code to calculate these summary statistics on MS coalescent simulations is available from the authors on request.

(g) *Inferential statistics*

Forty million coalescent simulations were run across a four-dimensional parameter space in the Malagasy settlement model. Values were drawn randomly from uniform distributions for $N_{foundation}$ (0, 2000), $N_{current}$ (0, 8041), $T_{settlement}$ (0, 5000 years) and $P_{Indonesian}$ (0, 1). Only simulations matching all summary criteria (described above) were retained. Because simulated parameters were drawn from random uniform distributions, accepted simulations form an unbiased estimate of the likelihood. The peaks (modes) of reported probability densities are maximum-likelihood estimates (MLEs), and were calculated using kernel density estimation via the density function in the base package of R [45]. For each demographic parameter, 95% CI were calculated as asymmetric [0.025, 0.975] quantiles of the observed probability density.

3. RESULTS

(a) *Malagasy motif in Madagascar*

The Polynesian motif (mtDNA haplotype B4a1a1a) occurs at variable frequency among the ethnic groups (*foko*) of Madagascar and has been shown to be a direct translocation from Island Southeast Asia [3,29]. In a 2010 study, the Polynesian motif was identified in 58 of 266 (22%) Malagasy individuals, although its frequency varied among three ethnic groups: 50 per cent in Merina, 22 per cent in Vezo and 13 per cent in Mikea [29]. Whole mtDNA genome sequencing found that the Polynesian motif in Madagascar is characterized by two additional mutations (1473 and 3423A), which have been identified in all Polynesian motif carriers screened in Madagascar so far [29]. For this reason, this new haplotype has been termed the Malagasy motif and occurs in Madagascar at an average frequency of 22 per cent.

(b) *Malagasy motif in Indonesia*

To identify the distribution of the Malagasy motif in Indonesia, the diagnostic markers of the Polynesian and Malagasy motifs were screened in a large number of individuals ($n = 2745$) from across the Indonesian archipelago (table 1). Only 45 of 2745 sampled individuals (2%) carried the Polynesian motif, specifically on the islands of Bali ($n = 6$), Sumba ($n = 4$), Alor ($n = 1$) and Timor ($n = 34$). This is consistent with previous results, which show the Polynesian motif reaching greatest frequency in remote Oceania, with much reduced frequencies in eastern Indonesia and only sporadic occurrences in western Indonesia [46].

The 45 Polynesian motif carriers were screened for the diagnostic markers of the Malagasy motif. None carried either of the two additional mutations (1473 and 3423A). This is consistent with published results of full mtDNA

Table 1. Allele frequencies of the Polynesian and Malagasy motifs across the Indonesian archipelago.

population	sample size	Polynesian motif (n, %)		Malagasy motif n, %	
Sumatra	40	0	—	0	—
Nias	62	0	—	0	—
Mentawai	126	0	—	0	—
Java	49	0	—	0	—
Bali	517	6	1.2	0	—
Sulawesi	188	0	—	0	—
Sumba	639	4	0.6	0	—
Flores	453	0	—	0	—
Lembata	92	0	—	0	—
Alor	23	1	4.4	0	—
Pantar	27	0	—	0	—
Timor	529	34	6.4	0	—
Total	2745	45	1.6	0	—

genome sequences; no Polynesian motif genomes outside Madagascar have been shown to carry the Malagasy motif [29,46]. We applied a Monte Carlo sampling error algorithm [31] to infer a possible upper bound on the allele frequency of the Malagasy motif across Indonesia as a whole. Even if present in Indonesia, the Malagasy motif is statistically unlikely to exceed a frequency of 0.1 per cent there. As this is only an inference, we do not use this value in our simulations. In its place, we focus on the concretely known frequency of the ancestral Polynesian motif in Indonesia. However, we conclude that (i) either the Malagasy motif arose in Madagascar and is therefore not present in Indonesia; or (ii) the Malagasy motif arose in Indonesia, but is either absent there or occurs only at an extremely low frequency there today.

(c) *Demographic inference*

A hierarchical series of coalescent models was developed to infer the demographic processes underpinning the initial settlement of Madagascar. Because this analysis is based on mtDNA evidence, we effectively report the history of Malagasy women. The inferential model was built around three especially unusual characteristics of the Malagasy and Indonesian mtDNA data: (i) the ancestral Polynesian motif occurs at very low frequency in Indonesia (this study); (ii) the derived Malagasy motif occurs at moderate-to-high frequency in Madagascar [29]; and (iii) the Malagasy motif differs from the Polynesian motif by two SNPs within the coding region [29]. Even given the relatively high rate of mtDNA mutations, two new polymorphisms would be expected to arise only once every approximately 5216 years [41]. Although this time frame is consistent with the estimated age of the ancestral Polynesian motif [46], it still far exceeds the colonization horizon of approximately 1500 years attested by archaeological and linguistic records for the settlement of Madagascar. Therefore, we set out to determine how often this unusual pattern of mtDNA diversity might occur under various demographic parametrizations of the settlement of Madagascar.

(d) *African–Indonesian framework*

To determine the demographic parameters of a Malagasy settlement model, we first had to formulate a basic demographic framework relating Africans and Indonesians, the parental groups of modern Malagasy. This relationship was recently inferred using coalescent simulations, and is described in great detail by Wollstein *et al.* [37]. Rather than duplicate their efforts, we accepted this inferred demographic model with one key exception. The Wollstein model was specifically inferred for a single population sample from Borneo, whereas our Indonesian sample was drawn from across the archipelago. We modified the Wollstein model to accommodate this difference. Coalescent datasets were simulated under a range of effective population sizes for Indonesia (as opposed to just Borneo). The same number of segregating sites as observed in the real Indonesian dataset ($S = 186$) was chosen as the acceptance criterion. The MLE for the effective population size of Indonesians was 4637 (95% CI: 2435–6823), which slightly exceeds the estimate of Wollstein *et al.* for the island of Borneo (4034; 95% CI: 1993–6497). Because we are sampling a larger geographical range of genetic diversity, the effective size of Indonesians is expected to be larger than the effective size of the single Borneo population by itself. This pan-Indonesian effective size estimate was applied in all subsequent modelling.

(e) *Malagasy settlement model*

The African–Indonesian framework was extended to accommodate the key aspects of Malagasy settlement (figure 1). In this more complex model, the Malagasy population is formed through admixture between African and Indonesian sources. At some time ($T_{\text{settlement}}$), the Malagasy population ($N_{\text{foundation}}$) is formed from some proportion of Indonesian ($P_{\text{indonesian}}$) and African ($1 - P_{\text{Indonesian}}$) settlers. The founding population subsequently expands until it reaches its current effective population size (N_{current}). Under this inferential model, we simulated coalescent datasets representing entire mtDNA genomes, which were accepted only if they possessed the key distinguishing characteristics of the Malagasy/Indonesian mtDNA dataset: (i) an ancestral haplotype ('Polynesian motif') in Indonesia with frequency less than or equal to 2 per cent; (ii) a derived haplotype ('Malagasy motif') in Madagascar with frequency greater than or equal to 22 per cent; (iii) ancestral and derived haplotypes that show no variation in the control region; (iv) ancestral and derived haplotypes that differ by two mutations in the coding region; and (v) 186 segregating sites in just the control region portion of the simulated Indonesian dataset.

Coalescent simulations were run on an eight-core UNIX cluster for 11 520 CPU hours (i.e. equivalent to one computer running for approx. 1.3 years). Only 671 simulations (0.0017%) were returned under all five acceptance criteria. To expand the acceptance rate, the segregating sites' criterion was extended to accommodate 5 per cent variance in the observed value ($S = 186 \pm 5\% = 177$–195). Under this more relaxed standard, 12 250 simulations (0.03%) were returned under all five acceptance criteria. As the two sets of results showed no substantive differences, demographic parameters are reported for the larger dataset.

Likelihood curves were plotted for all four demographic parameters in the Malagasy settlement model: foundation effective population size, current effective population size, time of settlement and proportion contribution from Indonesia (figure 2). MLEs and 95% CI are

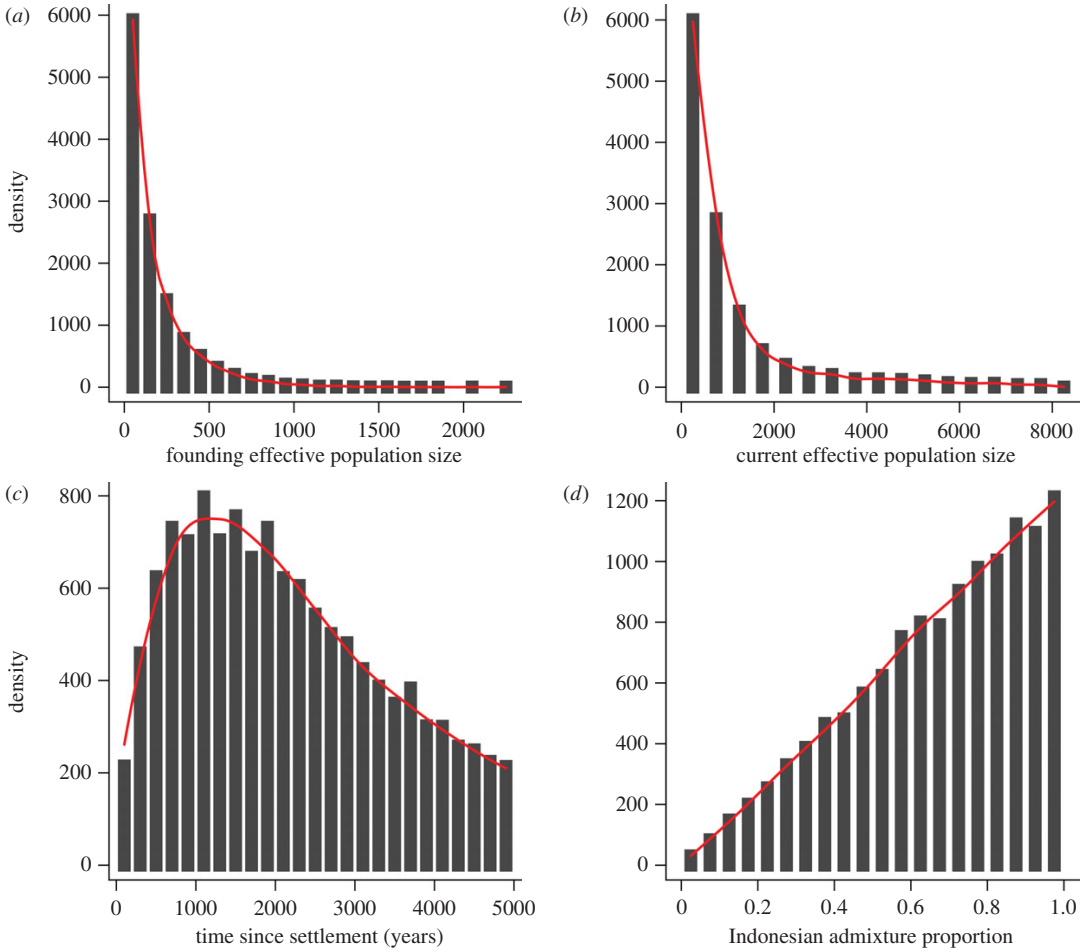

Figure 2. Likelihood curves inferred for demographic parameters in the Malagasy settlement model: (*a*) founding effective population size; (*b*) current effective population size; (*c*) time since settlement (years); (*d*) proportion of the founding population with Indonesian ancestry. Grey bars indicate the distribution of accepted Monte Carlo data points; red lines plot local regressions to the probability densities.

Table 2. Maximum-likelihood estimates (MLEs) and 95% CI of demographic parameters inferred for the Malagasy settlement model.

demographic parameter	label	MLE	95% CI
founding effective population size	$N_{\text{foundation}}$	30	3–746
current effective population size	N_{current}	193	29–5452
time since settlement (years)	$T_{\text{settlement}}$	1170	255–4721
Indonesian admixture proportion	$P_{\text{Indonesian}}$	0.93	0.16–0.99

reported in table 2. The most probable set of parameter values indicates that Madagascar was settled 1170 years ago (95% CI: 255–4721 years ago). The coalescent analysis infers that a very small number of women founded the Malagasy population—the best estimate is 30 (95% CI: 3–746). The proportion of this founding population having Indonesian ancestry was very high—0.93 (95% CI: 0.16–0.99). The Malagasy population subsequently grew approximately sixfold to reach its current effective population size (193; 95% CI: 29–5452). This effective size is very small compared with other globally distributed populations [47], but fits well with the fact that many Malagasy carry an identical mtDNA sequence (i.e. the Malagasy motif).

4. DISCUSSION

An extensive body of genetic, linguistic and anthropological research has now painted a picture of Malagasy history in broad brushstrokes. Humans reached Madagascar only relatively recently. In fact, Madagascar was among the last places on Earth to be settled—archaeological evidence suggests that the first sustained colonization did not occur until the middle of the first millennium AD. Today, Malagasy show clear evidence of genetic and linguistic contributions from Africa. Loanwords suggest contact with the Bantu-speaking peoples of east Africa [2], but the exact origin of these African settlers remains largely unknown. Contact with populations north of the Zambezi River is typically favoured [5,26]. Complementing this African connection, evidence

from multiple fields of study shows clear and substantial contributions from the Indonesian archipelago. The exact geographical provenance of this contribution is again unclear—arguments have been advanced for both west Indonesia (Ma'anyan language; Sanskrit, Malay and Javanese loanwords) and east Indonesia (South Sulawesi loanwords; prevalence of the Polynesian motif). The presence of Indonesian mtDNA and Y chromosomes in Madagascar argues for a mixed-sex founder population, and this Indonesian genetic component reaches substantial frequencies among Malagasy even today. Similarly, because African and Indonesian lineages differ little across all Malagasy ethnic groups screened to date, the main episode of African–Indonesian admixture likely occurred at the very beginning of Malagasy history.

Key detailing was recently added to this picture. The Polynesian motif in Madagascar carries two derived polymorphisms (1473 and 3423A), and this new haplotype—termed the Malagasy motif—is present in all Polynesian motif carriers screened in Madagascar so far [29]. However, the distribution of the Malagasy motif outside Madagascar has been unclear. Here, we show that this lineage is not widely distributed in Indonesia. Indeed, the Malagasy motif was not detected in 2745 individuals drawn from across the Indonesian archipelago—the largest Indonesian population dataset analysed to date. Even the Polynesian motif, the ancestral lineage of the Malagasy motif, is rare in Indonesia (2%), and with sporadic exceptions, this haplotype is restricted to eastern parts of the archipelago. It remains possible that the Malagasy motif may yet be found in a few small Indonesian communities, perhaps on the relatively understudied islands of Borneo and Sulawesi. Broader sampling in these regions will eventually answer this question. Alternately, the Malagasy motif may not be present in Indonesia at all. Perhaps it arose among the earliest Indonesian colonists to Madagascar, subsequently sweeping to high frequency either in Madagascar or somewhere along the path of their travel (e.g. putative stopping points along the east African coast). In either case, these scenarios hint at a substantial Indonesian founder event during the main settlement period in Madagascar, with strong genetic drift raising the frequency of the Malagasy motif to the elevated levels observed in Madagascar today.

To infer the parameters of this settlement model in more detail, a statistical simulation model was built around the unusual genetic patterns observed in the Malagasy/Indonesian mtDNA dataset: (i) a low frequency of the ancestral Polynesian motif lineage in Indonesia; (ii) two polymorphisms distinguishing the Polynesian and Malagasy motifs; and (iii) a high frequency of the derived Malagasy motif in Madagascar. We set out to ask: what range of settlement model parameters produce genetic patterns like the ones we observe?

Our coalescent modelling suggests several answers to this question. First, although the observed pattern of genetic diversity is relatively uncommon, it occurs more frequently under certain demographic parameters. Coalescent simulations best support settlement of Madagascar beginning around AD 830. This date is consistent with evidence from linguistics, which links the colonization of Madagascar to the expansion of Indonesian trading networks during the Srivijaya Empire [2]. Srivijaya reached its peak in the ninth century, but remained a major maritime power in the Indian Ocean until well into the thirteenth century, consistent with the time frame suggested by our simulations. Our estimates also support a model in which Madagascar was settled by a small effective founding population—estimated at only approximately 30 women, most of whom had Indonesian ancestry (93%). Although this number of founding women might seem surprisingly small, it fits well with estimates of the small number of women (approx. 70) who founded New Zealand, another island nation settled by related Austronesian speakers at around the same time period [48]. In contrast with the simulations of Tofanelli *et al.* [5], our more complex inferential modelling shows that founder events and drift dynamics are sufficient to explain the frequencies of the Polynesian and Malagasy motifs in Indonesia and Madagascar. Tofanelli *et al.* employed a one-deme model, simulated only a small range of founding population sizes and fixed (rather than inferred) the Malagasy growth rate. Although they simulated genetic drift, it is not clear that they explicitly modelled the Malagasy founder event. In comparison, our multi-deme simulation framework is far more exact about the colonization event itself. Our highest likelihood model describes a settlement process in which relatively few women, most travelling from Indonesia, founded the Malagasy population—with a much smaller, but just as important, biological contribution from Africa.

This is the first genetic analysis where statistical bounds have been placed on the demographic parameters of Malagasy settlement. The most likely model favours a small founding population, which brings into question the broader context in which the settlement of Madagascar took place. A recurring hypothesis is that Indonesian maritime traders initially settled Madagascar, either as a single colonization event or via repeated settlement waves from the same source population (a process known to the Malagasy as *ranto*) [2,49,50]. Indeed, merchants have plied coastal Indian Ocean trade routes between east Africa and northern China at least since the Roman era [12]. However, early written records imply that these trading voyages were dominated by men; there is no mention of women on board long-distance trading vessels [13]. There is no clear evidence suggesting that Madagascar was settled in multiple waves, but because *ranto* is mentioned in traditional Malagasy narratives, this may be a profitable direction for future simulations.

An alternative hypothesis is that Madagascar was settled as a formal trading colony, perhaps under the auspices of the Srivijaya Empire (although Malagasy are not Hindu today), and possibly resembling later Arab trading centres in the region. Or perhaps Madagascar was settled as an ad hoc centre for refugees, drawn from those who lost land and power during the rapid expansion of Srivijayan influence. Such colonies would be established to be self-sufficient, and therefore, might be expected to include Indonesian women. However, there is little evidence—historical, archaeological or biological—of other Indonesian bases around the Indian Ocean, including the east African coast, although there may have been a Malay trading post in what is now Sri Lanka [50]. Founding such a centre in any official capacity on Madagascar—at the far extreme of Indonesia's trading reach—therefore seems out of keeping with contemporary Indonesian trading practices. Along related lines, historical documents leave no record of refugees fleeing the Srivijaya Empire, although early

chroniclers seldom paid attention to powerless groups as these refugees would likely have been.

A third hypothesis is that Madagascar was settled via a direct sailing route across the Indian Ocean, perhaps even as the result of an unintended transoceanic voyage. This view is traditionally considered unlikely, but has recently been revived based on seafaring simulations using ocean currents and monsoon weather patterns [51]. Indeed, during the Second World War, wreckage from ships bombed in the vicinity of Sumatra and Java later washed up in Madagascar, including—in one instance—a survivor in a lifeboat [52]. Cargo ships were substantial vessels during the first millennium AD—up to 500 tonnes and manned by over 100 sailors [53,54]. It is therefore not beyond the realms of possibility that a single wayward vessel might have effected the settlement of Madagascar. This would certainly be consistent with the extremely small initial size of the Malagasy population, although perhaps not with the mixed-sex founding group suggested by Malagasy genetics.

Without further evidence, the exact nature of Madagascar's settlement must necessarily remain unresolved. Nevertheless, we can now place the settlement process on a much firmer statistical footing. Malagasy are the children of both east and west, with clear Indonesian and African antecedents. Madagascar was settled approximately 1200 years ago, primarily by a small cohort of Indonesian women, and this Indonesian contribution—of language, culture and genes—continues to dominate the nation of Madagascar even today.

We thank Golfiani Malik, Wuryantari Setiadi and Loa Helena Suryadi of the Eijkman Institute for Molecular Biology (Jakarta, Indonesia), and J. Stephen Lansing of the University of Arizona (Tucson, AZ, USA), together with Indonesian Public Health clinic staff, for sample collection. We also thank Harilanto Razafindrazaka (University of Toulouse) for helpful discussion. Research was supported by an Early Career Medal from Massey University and a Rutherford Fellowship from the Royal Society of New Zealand (RDF-10-MAU-001) awarded to M.P.C.

REFERENCES

1 Ferrand, G. 1908 L'origine africaine des Malgaches [The African origin of the Malagasy]. *J Asiat.* **10**, 353–500.
2 Adelaar, A. 2009 Loanwords in Malagasy. In *Loanwords in the world's languages: a comparative handbook* (eds M. Haspelmath & U. Tadmor), pp. 717–746. Berlin, Germany: De Gruyter Mouton.
3 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. (doi:10.1086/430051)
4 Serva, M., Petroni, F., Volchenkov, D. & Wichmann, S. 2011 Malagasy dialects and the peopling of Madagascar. *J. R. Soc. Interface* **9**, 54–67. (doi:10.1098/rsif.2011.0228)
5 Tofanelli, S., Bertoncini, S., Castrì, 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. (doi:10.1093/molbev/msp120)
6 Burney, D. A., Burney, L. P., Godfrey, L. R., Jungers, W. L., Goodman, S. M., Wright, H. T. & Jull, A. J. T. 2004 A chronology for late prehistoric Madagascar. *J. Hum. Evol.* **47**, 25–63. (doi:10.1016/j.jhevol.2004.05.005)
7 Gommery, D., Ramanivosoa, B., Faure, M., Guérin, C., Kerloc'h, P., Sénégas, F. & Randrianantenaina, H. 2011 Les plus anciennes traces d'activités anthropiques de Madagascar sur des ossements d'hippopotames subfossiles d'Anjohibe (Province de Mahajanga) [The oldest traces of human activity on Madagascar in the subfossil bones of hippopotamus in Anjohibe (Mahajanga Province)]. *C.R. Palevol.* **10**, 271–278. (doi:10.1016/j.crpv.2011.01.006)
8 Burney, D. A. 1987 Late Holocene vegetational change in central Madagascar. *Quat. Res.* **28**, 130–143. (doi:10.1016/0033-5894(87)90038-X)
9 Burney, D. A. 1993 Late Holocene environmental changes in arid southwestern Madagascar. *Quat. Res.* **40**, 98–106. (doi:10.1006/qres.1993.1060)
10 MacPhee, R. D. E. & Burney, D. A. 1991 Dating of modified femora of extinct dwarf Hippopotamus from southern Madagascar: implications for constraining human colonization and vertebrate extinction events. *J. Archaeol. Sci.* **18**, 695–706. (doi:10.1016/0305-4403(91)90030-S)
11 Wright, H. T., Rakotoarisoa, J. A., Heurtebize, G. & Vérin, P. 1993 The evolution of settlement system in the Efaho River Valley: a preliminary report on archaeological reconnaissances of 1983–86. *Bull. Indo-Pacific Prehistory Assoc.* **12**, 2–20.
12 Casson, L. 1989 *The periplus Maris Erythræi*. Princeton, NJ: Princeton University Press.
13 Tibbetts, G. 1979 *A study of the Arabic texts containing material on South-East Asia*. Leiden, The Netherlands: E.J. Brill.
14 Ravenstein, E. G. 1898 *The journal of the first voyage of Vasco da Gama, 1497–1499*. London, UK: Hakluyt Society.
15 Mariano, L. 1904 Relation du voyage de découverte fait à l'Ile Saint Laurent dans les années 1613–1614 [Account of the voyage of discovery made on the island of Saint Laurent in the years 1613–1614]. In *Collection des ouvrages anciens concernant Madagascar* (eds A. Grandidier & G. Grandidier), pp. 1–64. Paris, France: Comité de Madagascar.
16 Dahl, O. C. 1951 *Malgache et Maanyan: Une comparison linguistique* [*Malagasy and Ma'anyan: A linguistic comparison*]. Oslo, Norway: Egede Instituttet.
17 Dahl, O. C. 1977 La subdivision de la famille Barito et la place du Malgache [The subdivision of the Barito family and the place of Malagasy]. *Acta Orient.* **38**, 77–134.
18 Dahl, O. C. 1991 *Migration from Kalimantan to Madagascar*. Oslo, Norway: The Institute for Comparative Research in Human Culture, Norwegian University Press.
19 Adelaar, K. A. 1989 Malay influence on Malagasy: linguistic and culture-historical implications. *Ocean Linguist.* **28**, 1–46. (doi:10.2307/3622973)
20 Adelaar, K. A. 1995 Asian roots of the Malagasy: a linguistic perspective. *Bijdragen tot de Taal-, Land- en Volkenkunde.* **151**, 325–356.
21 Wright, H. T. & Rakotoarisoa, J.-A. 1997 Cultural transformations and their impacts on the environments of Madagascar. In *Natural change and human impact in Madagascar* (eds S. M. Goodman & B. D. Patterson), pp. 309–330. Washington DC: Smithsonian Institute Press.
22 Blench, R. M. 2010 Evidence for the Austronesian voyages in the Indian Ocean. In *The global origins and development of seafaring* (eds A. J. Anderson, J. H. Barrett & K. V. Boyle), pp. 239–248. Cambridge, UK: MacDonald Institute.
23 Perrier, X. *et al.* 2011 Multidisciplinary perspectives on banana (*Musa* spp.) domestication. *Proc. Natl Acad. Sci. USA* **108**, 11 311–11 318. (doi:10.1073/pnas.1102001108)
24 Fourquet, R., Sarthou, J., Roux, J. & Aori, K. 1974 Hémoglobine S et origines du peuplement de Madagascar:

nouvelle hypothèse sur son introduction en Afrique [Hemoglobin S and origins for the settlement of Madagascar: new hypothesis on its introduction to Africa]. *Arch. Inst. Pasteur Madagascar* **43**, 185–220.
25 Hewitt, R., Krause, A., Goldman, A., Campbell, G. & Jenkins, T. 1996 β-globin haplotype analysis suggests that a major source of Malagasy ancestry is derived from Bantu-speaking Negroids. *Am. J. Hum. Genet.* **58**, 1303–1308.
26 Soodyall, H., Jenkins, T., Hewitt, R., Krause, A. & Stoneking, M. 1996 The peopling of Madagascar. In *Molecular biology and human diversity* (eds A. Boyce & C. Mascie-Taylor), pp. 156–170. Cambridge, UK: Cambridge University Press.
27 Soodyall, H., Jenkins, T. & Stoneking, M. 1995 'Polynesian' mtDNA in the Malagasy. *Nat. Genet.* **10**, 377–378. (doi:10.1038/ng0895-377)
28 Cox, M. P. 2005 Indonesian mitochondrial DNA and its opposition to a Pleistocene era origin of proto-Polynesians in Island Southeast Asia. *Hum. Biol.* **77**, 179–188. (doi:10.1353/hub.2005.0037)
29 Razafindrazaka, H. *et al.* 2010 Complete mitochondrial DNA sequences provide new insights into the Polynesian motif and the peopling of Madagascar. *Eur. J. Hum. Genet.* **18**, 575–581. (doi:10.1038/ejhg.2009.222)
30 Msaidie, S., Ducourneau, A., Boetsch, G., Longepied, G., Papa, K., Allibert, C., Yahaya, A. A., Chiaroni, J. & Mitchell, M. J. 2011 Genetic diversity on the Comoros Islands shows early seafaring as major determinant of human biocultural evolution in the Western Indian Ocean. *Eur. J. Hum. Genet.* **19**, 89–94. (doi:10.1038/ejhg.2010.128)
31 Cox, M. P., Karafet, T. M., Lansing, J. S., Sudoyo, H. & Hammer, M. F. 2010 Autosomal and X-linked single nucleotide polymorphisms reveal a steep Asian-Melanesian ancestry cline in eastern Indonesia and a sex bias in admixture rates. *Proc. R. Soc. B* **277**, 1589–1596. (doi:10.1098/rspb.2009.2041)
32 Karafet, T. M., Hallmark, B., Cox, M. P., Sudoyo, H., Downey, S. S., Lansing, J. S. & Hammer, M. F. 2010 Major east-west division underlies Y chromosome stratification across Indonesia. *Mol. Biol. Evol.* **27**, 1833–1844. (doi:10.1093/molbev/msq063)
33 Karafet, T. M. *et al.* 2005 Balinese Y-chromosome perspective on the peopling of Indonesia: genetic contributions from pre-Neolithic hunter-gatherers, Austronesian farmers, and Indian traders. *Hum. Biol.* **77**, 93–114. (doi:10.1353/hub.2005.0030)
34 Lansing, J. S. *et al.* 2007 Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc. Natl Acad. Sci. USA* **104**, 16 022–16 026. (doi:10.1073/pnas.0704451104)
35 Lansing, J. S., Watkins, J. C., Hallmark, B., Cox, M. P., Karafet, T. M., Sudoyo, H. & Hammer, M. F. 2008 Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations. *Proc. Natl Acad. Sci. USA* **105**, 11 645–11 650. (doi:10.1073/pnas.0710158105)
36 Lansing, S. J., Cox, M. P., de Vet, T. A., Downey, S. S., Hallmark, B. & Sudoyo, H. 2011 An ongoing Austronesian expansion in Island Southeast Asia. *J. Anthropol. Archaeol.* **30**, 262–272. (doi:10.1016/j.jaa.2011.06.004)
37 Wollstein, A., Lao, O., Becker, C., Brauer, S., Trent, R. J., Nürnberg, P., Stoneking, M. & Kayser, M. 2010 Demographic history of Oceania inferred from genome-wide data. *Curr. Biol.* **20**, 1983–1992. (doi:10.1016/j.cub.2010.10.040)
38 Hudson, R. R. 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics.* **18**, 337–338. (doi:10.1093/bioinformatics/18.2.337)
39 Sigurðardóttir, S., Helgason, A., Gulcher, J. R., Stefansson, K. & Donnelly, P. 2000 The mutation rate in the human mtDNA control region. *Am. J. Hum. Genet.* **66**, 1599–1609. (doi:10.1086/302902)
40 Fenner, J. N. 2005 Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423. (doi:10.1002/ajpa.20188)
41 Soares, P. *et al.* 2009 Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am. J. Hum. Genet.* **84**, 740–759. (doi:10.1016/j.ajhg.2009.05.001)
42 Galtier, N., Nabholz, B., Glémin, S. & Hurst, G. D. D. 2009 Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Mol. Ecol.* **18**, 4541–4550. (doi:10.1111/j.1365-294X.2009.04380.x)
43 Watterson, G. A. 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276. (doi:10.1016/0040-5809(75) 90020-9)
44 Cox, M. P., Morales, D. A., Woerner, A. E., Sozanski, J., Wall, J. D. & Hammer, M. F. 2009 Autosomal resequence data reveal Late Stone Age signals of population expansion in sub-Saharan African foraging and farming populations. *PLoS ONE* **4**, e6366. (doi:10.1371/journal.pone.0006366)
45 R Development Core Team. 2012 R: a language and environment for statistical computing. See http://www.r-project.org.
46 Soares, P. *et al.* 2011 Ancient voyaging and Polynesian origins. *Am. J. Hum. Genet.* **88**, 239–247. (doi:10.1016/j.ajhg.2011.01.009)
47 Cox, M. P., Woerner, A. E., Wall, J. D. & Hammer, M. F. 2008 Intergenic DNA sequences from the human X chromosome reveal high rates of global gene flow. *BMC Genet.* **9**, e76. (doi:10.1186/1471-2156-9-76)
48 Penny, D., Murray-McIntosh, R. & Harrison, G. L. 2002 Estimating the number of females in the founding population of New Zealand: analysis of mtDNA variation. *J. Polyn. Soc* **111**, 207–221.
49 Beaujard, P. 2003 Les arrivées austronésiennes à Madagascar: vagues ou continuum? [Austronesian arrivals in Madagascar: Waves or continuum?]. *Étud. Océan Ind.* **35–36**, 59–147.
50 Ottino, P. 1974 Le moyen-âge de l'Océan Indien et le peuplement de Madagascar [The Indian Ocean in the Middle Ages and the settlement of Madagascar]. *Annu. Pays l'Océan Ind.* **1**, 197–221.
51 Fitzpatrick, S. M. & Callaghan, R. 2008 Seafaring simulations and the origin of prehistoric settlers to Madagascar. In *Islands of inquiry: colonization, seafaring and the archaeology of maritime landscapes* (eds G. Clark, F. Leach & S. O'Connor), pp. 55–66. Canberra, Australia: Australian National University Press.
52 Faublée, J. 1970 Les manuscrits arabico-malgaches du Sud-Est [Arab-Malagasy manuscripts of the Southeast]. *Revue Fr. d'histoire d'Outre-Mer* **57**, 268–287.
53 Adelaar, K. A. 2006 The Indonesian migrations to Madagascar: making sense of the multidisciplinary evidence. In *Austronesian diaspora and the ethnogenesis of people in Indonesian archipelago* (eds T. Simanjuntak, I. H. E. Pojoh & M. Hisyam), pp. 205–232. Jakarta, Indonesia: LIPI Press.
54 Chihara, D., Namikawa, R. & Hikata, R. 1971 *Borobudur.* Tokyo: Heibonsha.

7

AN ONGOING AUSTRONESIAN EXPANSION IN ISLAND SOUTHEAST ASIA

Lansing, J.S., **M.P. Cox**, T.A. de Vet, S.S. Downey, B. Hallmark and H. Sudoyo. 2011. An ongoing Austronesian expansion in Island Southeast Asia. *Journal of Anthropological Archaeology* 30:262-272.
http://doi.org/10.1016/j.jaa.2011.06.004

ABSTRACT The Austronesian expansion into Island Southeast Asia and the Pacific was the last and most far-reaching prehistoric human migration. Austronesian languages replaced indigenous languages over nearly half the globe, yet the absolute number of Austronesian colonists was small. Recently, geneticists have identified large geographic disparities in the relative proportions of Asian ancestry across different genetic systems (NRY, mitochondrial DNA, autosomes and X chromosomes) in Austronesian-speaking societies of Island Southeast Asia and the Pacific. Surprisingly, a substantial genetic discontinuity occurs in the middle of a continuous chain of islands that form the southern arc of the Indonesian archipelago, near the geographic center of the Austronesian world. In the absence of geographic barriers to migration, this genetic boundary and swathe of Austronesian language replacement must have emerged from social behavior. Drawing on decades of comparative ethnological research inspired by F.A.E. van Wouden's structural model of Austronesian social organization, later codified by Claude Lévi-Strauss as "House societies" ("sociétés à maison"), we propose a two-stage ethnographic model in which the appearance of matrilocal "House societies" during the initial phase of the Austronesian expansion, and the subsequent disappearance of "House societies" in lowland rice-growing regions, accounts for the observed linguistic, genetic and cultural patterns.

Journal of Anthropological Archaeology 30 (2011) 262–272

Contents lists available at ScienceDirect

Journal of Anthropological Archaeology

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

An ongoing Austronesian expansion in Island Southeast Asia

J. Stephen Lansing [b,*], Murray P. Cox [a], Therese A. de Vet [b], Sean S. Downey [b,c], Brian Hallmark [d], Herawati Sudoyo [e]

[a] Institute of Molecular Biosciences, Massey University, Private Bag 11 222, Palmerston North 4442, New Zealand
[b] School of Anthropology, University of Arizona, United States
[c] Institute of Archaeology, University College London, UK
[d] Division of Biotechnology, University of Arizona, United States
[e] Eijkman Institute for Molecular Biology, Jakarta, Indonesia

ARTICLE INFO

Article history:
Received 25 February 2011
Revision received 3 June 2011
Available online 13 July 2011

Keywords:
Austronesia
Matrilocal
Lévi-Strauss
House societies
Structuralism
Anti-chaos

ABSTRACT

The Austronesian expansion into Island Southeast Asia and the Pacific was the last and most far-reaching prehistoric human migration. Austronesian languages replaced indigenous languages over nearly half the globe, yet the absolute number of Austronesian colonists was small. Recently, geneticists have identified large geographic disparities in the relative proportions of Asian ancestry across different genetic systems (NRY, mitochondrial DNA, autosomes and X chromosomes) in Austronesian-speaking societies of Island Southeast Asia and the Pacific. Surprisingly, a substantial genetic discontinuity occurs in the middle of a continuous chain of islands that form the southern arc of the Indonesian archipelago, near the geographic center of the Austronesian world. In the absence of geographic barriers to migration, this genetic boundary and swathe of Austronesian language replacement must have emerged from social behavior. Drawing on decades of comparative ethnological research inspired by F.A.E. van Wouden's structural model of Austronesian social organization, later codified by Claude Lévi-Strauss as "House societies" ("sociétés à maison"), we propose a two-stage ethnographic model in which the appearance of matrilocal "House societies" during the initial phase of the Austronesian expansion, and the subsequent disappearance of "House societies" in lowland rice-growing regions, accounts for the observed linguistic, genetic and cultural patterns.

Introduction

Prior to AD 1500, Austronesian was the world's most widely dispersed language family, extending more than halfway around the globe from Madagascar to Easter Island. Despite decades of debate there is as yet no consensus on either the geographic origins of the Austronesian expansion, or the social mechanisms that led to the adoption of Austronesian language and cultural traits by so many societies across the Pacific. Among geneticists and linguists, much research has been directed towards clarifying the geographic pathways of migration and colonization, with less attention paid to the processes that led to the replacement of other languages and changes in the composition of populations. Until recently, this research has been hampered by the scarcity of data (genetic, linguistic and archaeological) for Island Southeast Asia (ISEA), which played a key role in the initial phase of the Austronesian expansion (Bellwood, 2001).

But recently more data has become available. This year our group completed analysis of nearly two thousand genetic samples from 13 Indonesian islands, along with associated linguistic data, and other researchers have published new investigations of genetic diversity in ISEA, Melanesia, Madagascar and Polynesia (Soares et al., 2011; Wollstein et al., 2010; Wilmshurst et al., 2011; Kayser et al., 2008; Kayser, 2010; Abdulla et al., 2009; Friedlaender et al., 2008; Kimura et al., 2008; Marck, 2008; Tabbada et al., 2010; Mona et al., 2009; Moodley et al., 2009; HUGO Pan-Asian SNP Consortium, 2009; Soares et al., 2008). Other studies have added to our knowledge of phylogenetic relationships within the Austronesian language family (Gray et al., 2009).

Here we propose a simple model for the Austronesian expansion that builds on these data as well as comparative ethnological research in ISEA. The model is based on an insight that dates from the earliest period of anthropological research in the region. In the 1930s, Dutch anthropologists began to argue that superficial variations in social structure – for example, patrilineality versus matrilineality – were probably not the result of successive migrations (vanWouden, 1968). Instead, Austronesian cultures shared a core set of ideas and institutions that found

* Corresponding author.
E-mail address: lansing@santafe.edu (J. Stephen Lansing).

0278-4165/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jaa.2011.06.004

expression in a cognitive classificatory system or "structure" linking social organization, cosmology and myth. As G.W. Lochter observed in 1968, "the great advance in understanding effected in the thirties was primarily the idea that accentuated matrilineal grouping, similarly marked patrilineal grouping, and double unilineal grouping could belong to one and the same structure (Lochter, 1968)." This structure later came to be identified with the concept of a "house society", developed by Claude Lévi-Strauss. In a series of lectures at the Collège de France from 1976 through 1981, Lévi-Strauss defined "house societies" by contrasting them with lineage-based social systems (Lévi-Strauss, 1983). His initial inspiration for these "sociétés à maison" was the noble houses of Europe: the historical house of Plantagenet, or the fictional house of Usher. Lévi-Strauss observed that houses may appear in hierarchical societies as durable social groupings, which 'reunite or transcend' opposing categories such as descent/alliance, patrilineal/matrilineal descent, hypergamy/hypogamy, and close/distant marriage (Lévi-Strauss, 1983).

While the original concept of "sociétés à maison" covers a lot of ground, subsequently ethnologists identified a much more specific set of characteristics associated with Austronesian house societies (Fox, 1993). These include cosmological dualism, with a pronounced emphasis on the complementarity of male and female principles (vanWouden, 1968); founder-focused ideology relating to both genitor and genetrix lines of descent (Bellwood, 1996); apical demotion of these lines of descent, with an emphasis on elder/younger relationships (Fox et al., 1995); and the pervasive use of botanical metaphors to express the concept of growth from an origin (Fox et al., 1995; Fox, 1980). The relationship between these concepts and their realization in social structure became a central theme in ethnographic research, published in scores of studies of the cultures of ISEA. Unsurprisingly, the highest retention of these "Austronesian" traits occurred in the most remote islands, which were less affected by later historical developments. Here we will focus on the Austronesian expansion on three such islands: Timor, Sumba and Nias, which span the breadth of Indonesia from east to west. They also encompass the full range of variation in social structure noted by the ethnologists of the 1930s, including patrilineal, matrilineal and double unilineal systems. As we will see, the house model explains the observed genetic and linguistic patterns on these islands. Interestingly, it also implies that in the Wehali region of central Timor, the Austronesian expansion continues today.

We begin with a brief overview of the genetic, archaeological and linguistic data that have been used to characterize the Austronesian expansion. Next, we define the house model, and use it to analyze this data. Finally, we consider the effects of a later historical development, the spread of irrigated rice cultivation in Western Indonesia, and the ensuing dissolution of "sociétés à maison" in this region. This two-stage historical analysis of social behavior offers an explanation for the cultural, linguistic and genetic patterns observed today across ISEA.

Austronesian genetics

Until recently, the prehistory of ISEA was usually characterized as shaped by two population dispersals: the initial Paleolithic colonization of Sahul ~45 thousand years ago, and a much later Neolithic expansion of Austronesian-speaking farmers from Taiwan ~4 thousand years ago. Recently we genotyped an extensive battery of Y chromosome markers, including 85 SNPs/indels and 12 Y-STRs, in a sample of 1917 men from 32 communities on 13 Indonesian islands. These results point to a more complex migration history. Although comprising a series of discontinuous processes, in broad terms we can summarize this history as a four-phase colonization model.

In the first phase, the arrival of the first anatomically modern humans in the late Pleistocene introduces basal NRY C and K lineages to the entire region, including Australia and Melanesia. Later dispersals of more hunter-gatherers introduced several major subclades of haplogroup O to Indonesia (e.g. O-M119, O-M95, O-P203 and O-M122) over an extended time period (e.g., from approximately 35 to 8 kya). The third stage corresponds to the Austronesian expansion, which we associate with haplogroup O-P201 and probably O-M110 and some O-P203. Elsewhere we present a detailed argument to justify this association, based on both new samples and the discovery of novel genetic markers that clarify the dispersal of the major O subclades in ISEA (Karafet et al., 2010). The fourth phase occurred in the historic era and introduced multiple haplogroups from southern Asia, Arabia and China. A key inference from this analysis is that only a small fraction of O subclades are associated with the Austronesians; other major subclades date to earlier population movements.

We also found that the paternal gene pool is sharply subdivided between Western and Eastern Indonesia, with a boundary running between the islands of Bali and Flores (Cox et al., 2010) (Fig. 1). Analysis of molecular variance reveals one of the highest levels of between-group variance yet reported for human Y chromosome data ($\phi_{ST} = 0.47$). That the Y chromosome genetic composition of these adjacent islands should be so dissimilar is surprising, because the break occurs in a narrow zone in a continuous chain of islands. Elsewhere, such a high level of population differentiation is usually associated with major geographic barriers, such as the Sahara desert or the Himalayas. The location of the division corresponds to a biogeographic frontier noted by Alfred Russell Wallace: not the famous "Wallace's Line", which traces the ancient sea barrier between Sahul and Sunda, but rather a second line located further to the east, between the islands of Sumbawa and Flores, which Wallace proposed based on his observation of phenotypic differences between human populations (Vetter, 2006).

Curiously, the sharp decline in Austronesian Y chromosomes to the east of Wallace's phenotypic line is not as clearly mirrored in the maternal gene pool. Mitochondrial DNA is passed from mothers to their children; consequently, markers on this molecule can be used to trace matrilineal descent. The eastward spread of Austronesian mitochondrial DNA was not halted by the Wallace Line, nor was the spread of Austronesian languages: both reached the far Pacific. In the 1990s, several studies showed that mitochondrial DNA in Polynesia is predominantly of Asian origin, while Y chromosomes are mostly Melanesian (Melton et al., 1995; Sykes et al., 1995). Later studies confirmed this pattern: about 94% of Polynesian mtDNA is ultimately of East Asian origin, while about 66% of Polynesian Y chromosomes are Melanesian (Cox et al., 2007; Kayser et al., 2008). The immediate predecessor of the "Polynesian motif" (mtDNA haplogroup B4a1a1a) has been found in Taiwanese aboriginals, with an estimated age of 13200 YBP (95% confidence interval: 9400–17000).

As well as Polynesia, this haplogroup is also found at the extreme western end of the Austronesian-speaking world, on the island of Madagascar, where a recent study of 266 Malagasy individuals found the Polynesian motif at levels ranging from 13% to 50% in three ethnic groups. These Polynesian motif carriers shared two polymorphisms not present elsewhere, thus defining a new Malagasy motif subclade (Razafindrazaka et al., 2010). Although molecular dating was largely uninformative, the presence of a variant of the Polynesian motif in Madagascar clearly indicates an Island Southeast Asian connection. Linguistic (Dahl, 1951; Dahl, 1977) and archaeological evidence (Burney et al., 2004; Dewar, 1996; Dewar and Wright, 1993) suggests that this linkage is relatively young (less than 1.5 kya). Like the

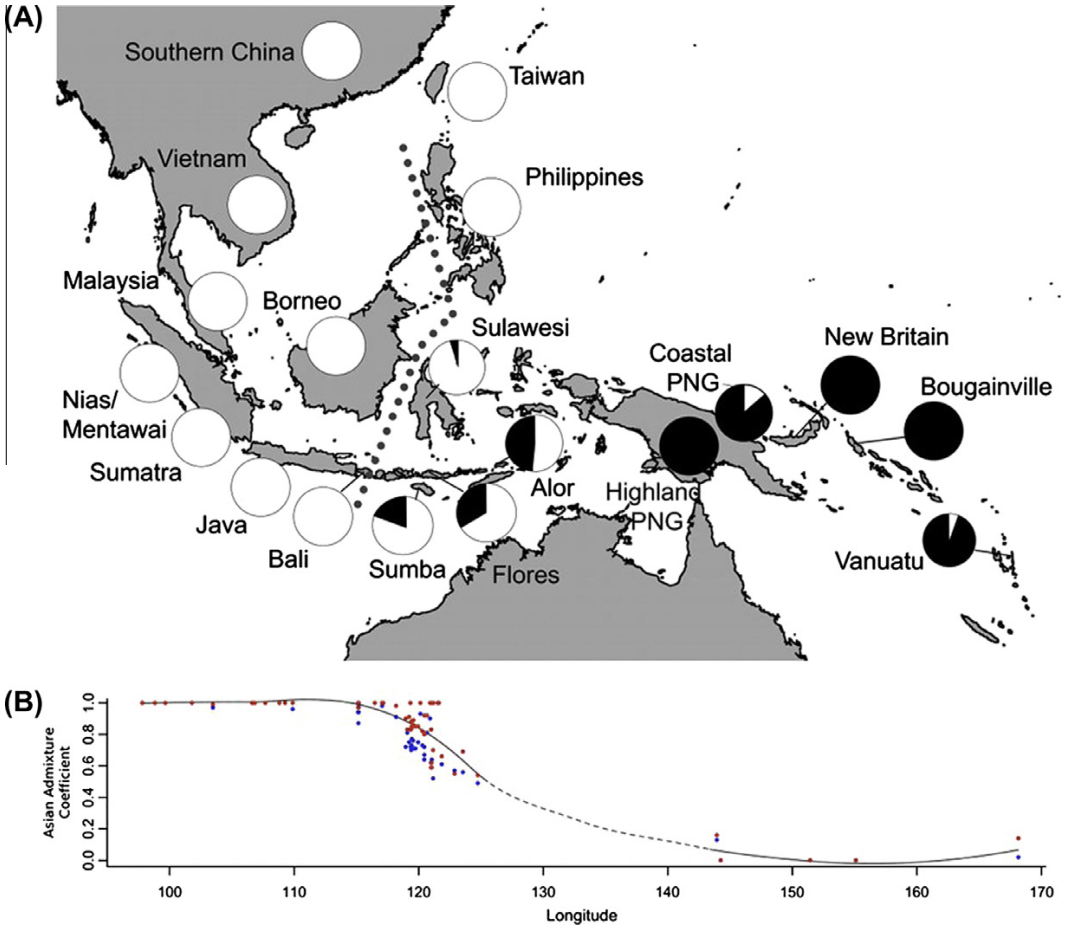

Fig. 1. Local admixture rates across the Indo-Pacific region. (A) Pie charts showing mean regional admixture rates (Asian component in white; Melanesian component in black). Wallace's biogeographical line is shown as a dotted line. Regional admixture rates are shown for data reduction purposes; admixture rates for all 60 populations (with confidence intervals) are listed in the Supplementary Information for Cox et al. (2010). (B) Change in Asian admixture rates calculated from all SNPs combined (black line). Asian admixture estimated from autosomal and X chromosomal SNPs are indicated by blue and red points, respectively. Note the decline in Asian admixture beginning in Eastern Indonesia, as well as preferential retention of X chromosomal (red) versus autosomal (blue) diversity. Regions with no data indicated by a dashed line; from other evidence, the decline in the Asian component may be more pronounced than this. Reproduced unmodified from Cox et al. (2010).

Austronesian language family, prior to AD 1500 the Polynesian motif was the world's most dispersed mitochondrial lineage.[1]

Thus a comparison of paternal and maternal genetic markers reveals two significant patterns. First, both point to an East Asian origin for the Austronesian expansion, consistent with the linguistic evidence that the Austronesian language family arose among Taiwanese aboriginal peoples. Second, there is a very pronounced sex bias, which appears to imply that few Austronesian men ventured beyond Wallace's phenotypic line, while their female relatives continued their colonizing expeditions through the islands of Melanesia, Micronesia and Polynesia. In 2003, Hage and Marck proposed that this genetic pattern could be explained as an effect of matrilocal residence and matrilineal descent in Austronesian Proto-Oceanic society (Hage and Marck, 2003). According to this model, matrilocal Austronesian communities accepted husbands from surrounding Papuan communities, and the male children of these marriages gradually displaced the Asian O clades. However, both mtDNA and NRY markers are strongly affected by genetic drift and founder effects, due to their small effective size (Cox, 2008). If the Austronesian expansion was carried out by small groups traveling in sailing canoes, as is assumed, then drift and founder effects could easily skew the demographic picture we see today. To address this issue, Cox et al developed a new sampling strategy designed to increase the statistical power of the analysis, and to clarify the sex bias revealed by the mtDNA and NRY. A small number of SNPs were identified as Ancestry Informative Markers (AIMs), which have increased power to distinguish between East Asian (southern Han Chinese) and Papuan populations (highland Papua New Guinea, which were largely isolated from Asian advances into the Pacific during the Holocene (evidence reviewed in Cox, 2008). To gain greater insight into the question of sex bias, the AIMs were chosen equally from autosomes and the X chromosome. The X chromosome spends two-thirds of its time in females and only one-third in males, whereas autosomes spend equal time in males and females, so in principle a comparison of diversity on autosomes and X chromosomes can help to reveal sex-biased migration and mixture processes (Hedrick, 2007). Thirty seven AIMs were genotyped in the largest panel of ISEA samples studied

[1] Soares et al. recently analyzed 157 complete mitochondrial genomes and suggest that the full Polynesian motif most likely originated in the vicinity of the Bismarck Archipelago. They hypothesize that the motif arose ~6 kya, but that Austronesian languages and culture arrived later, transmitted by small numbers of socially dominant Austronesian-speaking voyagers from ISEA in the Lapita formative period, ~3.5 kya (Soares et al., 2011). This conclusion is being debated, but we note that the model we propose here is consistent with either scenario.

to date: 1430 individuals from 60 populations, from mainland East Asia to Melanesia (Fig. 1). Consistent with the evidence for sex-biased admixture from the mtDNA and NRY studies, mean rates of Asian admixture are higher on the X chromosome than on the autosomes.

The genetic picture can be summed up as follows: all presumed "Austronesian" genetic markers (paternal, maternal and autosomal) are present among Taiwanese aborigines. This is consistent with the linguistic evidence that points unambiguously to Taiwan as the homeland of Austronesian languages. But it does not rule out the possibility that Austronesian genes might have traveled other routes into ISEA, or (as Soares et al propose) that there might have been several migrations from Taiwan along a "voyaging corridor". Second, there is a very striking differentiation in the distribution of NRY haplogroups between Western and Eastern Indonesia. Third, all genetic systems (NRY, autosomes, X chromosome and mtDNA) exhibit a pronounced sex bias east of Wallace's phenotypic line.

Austronesian archaeology

The relationship between the Austronesian expansion and the onset of the Neolithic in ISEA has been much debated by archaeologists. Some points are not in doubt: as Donohue and Denham note, "the earliest Malayo-Polynesian speakers were agricultural; this conclusion is apparent from historical linguistic evidence and accords with archaeological evidence from Taiwan" (2010, p. 249). Rice cultivation was well established in Taiwan by at least 5000 years ago (Zhang and Hung, 2010), and many words associated with rice and associated processing and storage facilities have been reconstructed to Proto Austronesian (PAN) (Pawley, 2007). But questions remain as to whether other Neolithic cultures might have already gained a foothold in ISEA before the arrival of Austronesian-speaking voyagers from Taiwan. There are two possibilities. First, perhaps Taiwan was not the sole or even the most important route by which Asian Neolithic innovations reached the islands. Second, several food crops were originally domesticated in Melanesia (bananas (Perrier et al., 2009), sugar cane (Grivet et al., 2004), greater yam (Malapa et al., 2005) and sago (Kjaer et al., 2004)). How far into ISEA had these crops spread, at the time of the Austronesian voyages? Had a Neolithic culture begun to flourish on some of the islands before the Austronesians arrived? A relatively sparse archaeological record has left room for much speculation. Ethnobotanical and linguistic evidence suggest a significant, pre-Austronesian westward dispersal of bananas and their cultivators from New Guinea into Eastern Indonesia and possibly even further west (Denham and Donohue, 2009). Evidence from pig mtDNA points to multiple distinct migrations both eastward out of Southeast Asia, and within Wallacea itself (Lum et al., 2006; Larson et al., 2005).

In Eastern Indonesia and Melanesia, and perhaps elsewhere, incoming Asian groups encountered other food-producing societies, leading to what has sometimes been described as a "Neolithic standoff". The rapid decline in Asian alleles in Eastern Indonesia may indicate where indigenous groups, present since the Pleistocene, were living in sufficiently large numbers to resist incursive populations spreading into the region during the mid-Holocene (Cox et al., 2010). Perhaps this demographic resistance was in part driven by indigenous agricultural traditions related to those found in New Guinea (Denham, 2005). Although only the east New Guinea highlands have yielded clear evidence for an autochthonous development of agriculture (Denham et al., 2003), indigenous agriculture may have been practiced more widely in this region than we currently have archaeological evidence for, an argument recently made from genetic evidence (Mona et al., 2007). Another possible explanation is that rice increasingly lacked power to drive the expansion of human populations as they moved into the changing climate of the equatorial belt (Cox, 2008). Alternately, perhaps the standoff reflects some transformation in the Austronesian social structure (see Jordan et al., 2009). To date, this pattern has yet to be satisfactorily explained.

Austronesian languages

Overwhelming linguistic evidence indicates that the Austronesian language family arose in Taiwan (Donohue and Denham, 2010). Nine of the ten primary subgroups of Austronesian are attested only on Taiwan. The tenth subgroup, Malayo-Polynesian, comprises all of the Austronesian languages spoken outside Taiwan, which number approximately one thousand. This subgroup is not a catchall, but has been defined on the basis of numerous shared innovations, both regular and irregular (Ross, 2009).

The phylogeny of Malayo-Polynesian languages offers several clues about the history of this language family. The phylogenetic structure is flat and rake-like, indicative of a recent rapid multidirectional expansion (Pawley, 1999). If language diversification (cladogenesis) is linked to population expansions, then expansion pulses should leave a series of short branches in the phylogenies because there will be little time for linguistic changes to accumulate before speech communities fragment. The same is true for genetic data. In contrast, when the geographic spread of cultures is constrained by physical or social boundaries, the rate of linguistic diversification should decrease, leading to longer branches (anagenesis). The Malayo-Polynesian languages conform to the first pattern. Gray et al. (1999) used lexical data and Bayesian phylogenetic methods to construct a phylogeny of 400 Malayo-Polynesian languages. This method predicted an origin of proto-Austronesian approximately 5230 years ago, and 3800–4500 years for the Malayo-Polynesan clade. The phylogeny of Malayo-Polynesian suggests very rapid geographic expansion, with four major expansion pulses and two pauses in Pacific settlement (Gray et al., 2009).

With regard to the first pause, the emergence of Malayo-Polynesian, Gray et al note that the invention of the outrigger canoe and its sail may have enabled the Austronesians to move across the 350-km Bashi channel between Taiwan and the Philippines before spreading rapidly over the 7000 km from the Philippines to Polynesia. This result is supported by linguistic reconstructions showing that the terminology associated with the outrigger canoe complex can only be traced back to Proto-Malayo-Polynesian and not Proto-Austronesian (Pawley and Pawley, 1994).

Once this expansion was under way, with few exceptions the Austronesians would have encountered islands that were already populated. An obvious question is why the Austronesian languages replaced nearly all of the pre-existing languages in ISEA. As Peter Bellwood asks, "Why are there not far more non-Austronesian [linguistic] enclaves surviving in ISEA, as there are in western Island Melanesia... ?" (Bellwood, 2010).

Austronesian house societies

In 1935, F.A.E. van Wouden surveyed the ethnographies of Eastern Indonesia and concluded that over the whole region, "in spite of the extreme unilineal character of the descent systems, both patrilineal and matrilineal descent are yet taken into account" (vanWouden, 1968). Van Wouden's analysis helped guide the work of later ethnologists, and in 1996 James J. Fox revisited this topic in a study of the transformation of progenitor and progenetrix lines of origin, noting that "although Eastern Indonesia may contribute a great deal to a model of a proto-Austronesian social world, nevertheless the region represents only one area of a vast Austronesian world." Fox's synthesis linked several analytical threads: the structuralist argument developed by van Wouden and later elaborated by Lévi-Strauss, the subsequent reconstruc-

tion of shared elements of Austronesian culture using comparative historical linguistics, and a vastly expanded ethnographic literature on Austronesian societies. Drawing on this synthesis, here we offer a structural model (sensu Lévi-Strauss) to account for the genetic, linguistic and archaeological patterns described above. For our purposes a structural model can be simply defined as the idea that "cosmos and society are organized in the same way", by means of a core set of binary symbolic oppositions (male/female, older/younger, treetrunk/tip, cosmos/society). The key difference between this approach and the comparative method of historical linguistics is a shift in analytical focus from the reconstruction of shared cognates, to the identification of a core collection of structural principles (binary symbolic oppositions) that form a "scheme of social categories... [that] serves as the model for an all-embracing classification" (Needham et al., 1968).

In Lévi-Strauss' original concept, sociétés à maison are a form of social structure intermediate between the elementary and complex structures that he had previously distinguished (Lévi-Strauss, 1949). Houses were defined by their possession of a 'domain' consisting of both material and immaterial wealth or honors; the extensive use of fictive kinship in alliance and adoption; and the transmission of the 'domain' – titles, prerogatives, and wealth – via women as well as men. Lévi-Strauss' proposal that the Austronesians introduced a particular form of "house society" to ISEA was first addressed in a collection of essays in 1987; subsequently the idea was debated in numerous articles and edited volumes. In 1993, James Fox and collaborators offered an historical perspective on Austronesian house societies, based on a comparison of contemporary ethnographic studies across ISEA, coupled with historical linguistics. As Fox noted, the reconstructed lexicon of Proto-Austronesian contains the word **Rumaq* which Blust glosses as a descent group or house (Blust, 1980). A second relevant term is Proto-Malayo-Polynesian **banua/*panua*, a more polysemous word whose glosses include inhabited territory, homeland, community and land-owing kin group (Blust, 1987). Blust also identified a large number of Proto-Austronesian terms for the physical architecture of the Malayo-Polynesian house. Along with these physical attributes, ethnographic studies refer to shared social and cosmological symbolism. Commonly, Austronesian houses define social groups and connect them to the past, using a vocabulary that emphasizes origins and founder rank. Typically, as Fox notes, the house is regarded as the ancestral embodiment of the group it represents, engaged in marital and affinal alliances with other houses.

In early twentieth century Eastern Indonesia, as van Wouden observed, the patrilineal principle dominated. But as he further noted, the underlying dualistic principle implies that female origins and descent remain significant. Houses need to form and retain alliances and to commemorate their origins. For those purposes "it is absolutely immaterial whether the principle of genealogical grouping is matrilineal or patrilineal." Instead, "one of the most striking facts is that in almost every one of the larger regions into which the area may be divided there is a people who are sharply distinguished from their patrilineal neighbors by their matrilineal descent groups." Van Wouden rejects the hypothesis of multiple migrations by noting two facts. First, "the various cultures in question exhibit too great a homogeneity to make it necessary to resort to migration-hypotheses." Second, even in the most extreme examples of unilineal descent, "both patrilineal and matrilineal descent are yet taken into account."

For example, at the extreme western edge of ISEA, communities on the island of Nias are organized as *banua* (villages) consisting of exogamous patrilineages (Schröder, 1917; Beatty, 1992). Four thousand kilometers to the east, near the eastern border of the archipelago, villages on the islands of Tanimbar consist of rows of named and unnamed houses (*uma*), linked by matrilateral alliances and affinal relations (McKinnon, 1995). In both of these societies, and in many others, a concept of cosmological dualism is expressed in the complementarity of father (**ama*) and mother (**ina*). Houses are ranked according to their distance from an origin, and the contrast between older/younger permeates the kinship system, social precedence and cosmological myths of origin (Fox, 1996). These attributes of Austronesian sociétés à maison sharply contrast with the social organization of neighboring Papuan societies, which (as Bellwood observes) "seem to lack totally any concept of genealogically-based ranking, whether of persons or descent groups... " (Bellwood, 1996).

House societies and the Austronesian expansion

Here we offer a model for the initial Austronesian expansion into ISEA, which accounts for the genetic and linguistic patterns described above. In our model, the Austronesian expansion begins with the spread of matrilocal Neolithic house societies into ISEA, which is already populated by hunter–gatherers. In Eastern Indonesia, these hunter–gatherers are Papuan. As Austronesian communities advance, their women sometimes accept husbands from neighboring Papuan communities.[2] As a consequence of matrilocal residence, the children of such marriages inherit their father's Papuan Y chromosome, their mother's Asian mitochondrial DNA and speak her Austronesian language. The simulation reflects the following assumptions:

- The initial population of both Austronesian colonists and indigenous hunter-gatherers is small.
- A Neolithic population expansion occurs in each Austronesian village.
- Small numbers of neighboring non-Austronesian males marry into the Austonesian matrilocal houses, at the rate alpha, for lengths of time that can vary at each settlement but are generally low.

The results of this simulation are shown in Fig. 2, in comparison with real data. Even if the migration rate (α) is quite low, there is ample time for a pronounced sex bias to develop. These results may be compared with genetic data from three ISEA populations we have studied, and with data from Polynesia. The observed genetic patterns will emerge after 50 generations if 2% of marriages are to non-Austronesians. This model accurately predicts the observed sex bias for all four genetic systems, and also accounts for the replacement of indigenous languages by Austronesian languages.

Equations for the model are as follows, where a = autosomes, x = X chromosome and y = non-recombining Y chromosome (NRY):

$$y(t) = (1-\alpha)^t \quad (1)$$

$$a(t) = (1-\alpha/2)^t \quad (2)$$

$$x(t) = (1-\alpha/3)^t \quad (3)$$

[2] The comparison is typically presented as Austronesians versus non-Austronesians. Following common practice, for simplicity we refer to the latter as Papuan. In Eastern Indonesia, non-Austronesians were likely the ancestors of modern Papuan peoples, who have been living continuously in this region for the past 50 kya (OConnell and Allen, 2004). In western Indonesia, the character of ancestral non-Austronesians is less clear. Borneo (40,000 BP; Kennedy, 1977; Bellwood, 1997), the Philippines (22–20,000 BP; Bellwood, 1997) and Java (4000 BP; Bellwood, 1978) likely hosted Australo-Melanesian populations from the late Pleistocene to the early Holocene, thus supporting Howell's (1976) contention of an "Old Melanesia" (Howells, 1976), a swathe of Australo-Melanesian populations that once stretched across much of modern Island Southeast Asia. However, recent genetic research indicates that peoples with Asian ancestry also have a long history in western Indonesia, stretching far back into the Pleistocene (Hill et al., 2007; Karafet et al., 2010). The nature of ancestral non-Austronesian populations in western Indonesia remains an outstanding question.

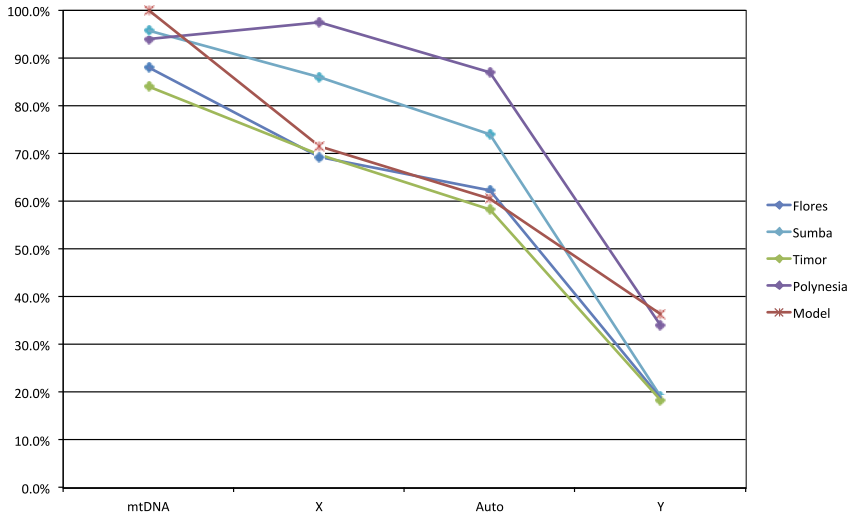

Fig. 2. Fraction of Asian DNA in four genetic systems compared with model results for $\alpha = 0.02$ and 50 generations. Sample sizes: Flores = 453, Sumba = 639, Timor = 529. Polynesian data from (Kayser et al., 2008; Cox et al., 2007; Kayser et al., 2008; Wollstein et al., 2010.).

The model describes a scenario that no longer exists in most of ISEA. By now, Austronesian languages have replaced earlier indigenous languages, and most contemporary societies are patrilineal (though they retain a dualistic emphasis on the complementarity of male and female principles). But the model scenario is closely approximated in contemporary Eastern Indonesia, in the Wehali region of central Timor. Wehali is an ancient matrilineal and matrilocal society, organized as a cluster of named houses that engage in marital alliances with each other. The women of Wehali also sometimes accept husbands from neighboring Papuan villages. The Papuan villages speak Papuan (non-Austronesian) languages, and Austronesian NRY haplotypes are less common in these communities. Thus the contemporary ethnographic situation in Wehali mirrors the conditions postulated in the model for the Austronesian expansion. To discover whether the genetic composition of Wehali houses is consistent with our model, we obtained genetic and linguistic samples from 476 men in ten villages in the Wehali region. As Fig. 2 shows, the distribution of Papuan and Austronesian haplotypes among these men agrees with the model.

Further insights into the historical demography of these villages can be obtained by calculating the effective population size of both men and women in the villages. If the effective population size of females is less than that of males, we can conclude that there has been more in-marriage of males, consistent with a long-term matrilocal sex bias. Conversely, a patrilocal and patrilineal community should exhibit a smaller effective population size for males. Table 1 presents the results of such a comparison between the matrilocal villages of Wehali, and 13 patrilocal villages on the neighboring island of Sumba. The results support the prediction: most Wehali villages show a smaller effective population size (Ne) for females than for males; the reverse is true for all 13 patrilocal Sumbanese villages (Fig. 3). This pattern persists in the con-

Table 1
Difference between effective population size calculated from haplotype data for mitochondrial DNA (mt Ne) and Y chromosome (Y Ne) for villages on Sumba and for the Wehali region of central Timor. Data and methods for estimation of effective population sizes from Lansing et al. (2008a).

Island	Village	mt N_e	Y N_e	mt N_e - Y N_e
Sumba	Anakalang	1468	190	1279
	Bilur Pangadu	3613	243	3370
	Bukambero	1998	214	1784
	Kodi	1384	175	1208
	Loli	885	371	514
	Lomboya	1601	274	1327
	Mahu	2705	279	2427
	Mamboro	1087	277	810
	Mbatakapidu	1327	183	1144
	Praibakul	1389	329	1060
	Rindi	11290	831	10459
	Waimangura	917	136	781
	Wanokaka	1566	334	1232
	Wunga	726	232	494
Timor	Besikama	1793	1192	601
	Fatuketi	885	884	2
	Kakaniuk	449	494	−46
	Kamanasa	3687	2985	702
	Kateri	936	560	376
	Kletek	1226	1803	−578
	Laran	1946	3890	−1944
	Raimanawe	1998	677	1322
	Tialai	635	621	14
	Umaklaran	516	802	−286
	Umanen Lawalu	765	2421	−1656

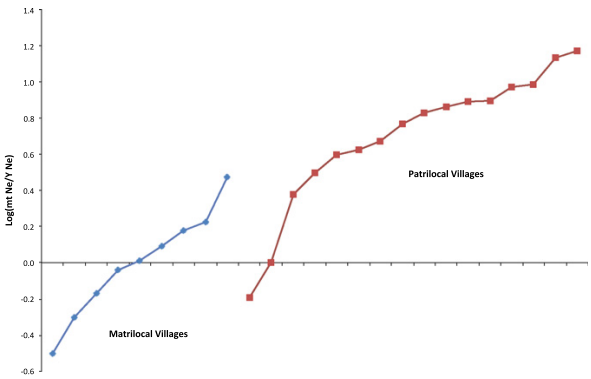

Fig. 3. Log distribution of the data shown in Table 1. This approximately continuous distribution is consistent with Lévi-Strauss' House model, which predicts that over time houses may modify their preferences for marital alliances based on genitor or genetrix. An alternative scenario of stable postmarital residence would produce distinct clumps rather than continuous variation. The largest bias for patrilocality belongs to Rindi in east Sumba, where patrilineal clans strongly favor asymmetric prescriptive alliance. The largest bias for matrilocality is found in the hamlets of Inner Wehali in Timor, site of the historic matrilocal ritual center of Wehali.

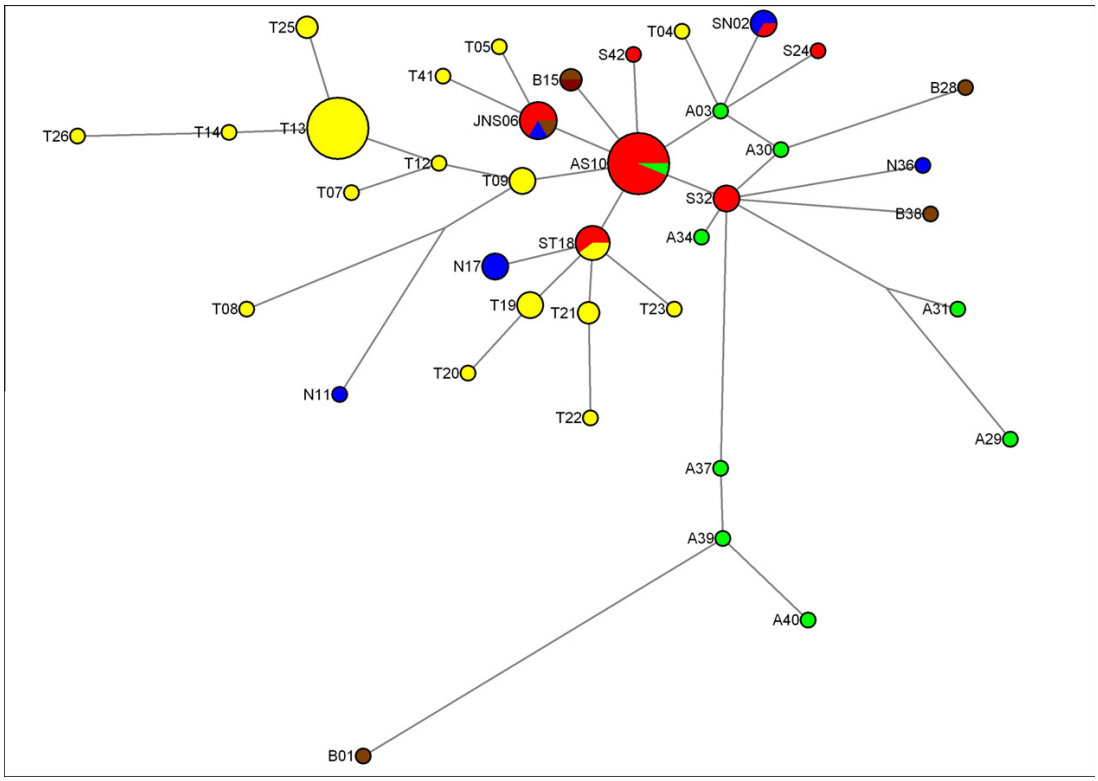

Fig. 4. Relationships within NRY haplogroup O-M110. Taiwanese aboriginals, green; Nias, blue; Java/Bali, brown; Sumba, red; and Timor, yellow. Note identical haplotypes shared by men from Taiwan (green), Nias (blue) and Sumba (red).

temporary population: survey data indicates that small numbers of men from neighboring Papuan communities have recently married into Wehali houses. The children of these marriages speak their mother's Austronesian language. Thus in Wehali, the Austronesian expansion described by the model is an ongoing process that has continued to the present day.

On Sumba, in recent times there have been no matrilineal/matrilocal communities. However, our model predicts that the original Austronesian colonists in Sumba belonged to a matrilineal/matrilocal house society, and that this form of social organization persisted for many generations. Consistent with this prediction, the concept of matrilineal descent is recognized by all Sumbanese societies, and named matrilineal descent groups still exist in some, as first noted by van Wouden and later confirmed by Rodney Needham's survey of Sumbanese kinship systems (Needham, 1987).

The Austronesian expansion into Western Indonesia

The model described above provides a concise explanation for the genetic, linguistic and ethnographic patterns observed in Eastern Indonesia. But what about the islands located to the west of Wallace's phenotypic line, where today only Austronesian languages are spoken? Here the genetic picture is more complex. As noted above, in this region Asian DNA is dominant, but nearly all NRY clades, including Asian lineages, appear to predate the Austronesian expansion. Thus the genetic evidence suggests that when Austronesian colonists arrived in western ISEA, they settled in a region already populated by indigenous hunter–gathers, mostly of Asiatic origin. The model dynamics are otherwise identical to those in Eastern Indonesia, with existing Asian hunter–gatherers taking the place of Papuans (as NANs, or non-Austronesians). Evidence for this scenario is provided by genetic, linguistic and ethnological materials from the island of Nias.

Nias is situated at the far western edge of the Indonesian archipelago. The language of Nias is an Austronesian outlier; it descends from one of the oldest branches of Malayo-Polynesian and is not closely related to other languages in the region. Nothofer has proposed that it may be a remnant from an originally larger Austronesian linguistic community, along with the language of Mentawei and some Batak languages of Sumatra (Nothofer, 1994). According to scholars of the colonial era, the island was reputed to be dangerous to mariners, and was seldom visited (Donleben, 1848; Schröder, 1917). The social organization of Nias closely resembles the patrilineal and patrilocal houses of Sumba, 3000 km to the east (Beatty, 1992; Marschall, 1977). Moreover, the architecture of Nias houses, and the custom of erecting dolmen in front of them, is strikingly similar to the houses and megaliths of Sumba (Viaro, 1981; Heine-Geldern, 1972).

We analyzed 60 samples from two villages in southern Nias and found that all of them belong to Austronesian Y chromosome haplogroups O-M110 (8 men) and O-P203 (52 men) (Karafet et al., 2010). Fig. 4 shows the distribution of haplotypes within O-M110 for Taiwanese aboriginals, as well as men from Nias, Java, Bali, Sumba and Timor. Note the clear evidence of shared haplotypes between Taiwan and Sumba (labeled in green and red) and Nias and Sumba (blue[3] and red). These genetic identities are observed across 14 microsatellite loci, and are therefore highly unlikely to reflect recurrent mutation. Instead, shared haplotypes between

[3] For interpretation of color in Figs. 1–5, the reader is referred to the web version of this article.

Table 2
Estimated time to most recent common ancestor (TMRCA) and standard deviations for the men who share haplogroup O-M110 (shown in Fig. 4), based on pedigree rate $\mu = 0.0021$.

	TMRCA	95% confidence
Taiwan	4700	976–8424
Sumba and Timor	2400	440–4360
Nias	780	388–1172
Bali	5129	2180–8060

Taiwan, Nias and Sumba indicate common ancestry. Furthermore, alrthough molecular dating has considerable uncertainty, this common ancestry occurs within the time frame predicted for the Austronesian expansion into Indonesia (Table 2).

Nias, like most of the societies of Eastern Indonesia, is now a patrilineal and patrilocal society. However, consistent with van Wouden's prediction, to the east, on the neighboring island of Sumatra there exist two large and closely related house societies, of which one (the Minangkabau) is matrilineal while the other (the Batak) is patrilineal (Singarimbun, 1975). In general, with few exceptions the contemporary cultures of western ISEA are Austronesian house societies, in which named descent groups engage in marital alliances with other houses: most but not all are patrilineal. There is, however, an important exception to this generalization.

The dissolution of "sociétés à maison" with irrigated rice culture

The only major region of Indonesia where house societies are not found today is in the rice-growing areas of Java and Bali. Interestingly, on both islands there are highland regions where irrigated rice is not grown, which retain more social and cultural attributes of traditional Austronesian house societies than the lowland rice-growing villages, as several ethnographers have noted (Hefner, 1990; Reuter, 2003). Earlier studies have proposed that the productivity of rice gardening played an important role in propelling the Austronesians into ISEA. The current eastward limit of rice is to the east of Wallace's line, which also marks the sharp subdivision in the human gene pool between Western and Eastern Indonesia (Fig. 1). It has been suggested that natural climatic variation could underpin the change from rice agriculture to tuber and palm-based economies, and the decreasing ability of rice horticulture to propel the Austronesian expansion into the territory of Papuan peoples (Cox et al., 2010).

But rice is successfully grown today in garden plots in Eastern Indonesia, and appears to be a traditional crop. We suggest that the key development occurred on the other side of Wallace's line, with the development of irrigated agriculture in Java and Bali around the sixth century C.E. (Christie, 2007). The advent of wet-rice cultivation led to the dissolution of house societies on these islands as a consequence of three demographic processes: a shift to endogamous marriage, population growth, and reduced mobility. These changes were concentrated in the regions where irrigated rice could be grown. House societies vanished from the lowlands, where endogamous farming communities grew up along the rivers and irrigation canals. In Bali, these communities adapted to population growth by creating new settlements downstream (Lansing et al., 2008b; Lansing et al., 2009). The effect of these changes was to bring large-scale (inter-island) population movement to a standstill, and to magnify existing genetic differences on either side of Wallace's phenotypic line.

A shift in marriage patterns

Evidence for a shift in marriage patterns is apparent from a comparison of demographic skew in effective population sizes of Balinese villages, as compared to skew in Eastern Indonesian house societies described above (Fig. 5). Overall the mean skew in Balinese rice-growing villages is weakly patrilocal. This may seem puzzling, since patrilocal residence is the norm in these communities. We suggest that the explanation lies in the combination of preferential endogamy and restricted mobility: both men and women tend to marry within the village where they were born. Supporting evidence for this explanation comes from two sources: surveys of contemporary marriage practices, and analysis of haplotype variation within villages.

With regard to the first point, in a survey of 252 men in 13 rice-growing villages, 84% married within their natal village. This preference fell to 34% in two highland villages, where rice is not grown.[4] With regard to the second point, genetic analysis of 587 Balinese men indicated patrilocal residence with very little movement on the landscape except for occasional micro-movements to nearby daughter settlements (see Lansing et al. (2008b) for a detailed analysis).

The argument, in sum, is that over the past 1500 years, the spread of wet-rice cultivation produced major changes to the demography of Bali, where the steep genetic cline shown in Fig. 1 begins. These changes were concentrated in the regions where irrigated rice could be grown (the highlands retained more of the ancient cultural features of 'house societies'). House societies vanished from the lowlands, where small, highly endogamous farming communities grew up along the rivers and irrigation canals. These communities adapted to population growth by creating new settlements downstream. These developments brought population movements across the Wallace line to a standstill, and accelerated genetic drift in the small, stationary rice-growing villages of Bali.

Summary and conclusions

Recent studies of the genetic and linguistic traces left behind by the colonizing voyages of the Austronesians have revealed several remarkable features. Here we explain these features by modeling the demographic and linguistic consequences of a sociocultural innovation: the appearance and disappearance of house societies among Austronesian populations.

The spread of Austronesian languages

It is clear from the genetic and archaeological evidence that ISEA was already populated when the Austronesians began their colonizing voyages. The absence of other languages in ISEA prompted Peter Bellwood's question: "Why are there not far more non-Austronesian [linguistic] enclaves surviving in ISEA, as there are in western Island Melanesia... ?" (Bellwood, 2010) Our model provides an explanation based on three assumptions: the initial population of both indigenous peoples and Austronesian colonists was small; the Austronesian settlements underwent a Neolithic population expansion; and Austronesian women occasionally married men from surrounding non-Austronesian villages. The children of these marriages would have spoken their mother's Austronesian language, while the total population of Austronesian speakers grew.

[4] The very high rate of endogamous marriage was also noticed by Leopold Howe in his study of the Balinese village of Pujung in the 1970s. He writes: "Village endogamy is very high indeed. Out of a total of 399 marriages for which I have reliable information, 82% were village endogamous. Of the 72 marriages contracted outside the village 34 were into the nearby villages of Ked and Bonjaka, and both of these were thought to be offshoots of Pujung (Bonjaka indeed is, since it was settled within living memory by people from Pujung). There is therefore good reason to include these as endogamous marriages, in which case the total goes up to 90.5%". Leopold E.A. Howe, Pujung: An Investigation into the Foundations of Balinese Culture. Doctoral thesis, University of Edinburgh, 1980.

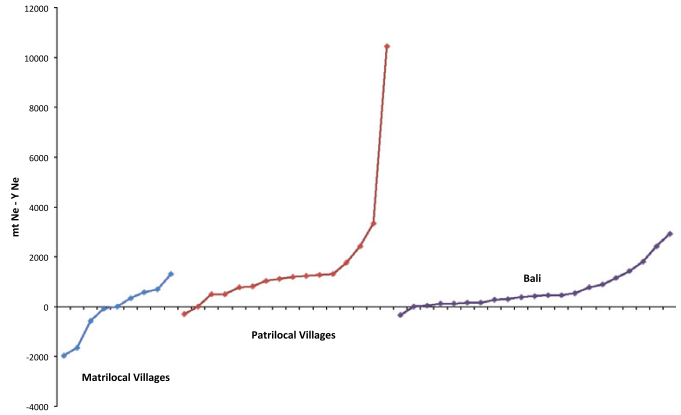

Fig. 5. Difference between effective population size calculated from haplotype data for mitochondrial DNA (mt Ne) and Y chromosome (Y Ne) for villages on Bali and Eastern Indonesia.

Given sufficient time, this process will eventually lead to the replacement of non-Austronesian languages. The smaller the indigenous population at the time of Austronesian colonization, the faster this will occur. Consistent with this prediction, in an earlier study we found that on the island of Sumba, the rates of retention of Papuan words and genes vary systematically, with higher retention in areas with more dense Papuan populations at the time of the Austronesian colonization (Lansing, 2007). This analysis can be extended to explain the presence of Papuan-speaking enclaves on the island of Timor, to the east of Sumba. Timor is the largest island in Eastern Indonesia, so the ratio of Austronesian colonists to the Papuan-speaking population would have been quite small during the initial Austronesian expansion. Thus in central Timor, the Austronesian expansion continues today, as Papuan-speaking men occasionally marry into the matrilocal Austronesian villages of Wehali. Over time, this continuing process will tilt the balance in favor of Austronesian languages in the region, albeit at a much slower rate than in the initial phase of the Austronesian expansion.

Sex-biased genetic skew

Pronounced sex-biased variation in all chromosomes was initially observed for Oceania, and subsequently found in ISEA. In 2003, Hage and Marck proposed that it could be the result of matrilocal residence, and in 2009 Jordan et al used linguistic reconstruction to argue that matrilocal residence is ancestral in Austronesian societies (Jordan et al., 2009). In 2010, Gray et al argued that "the high levels of male-biased admixture detected in Polynesian genetic studies must either have occurred over this very short time span (approximately four generations), with Papuan males actively incorporated into the Austronesian expansion, or there was extended post-settlement contact between Near Oceania and Polynesia."

Our model provides an alternative explanation, which does not require additional assumptions about either unusually high rates of admixture, or post-settlement contact between Near Oceania and Polynesia. Instead, the model predicts that the same processes of admixture occurred along the whole colonization route. Because the demographic effects are cumulative, they appear most strongly in Polynesians who live at the end of the chain of colonization.

The sharp genetic cline along Wallace's phenotypic line

It is clear that the sharp genetic cline along Wallace's phenotypic line is the result of social processes rather than geography, because it occurs along a continuous chain of islands that have been populated for tens of thousands of years. Our model explains this break by the disappearance of house societies in lowland regions in the western islands, where irrigated rice cultivation triggered a population explosion and also brought population movement to a standstill.

Conclusion: butterfly effects

Claude Lévi-Strauss did not undertake ethnographic research in ISEA, but his reflections on house societies had a profound influence on comparative anthropological studies in the region. Subsequent ethnographic studies showed that widely dispersed Austronesian societies share not only closely related languages, but a gendered cosmology and an emphasis on origins that is not found in neighboring Melanesian societies. Today house societies are found in many parts of ISEA, and phylogenetic analysis suggests that they are not a recent innovation, but have existed since the onset of the Austronesian expansion.

The model we have proposed here traces the genetic, linguistic, demographic and cultural consequences of the prevalence of this form of social organization. In Lévi-Strauss' original conception, house societies sometimes emerge as a transitional phase in the evolution of complex societies, when competition for social rank weakens social ties based on shared unilineal descent. The house provides a way to "solidify" the "unstable relation of alliance" (Lévi-Strauss, 1987), by sanctioning ties based on either genitor or genetrix (or as the Austronesians say, **ama* and **ina*). At any given moment, a house may emphasize one relationship (for example, social ties based on shared relationship to a genitor), but there is always the possibility to strengthen alliances based on the complementary principle. Dualistic cosmologies recognize and acknowledge this potential. Thus for Lévi-Strauss, house societies are an emergent and dynamical form of social organization, which actively reshape their social environment, while varying in their immediate emphasis on patrilineal or matrilineal social relations.

But both Lévi-Strauss and van Wouden assumed that there was no inherent trajectory to the evolution of "sociétés à maison." The genetic data reviewed here suggests otherwise for the Austronesian case. As the model shows, over a time scale of tens of generations a seemingly trivial shift in marriage preferences can produce a seismic change in language, culture and demography. The data we have reviewed here provide strong evidence that at a first approximation, this transformative potential was realized in the wake of the Austronesian colonization of ISEA.

The idea that such a subtle process could have such profound consequences seems counter-intuitive, and leads us to offer a final remark. Edward Lorenz' "butterfly effect" quickly became the canonical example of chaotic behavior in dynamical systems due to sensitive dependence on initial conditions (Hilborn, 2003). The mathematical basis of the butterfly effect is the Lorenz equations, three first order differential equations in which the iteration of initially tiny variation in initial parameterization quickly produces chaotic dynamics. Whereas in our model, the iteration of small α over many generations creates new regimes of order.

Acknowledgments

This research was supported by the US National Science Foundation and the Eijkman Institute for Molecular Biology, Jakarta Indonesia. M.P.C. was supported by a Rutherford Fellowship from the Royal Society of New Zealand (RDF-10-MAU-001). We gratefully acknowledge the assistance of Tatiana Karafet and Meryanne Tumonggor with the genetic analysis.

References

Abdulla, M.A. et al., 2009. Mapping human genetic diversity in Asia. Science 326, 1541–1545.
Beatty, A., 1992. Society and Exchange in Nias. Clarendon Press, Oxford.
Bellwood, P., 1978. Man's Conquest of the Pacific: The Prehistory of Southeast Asia and Oceania. Collins, Auckland (p. 76).
Bellwood, P., 1996. "Hierarchy, founder ideology and Austronesian expansion". In: Fox, James J., Clifford, Sather (Eds.), Origins, Ancestry and Alliance: Explorations in Austronesian Ethnography. Research School of Pacific and Asian Studies, Australian National University, Canberra, p. 22.
Bellwood, P., 1996. Hierarchy, founder ideology and Austronesian expansion. In: Fox, James J., Clifford, Sather (Eds.), Origins, Ancestry and Alliance: Explorations in Austronesian Ethnography. Research School of Pacific and Asian Studies, Australian National University, Canberra, p. 22.
Bellwood, P., 1997. Prehistory of the Indo-Malaysian Archipelago. University of Hawai Press, Honolulu, pp. 84–87.
Bellwood, P., 2001. Foregone conclusions? Current anthropology comment. Current Anthropology 42, 107–108.
Bellwood, P., 2010. Response to "Farming and language in Island Southeast Asia: reframing Austronesian history". Current Anthropology 51, 240–241.
Blust, R., 1980. Early Austronesian social organization: the evidence of language. Current Anthropology 21, 205–247.
Blust, R., 1987. Lexical reconstruction and semantic reconstruction: the case of Austronesian house words. Diachronica IV (1/2), 79–106 (28).
Burney, D.A., Burney, L.P., Godfrey, L.R., Jungers, W.L., Goodman, S.M., Wright, H.T., Jull, A.J.T., 2004. A chronology for late prehistoric Madagascar. Journal of Human Evolution 47, 25–63.
Christie, J.W., 2007. Water and rice in early Java and Bali. In: Boomgaard, P. (Ed.), A World of Water: Rain, Rivers and Seas in Southeast Asian Histories. KITLV Press, Leiden, pp. 23–58.
Cox, M.P., 2008. The genetic environment of Melanesia: clines, clusters and contact. In: Koven, V.T. (Ed.), Population Genetics Research Progress. Nova Science Publishers, New York, pp. 45–83.
Cox, M.P., Redd, A.J., Karafet, T.M., Ponder, C.A., Lansing, J.S., Sudoyo, H., Hammer, M.F., 2007. A Polynesian motif on the Y chromosome: population structure in remote Oceania. Human Biology 79, 525–535.
Cox, M.P., Karafet, T.M., Lansing, J.S., Sudoyo, H., Hammer, M.F., 2010. Autosomal and X-linked SNPs reveal a sharp transition from Asian to Melanesian ancestry in Eastern Indonesia and a female-bias in admixture rates. Proceedings of the Royal Society of London Series B 277, 1589–1596.
Dahl, O.C., 1951. Malgache et Maanyan: Une comparison linguistique. Egede Instituttet, Oslo.
Dahl, O.C., 1977. La subdivision de la famille Barito et la place du Malgache. Acta Orientalia 38, 77–134.
Denham, T., 2005. Envisaging early agriculture in the highlands of New Guinea: landscapes, plants and practices. World Archaeology 37, 290–306.
Denham, T., Donohue, M., 2009. Pre-Austronesian dispersal of banana cultivars West from New Guinea: linguistic relics from Eastern Indonesia'. Archaeology in Oceania 44, 18–28.
Denham, T.P., Haberle, S.G., Lentfer, C., Fullager, R., Field, J., Therin, M., Porch, N., Winsborough, B., 2003. Origins of agriculture at Kuk Swamp in the highlands of New Guinea. Science 301, 189–193.
Dewar, R.E., 1996. The archaeology of the early colonization of Madagascar. In: Reade, J. (Ed.), The Indian Ocean in Antiquity. Kegan Paul and the British Museum, London, pp. 471–486.
Dewar, R.E., Wright, H.T., 1993. The culture history of Madagascar. Journal of World Prehistory 7, 417–466.
Donleben, J.F., 1848. Bijdragen tot de kennis van het eiland Nias. Tijdschrift voor Nederlandsch-Indie 10, 171–199.
Donohue, M., Denham, T., 2010. Farming and language in Island Southeast Asia: reframing Austronesian history. Current Anthropology 51 (2), 223–256.
Fox, J.J. (Ed.), 1980. The Flow of Life: Essays on Eastern Indonesia. Harvard University Press, Cambridge, MA.
Fox, J.J., 1993. Comparative perspectives on Austronesian houses. in Fox, James J. (Ed.), Inside Austronesian Houses: Perspectives on Domestic Designs for Living. Research School of Pacific and Asian Studies, Australian National University, Canberra, p. 1.
Fox, James J., Clifford, Sather (Eds.), 1996. Origins, Ancestry and Alliance: Explorations in Austronesian Ethnography. Research School of Pacific and Asian Studies, Australian National University, Canberra, p. 223.
Fox, J.J., Bellwood, P., Tryon (Eds.), 1995. The Austronesians: Historical and Comparative Perspectives. Anthropology, Research School of Pacific and Asian Studies, Canberra, p. 223.
Friedlaender, J.S., Friedlaender, F.R., Reed, F.A., Kidd, K.K., Kidd, J.R., Chambers, G.K., Lea, R.A., Loo, J.-H., Koki, G., Hodgson, J.A., Merriwether, D.A., Weber, J.L., 2008. The genetic structure of Pacific Islanders. PLoS Genetics 4, e19.
Gray, R.D. et al., 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science 323 (479), 479–483.
Grivet, L. et al., 2004. A review of recent molecular genetics evidence for sugarcane evolution and domestication. Ethnobotany Research & Applications 2, 9–17.
Hage, P., Marck, J., 2003. Matrilineality and the Melanesian origin of Polynesian Y chromosomes. Current Anthropology 44 (1), S121–S127 (F S123).
Hedrick, P.W., 2007. Sex: Differences in mutation, recombination, selection, gene flow, and genetic drift. Evolution 61, 2750–2771.
Hefner, R., 1990. Hindu Javanese. Princeton University Press, Princeton.
Heine-Geldern, R., 1972. The archaeology and art of Sumatra. In: Loeb, Edwin M. (Ed.), Sumatra: Its History and People. Oxford University Press, Kuala Lumpur (1935).
Hilborn, R.C., 2003. Sea gulls, butterflies, and grasshoppers: a brief history of the butterfly effect in nonlinear dynamics. American Journal of Physics 72 (4), 425–427.
Hill, C., Soares, P., Mormina, M., Macaulay, V., Clarke, D., Blumbach, P.B., Vizuete-Forster, M., Forster, P., Bulbeck, D., Oppenheimer, S., Richards, M., 2007. A mitochondrial stratigraphy for Island Southeast Asia. American Journal of Human Genetics 80, 29–43.
Howells, W.W., 1976. Physical variation and history in Melanesia and Australia. American Journal of Physical Anthropology 45, 641–649.
HUGO Pan-Asian SNP Consortium, 2009. Mapping human genetic diversity in Asia. Science 326, 1541–1545.
Jordan, F.M., Gray, R.D., Greenhill, S.J., Mace, R., 2009. Matrilocal residence is ancestral in Austronesian societies. Proceedings of the Royal Society B 276 (1664), 1957–1964.
Karafet, T.M., Hallmark, B., Cox, M.P., Sudoyo, H., Downey, S., Lansing, J.S., Hammer, M.F., 2010. Major east–west division underlies Y chromosome stratification across Eastern Indonesia. Molecular Biology and Evolution 27, 1833–1844.
Kayser, M. et al., 2008. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. Molecular Biology and Evolution 23 (11), 2234–2244.
Kayser, M., 2010. The human genetic history of Oceania: near and remote views of dispersal. Current Biology 20 (4), R194–R201.
Kayser, M., Choi, Y., van Oven, M., Mona, S., Brauer, S., Trent, R.J., Suarkia, D., Schiefenhovel, W., Stoneking, M., 2008. The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. Molecular Biology and Evolution 25, 1362–1374.
Kayser, M., Lao, O., Saar, K., Brauer, S., Wang, X., Nürnberg, P., Trent, R.J., Stoneking, M., 2008. Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. American Journal of Human Genetics 82, 194–198.
Kennedy, K.A.R., 1977. The deep skull of Niah. Asian Perspectives 20, 32–50.
Kimura, R., Ohashi, J., Matsumura, Y., Nakazawa, M., Inaoka, T., Ohtsuka, R., Osawa, M., Tokunaga, K., 2008. Gene flow and natural selection in Oceanic human populations inferred from genome-wide SNP typing. Molecular Biology and Evolution 25 (8), 1750–1761.
Kjaer, A. et al., 2004. Investigation of genetic and morphological variation in the sago palm (*Metroxylon sagu*; Arecaceae) in Papua New Guinea. Annals of Botany (London) 94, 109–177.
Lansing, J.S. et al., 2007. Coevolution of languages and genes on the Island of Sumba, Eastern Indonesia. Proceedings of the National Academy of Sciences of the United States of America 104 (41), 16022–16026.
Lansing, J.S., Watkins, J.C., Hallmark, B., Cox, M.P., Karafet, T.M., Sudoyo, H., Hammer, M.F., 2008a. Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations. Proceedings of the National Academy of Sciences of the United States of America 105, 11645–11650 (See SOM for calculations of effective population size).
Lansing, J.S., Karafet, T.M., Schoenfelder, J., Hammer, M.F., 2008b. A DNA signature for the expansion of irrigation in Bali? In: Sanchez-Mazas, A., Blench, R., Ross, M., Peiros, I., Lin, M. (Eds.), Past Human Migrations in East Asia and Taiwan: Matching Archaeology, Linguistics and Genetics. Routledge, London, pp. 76–394.
Lansing, J.S., Downey, S.S., Jannsen, M., Schoenfelder, J., 2009. A robust budding model of Balinese water temple networks. World Archaeology 41 (1), 112–133.

Larson, G., Dobney, K., Albarella, U. et al. (13 co-authors). 2005. Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. Science 307:1618-1621.
Lévi-Strauss, C., 1949. Les structures Elementaires de la parenté. Presses Universitaires de France, Paris.
Lévi-Strauss, C., 1983. Le regard éloigné. Plon, Paris.
Lévi-Strauss, C., 1987. The View from Afar. Basic Books. Orig. Le regard éloigné (1983), Paris, Plon, pp. 155–159.
Lochter, G.W., 1968. Introduction to the English translation. In: van Wouden, F.A.E. (Ed.), Types of Social Structure in Eastern Indonesia. Martinus Nijhoff, The Hague (original 1935).
Lum, J.K., McIntyre, J.K., Greger, D.L., Huffman, K.W., Vilar, M.G., 2006. Recent Southeast Asian domestication and Lapita dispersal of sacred male pseudohermaphroditic "tuskers" and hairless pigs of Vanuatu. Proceedings of the National Academy of Sciences of the United States of America 103 (46), 17190–17195.
Malapa, R. et al., 2005. Genetic diversity of the greater yam (Dioscorea alata L.) and relatedness to D. nummularia Lam. and D. transversa Br. as revealed with AFLP markers. Genetic Resources and Crop Evolution 52, 919–929.
Marck, J., 2008. Proto Oceanic society was matrilineal. Journal of the Polynesian Society 117, 345–382.
Marschall, W., 1977. Der Berg des Herren der Erde. Deutscher Taschenbuch Verlag, Munich.
McKinnon, S., 1995. Houses and hierarchy: the view from a South Moluccan Society. In: Carsten, J., Hugh Jones, S. (Eds.), About the House: Lévi-Strauss and Beyond. Cambridge University Press, Cambridge, p. 175.
Melton, T.R., Redd, A.J., et al., 1995. Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. American Journal of Human Genetics 57, 403–414.
Mona, S., Tommaseo-Ponzetta, M., Brauer, S., Sudoyo, H., Marzuki, S., Kayser, M., 2007. Patterns of Y-chromosome diversity intersect with the Trans-New Guinea hypothesis. Molecular Biology and Evolution 24, 2546–2555.
Mona, S., Grunz, K.E., Brauer, S., Pakendorf, B., Castr, L., Sudoyo, H., Marzuki, S., Barnes, R.H., Schmidtke, J., Stoneking, M., Kayser, M., 2009. Genetic admixture history of Eastern Indonesia as revealed by Y-chromosome and mitochondrial DNA analysis. Molecular Biology and Evolution 26, 1865–1877.
Moodley, Y., Linz, B., Yamaoka, Y., Windsor, H.M., Breurec, S., Wu, J.-Y., Maady, A., Bernhoft, S., Thiberge, J.-M., Phuanukoonnon, S., Jobb, G., Siba, P., Graham, D.Y., Marshall, B.J., Achtman, M., 2009. The peopling of the Pacific from a bacterial perspective. Science 323, 527–530.
Needham, R., 1987. Mamboru: History and Structure in a Domain of Northwestern Sumba. Oxford University Press.
Needham R., 1968. Foreword to F.A.E. van Wouden, Types of Social Structure in Eastern Indonesia. Martinus Nijhoff, The Hague, p. 2 (original 1935).
Nothofer, Bernd, 1994. The relationship between the languages of the Barrier Islands and the Sulawesi-Philippine languages. In: Tom, Dutton, Darrell, T. Tryon (Eds.), Language Contact and Change in the Austronesian World. Mouton, de Gruyter, Berlin, New York, pp. 389–409.
O Connell, J.F., Allen, J., 2004. Dating the colonization of Sahul (Pleistocene Australia New Guinea): a review of recent research. Journal of Archaeological Science 31, 835–853.
Pawley, A., 1999. In: Bellwood, P., Renfrew, C. (Eds.), Examining the Farming/Language Dispersal Hypothesis. McDonald Institute for Archaeological Research, Cambridge, pp. 251–274 (2002).
Pawley, A., 2007. The origins of early Lapita culture: the testimony of historical linguistics. In: Bedford, S., Sand, C., Connaughton, S.P. (Eds.), Oceanic Explorations: Lapita and Western Pacific Settlement. Australian National University, Canberra, pp. 17–49.
Pawley, A., Pawley, T., 1994. In: Pawley, A., Ross, M. (Eds.), Austronesian Terminologies: Continuity and Change. Australian National University, Canberra, pp. 329–361.
Perrier, X. et al., 2009. Combining biological approaches to shed light on the evolution of edible bananas. Ethnobotany Research & Applications 7, 199–216.
Razafindrazaka, H. et al., 2010. Complete mitochondrial DNA sequences provide new insights into the Polynesian motif and the peopling of Madagascar. European Journal of Human Genetics 18, 575–581.
Reuter, T., 2003. The House of Our Ancestors: Precedence and Dualism in Highland Balinese Society. University of Washington Press.
Ross, M., 2009. Proto Austronesian verbal morphology: a reappraisal. In: Adelaar, Alexander, Pawley, Andrew (Eds.), Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust. Pacific Linguistics, Canberra, pp. 95–326.
Schröder, E.E.W. Gs., 1917. Nias: ethnographische, geographische en historische aantekeningen en studien, vol. 2. E.J. Brill, Leiden.
Singarimbun, M., 1975. Kinship, Descent and Alliance among the Karo Batak. University of California Press, Berkeley.
Soares, P. et al., 2011. Ancient Voyaging and Polynesian Origins. American Journal of Human Genetics 88, 239–247.
Soares, P., Trejaut, J.A., Loo, J.-H., Hill, C., Mormina, M., Lee, C.-L., Chen, Y.-M., Hudjashov, G., Forster, P., Macaulay, V., Bulbeck, D., Oppenheimer, S., Lin, M., Richards, M.B., 2008. Climate change and postglacial human dispersals in Southeast Asia. Molecular Biology and Evolution 25, 1209–1218.
Sykes, B.A. et al., 1995. The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. American Journal of Human Genetics 57, 1463–1475.
Tabbada, K.A., Trejaut, J., Loo, J.-H., Chen, Y.-M., Lin, M., Mirazn-Lahr, M., Kivisild, T., De Ungria, M.C.A., 2010. Philippine mitochondrial DNA diversity: a populated viaduct between Taiwan and Indonesia? Molecular Biology and Evolution 27, 21–31.
van Wouden, F.A.E., 1968. Types of Social Structure in Eastern Indonesia. Martinus Nijhoff, The Hague (original 1935) p. 153.
Vetter, J., 2006. Wallace's other line: human biogeography and field practice in the eastern colonial tropics. Journal of the History of Biology 39, 89–123.
Viaro, A., 1981. Urbanisme et architecture traditionnels du sud de l'isle de Nias. Etablissements humains et environnement socio-culturel #21. UNESCO, Paris.
Wilmshurst, J.M., Hunt, T.L., Lipo, C.P., Anderson, A.J., 2011. High-precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. PNAS 108 (5), 1815–1820.
Wollstein, A. et al., 2010. Demographic history of oceania inferred from genome-wide data. Current Biology 20, 1983–1992.
Zhang, C., Hung, Hsiao-chun, 2010. The emergence of agriculture in southern China. Antiquity 84 (323), 11–25.

8

AUTOSOMAL AND X-LINKED SINGLE NUCLEOTIDE POLYMORPHISMS REVEAL A STEEP ASIAN-MELANESIAN ANCESTRY CLINE IN EASTERN INDONESIA AND A SEX BIAS IN ADMIXTURE RATES

Cox, M.P., T.M. Karafet, J.S. Lansing, H. Sudoyo and M.F. Hammer. 2010. Autosomal and X-linked single nucleotide polymorphisms reveal a steep Asian-Melanesian ancestry cline in eastern Indonesia and a sex bias in admixture rates. *Proceedings of the Royal Society B* 277:1589-1596.
http://doi.org/10.1098/rspb.2009.2041

ABSTRACT The geographical region between mainland Asia and New Guinea is characterized by numerous small islands with isolated human populations. Phenotypically, groups in the west are similar to their neighbours in mainland Southeast Asia, eastern groups near New Guinea are similar to Melanesians, and intervening populations are intermediate in appearance. A long-standing question is whether this pattern primarily reflects mixing between groups with distinct origins or whether natural selection has shaped this range of variation by acting differentially on populations across the region. To address this question, we genotyped a set of 37 single nucleotide polymorphisms that are evolutionarily independent, putatively neutral and highly informative for Asian-Melanesian ancestry in 1430 individuals from 60 populations spanning mainland Asia to Melanesia. Admixture analysis reveals a sharp transition from Asian to Melanesian genetic variants over a narrow geographical region in eastern Indonesia. Interestingly, this admixture cline roughly corresponds to the human phenotypic boundary noted by Alfred Russell Wallace in 1869. We conclude that this phenotypic gradient probably reflects mixing of two long-separated ancestral source populations – one descended from the initial Melanesian-like inhabitants of the region, and the other related to Asian groups that immigrated during the Paleolithic and/or with the spread of agriculture. A higher frequency of Asian X-linked markers relative to autosomal markers throughout the transition zone suggests that the admixture process was sex-biased, either favouring a westward expansion of patrilocal Melanesian groups or an eastward expansion of matrilocal Asian immigrants. The matrilocal marriage practices that dominated early Austronesian societies may be one factor contributing to this

observed sex bias in admixture rates.

SUPPLEMENTARY MATERIALS

http://rspb.royalsocietypublishing.org/content/277/1687/1589.figures-only

PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2010) **277**, 1589–1596
doi:10.1098/rspb.2009.2041
Published online 27 January 2010

Autosomal and X-linked single nucleotide polymorphisms reveal a steep Asian–Melanesian ancestry cline in eastern Indonesia and a sex bias in admixture rates

Murray P. Cox[1,3], Tatiana M. Karafet[1], J. Stephen Lansing[2,4], Herawati Sudoyo[5] and Michael F. Hammer[1,2,*]

[1]*ARL Division of Biotechnology, Life Sciences South 231, and* [2]*School of Anthropology, University of Arizona, Tucson, AZ 85721, USA*
[3]*Allan Wilson Centre for Molecular Ecology and Evolution, and the Bio-Protection Centre, Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand*
[4]*Santa Fe Institute, Santa Fe, NM 87521, USA*
[5]*Eijkman Institute for Molecular Biology, Diponegoro 69, Jakarta, Indonesia*

The geographical region between mainland Asia and New Guinea is characterized by numerous small islands with isolated human populations. Phenotypically, groups in the west are similar to their neighbours in mainland Southeast Asia, eastern groups near New Guinea are similar to Melanesians, and intervening populations are intermediate in appearance. A long-standing question is whether this pattern primarily reflects mixing between groups with distinct origins or whether natural selection has shaped this range of variation by acting differentially on populations across the region. To address this question, we genotyped a set of 37 single nucleotide polymorphisms that are evolutionarily independent, putatively neutral and highly informative for Asian–Melanesian ancestry in 1430 individuals from 60 populations spanning mainland Asia to Melanesia. Admixture analysis reveals a sharp transition from Asian to Melanesian genetic variants over a narrow geographical region in eastern Indonesia. Interestingly, this admixture cline roughly corresponds to the human phenotypic boundary noted by Alfred Russell Wallace in 1869. We conclude that this phenotypic gradient probably reflects mixing of two long-separated ancestral source populations—one descended from the initial Melanesian-like inhabitants of the region, and the other related to Asian groups that immigrated during the Paleolithic and/or with the spread of agriculture. A higher frequency of Asian X-linked markers relative to autosomal markers throughout the transition zone suggests that the admixture process was sex-biased, either favouring a westward expansion of patrilocal Melanesian groups or an eastward expansion of matrilocal Asian immigrants. The matrilocal marriage practices that dominated early Austronesian societies may be one factor contributing to this observed sex bias in admixture rates.

Keywords: admixture; sex-biased; ancestry; Indonesia; Austronesian

1. INTRODUCTION

Two broad spheres of cultural influence dominate Indo-Pacific pre-history. Modern humans first settled in the area around 45 thousand years ago (O'Connell & Allen 2004; Barker 2005). Melanesians, who probably resembled these earliest settlers (Brothwell 1960; Howells 1973; Krigbaum & Datan 1999; Bellwood 2007) and are today largely restricted to New Guinea and its surrounding islands, are thought to be their direct descendents. However, Melanesian groups may once have ranged more widely across Island Southeast Asia (ISEA) (Howells 1973). Conversely, populations to the west and north (e.g. in western Indonesia, Borneo, Sulawesi, the Philippines and mainland Southeast Asia) (figure 1) are characterized by Asian features. The arrival time of these populations in ISEA is not well known, nor is the extent to which they contributed genetically to the then-isolated Melanesian populations. Certainly, the spread of some Asian groups was relatively recent, coinciding with the first agricultural settlements in ISEA (Bellwood 2005). This process is often termed the Austronesian expansion, and is putatively linked to demic dispersals from mainland China during the Mid-Holocene.

Mitochondrial DNA (mtDNA) and Y chromosome (NRY) data suggest that the Austronesian expansion differentially affected populations in ISEA. Broadly speaking, populations in the west have considerable Asian ancestry (Karafet *et al.* 2005); Asian lineages occur less frequently further east (Friedlaender *et al.* 2007); and in the remote highlands of New Guinea, where Austronesian languages and cultural items are absent (Bellwood 2007), Asian lineages have not been found at all. This pattern has been attributed to groups expanding recently out from mainland Asia/Taiwan and mixing with pre-existing populations in Melanesia.

* Author for correspondence (mfh@u.arizona.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2009.2041 or via http://rspb.royalsocietypublishing.org.

Received 9 November 2009
Accepted 6 January 2010

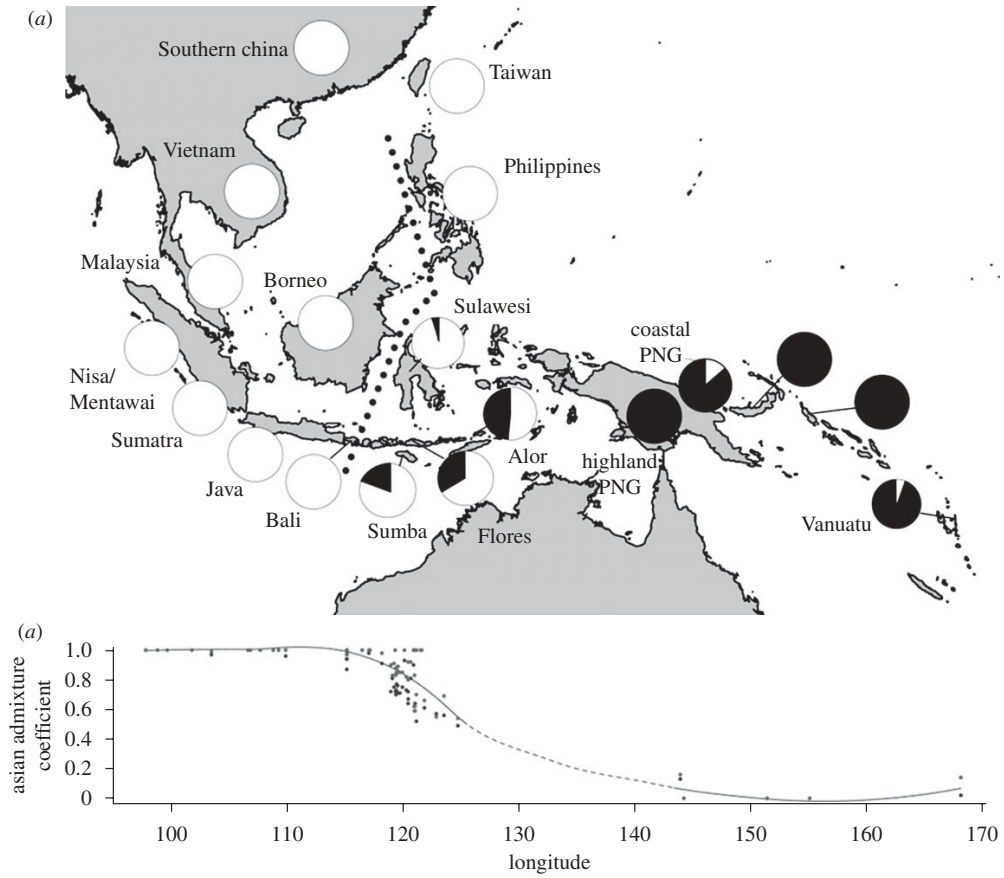

Figure 1. Local admixture rates across the Indo-Pacific region. (*a*) Pie charts showing mean regional admixture rates (Asian component in white; Melanesian component in black). Wallace's biogeographic line is shown as a dotted line. Regional admixture rates are shown for data reduction purposes; admixture rates for all 60 populations (with confidence intervals) are listed in the electronic supplementary material. (*b*) Change in Asian admixture rates calculated from all SNPs combined (black line). Regions with no data indicated by a dashed line (exact gradient unknown). Asian admixture estimated from autosomal and X chromosomal SNPs are indicated by black and grey points, respectively. Note the decline in Asian admixture beginning in eastern Indonesia, as well as preferential retention of X chromosomal (grey) versus autosomal (black) diversity.

Furthermore, haploid data are discordant with respect to the relative frequencies of lineages with putatively Asian and Melanesian origins, which has been attributed to sex-biased admixture. For instance, mtDNA lineages with Asian affinity are often found at higher frequency than Asian Y chromosomes in eastern Indonesia and Oceania (Mona *et al.* 2009), thereby suggesting that the admixture process favoured Asian women (Hage & Marck 2003). Because the entire haploid mtDNA and NRY are strongly affected by genetic drift and founder events owing to their small effective size (Cox 2006), and possibly by natural selection that may have acted on functional sites anywhere within these non-recombining systems, some caution is needed when interpreting this result. Studies of multiple, putatively neutral regions of the nuclear genome have substantially more power to address questions of this nature (Ellegren 2009); however, only three such published studies have relevance for the Indo-Pacific region (Friedlaender *et al.* 2008; Kayser *et al.* 2008; Kimura *et al.* 2008). While these three studies differ in the kind of marker genotyped (autosomal STRs or single nucleotide polymorphisms (SNPs)) and the particular population(s) sampled (mostly Pacific Islanders from different island groups), they tend to agree that approximately 80 per cent of the Polynesian autosomal gene pool is of East Asian origin and approximately 20 per cent is of Melanesian origin. The largest of the three studies also inferred that Taiwanese aboriginals have a predominantly (approx. 100%) Asian origin, while island Melanesian groups share fewer than 20 per cent of Asian markers (Friedlaender *et al.* 2008). Because each of these studies targeted individual populations or small geographical regions, we still do not have a good understanding of the spatial distribution of Asian–Melanesian ancestry across the Indo-Pacific region. Important questions remain. Does the extent of Asian–Melanesian ancestry really differ across the Indo-Pacific region? Do patterns of ancestry in ISEA mirror those in flanking regions (i.e. mainland Asia and Oceania)? Has incursive Asian gene flow produced a recognizable geographical pattern? And do these admixture rates have a sex-specific bias?

To address these questions, we adopted a different strategy to previous studies. Rather than sampling a large number of randomly identified genomic markers, most of which contain little information about admixture,

we assembled a relatively small panel of SNPs ($n = 37$) that exhibit high F_{ST} between representative populations in Asia (southern Han Chinese) and Melanesia (Papua New Guinea (PNG) highlanders). These two groups were chosen because (i) their nuclear genetic diversity has been at least partially characterized (Jakobsson *et al.* 2008; Li *et al.* 2008), (ii) they represent two extremes of population differentiation among Southeast Asian populations (Bellwood 2007), and (iii) there is little evidence for pre-historic contact between them—PNG highland populations were largely insulated from Asian advances into the Pacific during the Holocene (evidence reviewed in Cox 2008). We then genotyped these markers in the largest panel of ISEA samples studied to date: 1430 individuals from 60 populations spanning mainland Asia to Melanesia. This approach—using targeted ancestry informative markers (AIMs) chosen from putative source populations outside our study range—allows us to use a smaller panel of markers without significant loss of power. Furthermore, we chose SNPs equally from the autosomes and the X chromosome to address the question of sex-specific admixture. This is possible because the X chromosome spends two-thirds of its time in females and only one-third of its time in males, whereas the autosomes spend equal time in males and females. Therefore, sex-biased migration and mixture processes can, in principle, be detected by examining patterns of diversity on the autosomes and the X chromosome (Hedrick 2007). We also chose our SNPs so that they were effectively unlinked (i.e. statistically independent) and are located in intergenic regions of the genome to minimize the confounding effects of natural selection.

2. MATERIAL AND METHODS

(a) *Ancestry informative markers*

We identified SNPs that show high F_{ST} between southern Han Chinese and PNG highlanders. F_{ST} was calculated as described previously (Cox *et al.* 2008). SNPs were chosen from two sources: the HOMINID dataset, a collection of resequenced putatively neutral regions distributed across the human genome (Wall *et al.* 2008), and the Jakobssen dataset, a collection of 500 000 SNPs typed in the HGDP-CEPH panel (Jakobsson *et al.* 2008). To address the question of sex-specific admixture, SNPs ($n = 37$) were selected from both the autosomes and the X chromosome (electronic supplementary material, table S1). All SNPs showed high F_{ST} between the two populations (autosome mean = 0.76, X chromosome mean = 0.72; a difference of less than 6%). Moreover, to minimize the effects of natural selection, all SNPs were located away from genes (including introns, UTRs and immediate flanking regions) and are more than 1 cM distant from other SNPs in the panel (i.e. they are evolutionarily independent).

(b) *Samples*

We assayed 1430 DNA samples from 60 populations across the Indo-Pacific region (mainland China to Vanuatu). Sample sizes varied, but averaged to 33 genotypes per SNP per population (electronic supplementary material, table S2). Sample details are available elsewhere (Karafet *et al.* 2005; Lansing *et al.* 2007, 2008), with the exception of New Britain, which is a composite sample of individuals from two inland, Papuan-speaking regions in eastern New Britain—Mali (Marabu) and Kaket (Rangulit and Malasait); Long Gi-Berau, which consists mostly of ethnic Dayak Kenyah; Long Soloy-Berau, which consists mostly of ethnic Dayak Punan; and Timur Hadakewa, which has been included in our greater Flores group, although these individuals were sampled on the neighbouring (but culturally related) island of Lembata.

(c) *Data generation*

We generated more than 76 000 genotypes for the sample panel. SNPs were genotyped using two different methods: TaqMan, a probe-based PCR assay, and Sequenom, a multiplexed single-base primer extension platform. All genotyping was performed by the University of Arizona Genetics Core facility (http://uagc.arl.arizona.edu/). Appropriate controls were run for both methods, and a subset of SNPs was validated against DNA sequences generated previously (Wall *et al.* 2008).

(d) *Admixture estimates*

Because we ascertained SNPs using specific criteria rather than sampling an unbiased range of SNP diversity (i.e. we did not draw from the full site frequency spectrum), many methods of inferring admixture could not be applied here. This includes recently developed coalescent approaches (Bertorelle & Excoffier 1998; Chikhi *et al.* 2001; Wang 2003). Instead, we inferred admixture rates using a modified weighted least-squares estimator (Chakraborty *et al.* 1992). This method has the advantage of computational speed, and is therefore readily amenable to resampling approaches and power analyses. We scripted the algorithm in R (code available on request) and validated it against the original test datasets (Chakraborty *et al.* 1992). The method was modified to account for the sampling error in each of the 'parental' (P_1 and P_2) and 'hybrid' (H) populations by inferring a frequency density for P_1, P_2 and H at each SNP (i.e. a weighted histogram that returns the observed allele frequency). Drift, a stochastic process, was addressed by examining a large number of unlinked (i.e. statistically independent) SNPs. Using a computationally intensive resampling approach, random variables were drawn from the three frequency distributions, admixture was calculated using the least-squares estimator and the process was repeated 10^5 times. The median admixture rate with 95 per cent confidence intervals (i.e. 0.025 and 0.975 quantiles) was calculated from the distribution of resampled admixture rates. Code was parallelized and run on a UNIX-based high-performance computing grid at the University of Arizona.

3. RESULTS

To obtain Asian admixture rates, we typed 37 SNPs: 18 from the autosomes and 19 from the X chromosome (electronic supplementary material, table S1). All SNPs have $F_{ST} > 0.50$ between southern Han Chinese and PNG highlanders, which places them in the top 2 per cent of all genomic polymorphisms (Jakobsson *et al.* 2008). Autosomal SNPs had an average F_{ST} of 0.76 (range: 0.51–0.95); X chromosomal SNPs had an average F_{ST} of 0.72 (range: 0.50–0.85). All SNPs are from intergenic regions (i.e. map far from genes and known functional regions) and are effectively unlinked (i.e. each SNP is more than 1 cM away from any other marker in the panel). These 37 markers were typed in 1430 individuals from 60 Indo-Pacific populations

Table 1. Regional admixture rates. See electronic supplementary material, tables S3–S6, for detailed information on individual populations.

region	location	population	Asian admixture			difference	female preference
			all	A	X		
mainland Asia	China	Chinese	1.00	1.00	1.00	0	0
	Vietnam	Vietnamese	1.00	1.00	1.00	0	0
	Malaysia	Malay	1.00	0.97	0.99	0.03	+
ISEA	Taiwan	Aboriginal	1.00	1.00	1.00	0	0
	Philippines	Aeta	1.00	1.00	0.83	−0.17	−
	Philippines	Filipino	1.00	1.00	1.00	0	0
Indonesia	Sumatra	Toba	1.00	1.00	1.00	0	0
	Nias	Nias	1.00	1.00	1.00	0	0
	Mentawai	Mentawai	1.00	1.00	1.00	0	0
	Java	Dieng	0.98	0.96	1.00	0.04	+
	Java	Javanese	1.00	1.00	1.00	0	0
	Borneo	Borneo	1.00	0.99	1.00	0.01	+
	Bali	Bali	0.99	0.95	1.00	0.04	+
	Sulawesi	Sulawesi	0.97	0.91	0.99	0.08	+
	Flores/Lembata	Flores/Lembata	0.66	0.62	0.69	0.08	+
	Sumba	Sumba	0.81	0.74	0.86	0.12	+
	Alor	Alor	0.51	0.49	0.54	0.04	+
Melanesia	PNG	Coastal	0.14	0.13	0.16	0.03	+
	Bismarck Archipelago	New Britain	0	0	0	0	0
	Bougainville	Nasioi	0	0	0	0	0
	Vanuatu	Maewo	0.06	0.02	0.14	0.12	+

(i.e. approx. 33 genotypes per SNP per population) (electronic supplementary material, table S2). Most differences in these SNP frequencies help to distinguish between Asian and Melanesian populations; for instance, over half of the observed variance of our dataset is carried on the first principal component (electronic supplementary material, figure S1). These AIMs are thus well chosen to capture the major variation along the Asian–Melanesian gradient, although we emphasize that other important components of variation are likely to exist in ISEA.

We assembled a dataset of 37 allele frequencies for 60 populations (i.e. a matrix containing 2220 entries). To reduce this dataset to more manageable proportions, we used the allele frequencies to estimate admixture rates for each population. Admixture rates were calculated using an approach modified from Chakraborty *et al.* (1992) (see §2 for details). We defined two representative 'parental' populations: one for Asia (southern Han Chinese) and one for Melanesia (PNG highlanders). SNP frequencies are similar in all seven Chinese populations (southern Han, Taiwanese Hakka, Miao, She, Tujia, Yao and Yi), and indeed, most of our mainland Asian populations. Admixture estimates (see subsequently) reinforce this shared history. Similarly, no Asian admixture was inferred for the ancestors of all the Papuan-speaking populations in this study (indigenous groups from PNG highlands, New Britain and Bougainville). Consequently, the choice of populations selected to represent the extremes of Asian–Melanesian diversity has little effect on admixture estimates.

We estimated admixture rates for the autosomal and X chromosomal SNPs combined (electronic supplementary material, table S3), as well as for the autosomes (electronic supplementary material, table S4) and X chromosome separately (electronic supplementary material, table S5). For ready comparison, regional admixture rates (i.e. summarized for major island groups) are listed in table 1. In brief, populations from Borneo, Bali and westward exhibit extremely high rates of Asian ancestry (effectively fixed at 100%). While Asian ancestry still predominates across Wallace's biogeographic line in Sulawesi (approx. 97%), the extent of the Asian contribution drops off rapidly in the islands further east: approximately 81 per cent on Sumba, approximately 66 per cent on Flores and approximately 51 per cent on Alor (figure 1*a*). Further east in Melanesia, only the Austronesian-speaking regions of coastal PNG (approx. 14%) and Vanuatu (approx. 6%) show any evidence of Asian admixture. Asian admixture was not observed in our Papuan-speaking populations from New Britain and Bougainville (i.e. Nasioi). The change from predominantly Asian to predominantly Melanesian genomic ancestry occurs rapidly within a relatively small area of eastern Indonesia that falls roughly between 120 and 145 east longitude (figure 1*b*) (Cox 2008). We have no samples from 125 to 145 east longitude, but based on published haploid loci from these regions (Mona *et al.* 2007), the Melanesian genomic component is probably dominant beyond 130 east longitude (i.e. from New Guinea and further east).

Admixture rates also differ between the autosomes and the X chromosome (figure 1*b*, black and grey dots, respectively). We observe higher mean rates of Asian admixture on the X chromosome, which is consistent with approximately 7 per cent greater contribution on average from Asian women during the admixture process (electronic supplementary material, table S6). This difference in admixture rates reaches statistical significance for only a few populations because the confidence intervals on our admixture rates are generally large. (The Aetas of the Philippines are an important exception.) To

check whether admixture rates are, in a broad sense, higher on the X chromosome, we instead compared our data with expectations under the binomial distribution. Across 43 Indonesian populations (where we have greater control of sampling quality and coverage), we observe 31 cases where admixture is higher on the X chromosome than on the autosomes. This outcome is statistically highly unlikely ($p = 0.0027$) if this admixture ratio were fluctuating by chance alone. When we exclude cases where Asian ancestry reaches 100 per cent on both the autosomes and the X chromosome (i.e. there is zero difference in admixture rates), the observed pattern is even less likely ($p < 0.00001$). We conclude that Asian women made a larger contribution than Asian men to the ancestors of modern ISEA populations during the admixture process.

Finally, we explored whether this bias in admixture rates is structured spatially across Indonesia. Visual inspection of the data suggests a higher level of bias in the southeast (electronic supplementary material, figure S2). To check whether this pattern differs statistically from a uniform expectation, we selected the k nearest neighbours of each sampling location, and determined whether neighbouring points exhibited lower (-1), identical (0) or greater ($+1$) admixture on the X chromosome relative to the autosomes. We also set the number of cases where Asian ancestry reaches 100 per cent on both the autosomes and the X chromosome to the observed level ($n = 7$). The empirical dataset was then compared with simulations where admixture rates were permitted to fluctuate randomly (i.e. higher or lower in each sampling location). We found that the observed distribution was an outlier under all of these simulations, regardless of the choice of k (all $p < 0.002$). Therefore, the bias towards Asian women is significantly structured across Indonesia; neighbouring populations have similar admixture biases, and this effect is increasingly prevalent towards the southeast.

4. DISCUSSION

An Asian–Melanesian phenotypic gradient across the Indo-Pacific region has long been recognized (Wallace 1869); however, the proportion of nuclear loci with Asian and Melanesian ancestry in populations across this region has not been well studied. Here, we perform the first survey of multiple, putatively neutral DNA polymorphisms in a large set of populations to infer admixture dynamics across ISEA. We first genotyped a series of AIMs that distinguish Asian and Melanesian diversity, and then determined the geographical pattern and extent of Asian–Melanesian ancestry. Rather than appearing as a broad cline across Indonesia, we find that a dramatic change from Asian to Melanesian ancestry occurs within a relatively narrow geographical window in the far east—essentially within the Indonesian province of Nusa Tenggara Timur. In particular, to the west of Sumba and Flores, Asian ancestry approaches 100 per cent; while east of Alor, evidence of Asian ancestry diminishes dramatically (see below). Alor is among the most westerly locations where Papuan languages are spoken, and Melanesian Y chromosome lineages are common on all of these islands (Lansing *et al.* 2007, 2008), but not further west (Karafet *et al.* 2005). Interestingly, this transition is shifted eastward relative to Wallace's line—a boundary that separates the biogeographic regions of Asia and Wallacea. At its southern limit, Wallace's line falls between the islands of Bali and Lombok (figure 1), which are separated by a deep-water sea channel that marks the southern edge of the Sunda Shelf. During ice-age glacial advances, the Sunda land mass included Borneo, Bali, Java and Sumatra, together with mainland Southeast Asia. However, even in periods of low sea level, deep water in Wallacea separated the Sunda shelf from the eastern landmass of Sahul (connecting New Guinea and Australia). While the distribution of many flora and fauna conforms to Wallace's line, the seafaring capabilities of human settlers to this region undoubtedly overcame this barrier to dispersal. Indeed, Asian ancestry exceeds 50 per cent as far east as the island of Alor, which is well within Wallacea and approximately 1000 km east of Bali, as well as on the island of Sulawesi, which is located east of Wallace's line in the north (figure 1). Curiously, Wallace himself noted this difference, positing a second line in eastern Indonesia corresponding to changes in human phenotype (Wallace 1869; Cox 2008). Wallace's second 'phenotypic' line broadly parallels the rapid decline in Asian admixture identified here.

The historical processes underlying this sharp transition from Asian to Melanesian ancestry are not completely clear. Human genetic diversity is typically partitioned over geography in more gradual clines observed at the level of continents (Serre & Paabo 2004). Steep and narrow clines are more unusual (Novembre & Di Rienzo 2009), partly because long-term stability requires large initial gene frequency differences between source populations, and repeated gene flow tends to destablize them (Wijsman & Cavalli-Sforza 1984). A major question emerging from this study relates to the age of the cline in eastern Indonesia. Was it established in the Paleolithic by the encounter of genetically differentiated hunter–gatherer groups (Hill *et al.* 2007), or did it arise more recently with the mixing of Austronesian farmers and local populations in eastern Indonesia? Climatic changes following the last glacial maximum (approx. 18 kya; Mulvaney & Kamminga 1999) may have spurred expansions of Asian hunter–gatherers into ISEA from further north on the mainland (Soares *et al.* 2008). Indeed, the spread of the Southeast Asian Hoabinhian culture into Sumatra is one tangible marker of these movements (Bellwood 2007). Dispersals of Asian hunter–gatherers radiating over an extended period of time during the Paleolithic (e.g. 35–8 kya) may have introduced a proportion of the Asian alleles that we detect in western ISEA.

Alternatively, as pointed out in many previous studies based on both genetic (Cox 2005, 2006, 2008; Karafet *et al.* 2005; Hill *et al.* 2007; Lansing *et al.* 2007; Mona *et al.* 2009) and archaeological data (Bellwood 2005, 2007), there is good evidence that significant Asian contact occurred in eastern Indonesia and Melanesia during the Austronesian expansion. Indeed, much of the pattern of admixture we observe in this study, especially in Wallacea and Near Oceania, may well reflect cultural processes associated with the expansion of Austronesian farmers into the territory previously occupied by Melanesian hunter–gatherers (given subsequently). A

third alternative involves a combination of Paleolithic and Neolithic migrational processes. Despite the utility of this set of highly informative autosomal and X-linked SNPs for obtaining admixture rates, more extensive genetic data distributed across the genome would be required to infer the timing of different waves of migration (Hellenthal *et al.* 2008). Furthermore, the observed variance in admixture rates among individual communities may well be caused by a variety of demographic factors, such as genetic drift and repeated founder events, during the admixture process. The exact nature and mode of action of these factors at the community level remain unclear.

A major finding of this study is that mean rates of Asian admixture are higher on the X chromosome than on the autosomes, suggesting that Asian women made a approximately 7 per cent greater contribution on average during the admixture process(es). This is consistent with previous studies noting a higher proportion of Asian mtDNA versus Y chromosome lineages in many Indo-Pacific populations from eastern Indonesia (Mona *et al.* 2009). Our results, derived from unlinked and highly informative nuclear markers, are concordant with this finding. In our study, this sex bias in admixture rate appears more prevalent towards the southeast. What cultural processes might underlie this pattern? Further east in Oceania, it has been attributed to the role of matrilocal communities during the Austronesian expansion (Hage 1999; Hage & Marck 2003; Kayser *et al.* 2008). Matrilocal residence has been inferred as the ancestral state in early Austronesian societies (Jordan *et al.* 2009), including many whose descendent communities are analysed here. In matrilocal groups, husbands live with their wife's kin, and therefore local Melanesian men would be preferentially incorporated into expanding Asian groups. Over time, this process tends to increase Asian maternal versus paternal ancestry (i.e. increase the frequency of Asian X-linked and mtDNA markers, and decrease the frequency of Asian autosomal and Y chromosomal markers), and probably underpins the bias towards higher rates of Asian admixture on the X chromosome relative to the autosomes. Conversely, an unbiased admixture process would not cause an imbalance in autosome and X chromosome admixture rates.

Other evidence also links the observed pattern in the eastern portion of this range to the Austronesian expansion. The only Melanesian groups in our study that speak Papuan languages—that is, New Britain and Bougainville—show no evidence of Asian admixture, in accordance with earlier studies (Friedlaender *et al.* 2007). Asian admixture is also infrequent in the Austronesian-speaking populations of Melanesia; that is, coastal PNG and Vanuatu. Indeed, Maewo islanders are one of only two Vanuatu populations where Austronesian Y chromosome lineages have been detected (at a frequency of approx. 10%; Cox 2006). Even the low rates of Asian admixture that we infer for these Austronesian-speaking communities (approx. 10%) may therefore be elevated relative to surrounding groups. The Aeta, a hunter–gatherer Negrito group from the Philippines, is a key exception to general trends. Here, Asian admixture is significantly lower on the X chromosome relative to the autosomes ($p > 0.05$; electronic supplementary material, tables S4 and S5), suggesting that Asian admixture in the Aeta was biased towards males. This deviation in autosome and X chromosome admixture rates (approx. 17%) is the most extreme of any in our study (electronic supplementary material, table S6). However, Negrito populations probably experienced the flip side of the Austronesian expansion; older communities such as the Aeta adopted Austronesian languages only recently, and local hunter–gatherer women may have preferentially married men from neighbouring agricultural communities. For completeness, we note that an alternative hypothesis for the admixture pattern we observe is a westward expansion of Melanesians practising patrilocality into a territory previously occupied by populations of predominantly Asian ancestry. Some evidence for such a westward expansion comes from archaeobotanical and botanical data supporting a pre-Austronesian dispersal of banana cultivars from New Guinea into eastern Indonesia (Denham & Donohue 2009).

A remaining question is why Asian–Melanesian ancestry changes over such a small area in eastern Indonesia. It may mark the region where indigenous Papuan groups were large enough to resist incursive Austronesian populations during the Mid-Holocene. Such demographic resistance may have its foundation in pre-existing agricultural traditions related to those found in New Guinea (Denham 2005). While the only clear evidence of pre-Austronesian agriculture in the region occurs in the highlands of eastern New Guinea (Denham *et al.* 2003), indigenous agriculture may have been practiced more widely than current archaeological evidence suggests. Patterns of Y chromosome diversity have been presented to support this model (Mona *et al.* 2007). Alternately, the rapid phenotypic cline may instead result from changes in Austronesian agricultural practices (Cox 2008). The domestication of rice has long been credited as a presumptive trigger of the Austronesian expansion (Bellwood 1978), but rice agriculture becomes less important from west to east across Indonesia. The current eastward limit of rice falls in eastern Indonesia, where the seasonal tropical climate of more northerly latitudes is transformed into the 'season-less' monotony of the equatorial zone (Dewar 2003). This natural climatic variation probably underpins the change from rice agriculture to tuber-based economies (Spriggs 2000). Therefore, the rapid admixture gradient that we observe may reflect the decreasing ability of rice agriculture to propel the Austronesian expansion into the long occupied territory of indigenous Papuan groups (Cox 2008).

Here, we present the first regional picture of Asian admixture rates across the Indo-Pacific. Rates of Asian ancestry vary both geographically and in a sex-specific manner: we infer Asian ancestry to be approximately 100 per cent for most ISEA populations in the west and north, whereas most Melanesian groups show little Asian ancestry—an absence that is especially notable in Papuan-speaking groups. Instead of a gradual cline, we show that the relative contributions of these neutral markers change rapidly over a small area of eastern Indonesia (i.e. near the lesser Sunda islands of Flores, Sumba, Lembata and Alor). Because our markers are distributed across the human genome, the observed cline better fits a demographic scenario, and the primary cause of the phenotypic boundary first identified by Alfred Wallace is most probably the mixing of two phenotypically distinct

populations, not natural selection. Additionally, we show that admixture is biased towards the X chromosome relative to the autosomes, which indicates that Asian women were probably favoured in this admixture process. The same general pattern dominates ISEA, with increasing prevalence towards the southeast. Here, in eastern ISEA at least, it may reflect the matrilocal residence system of ancestral Austronesian societies.

Indonesian samples were obtained in collaboration with the Eijkman Institute for Molecular Biology, Jakarta, Indonesia, with the assistance of Indonesian Public Health clinic staff, following protocols for the protection of human subjects established by both the Eijkman Institute and the University of Arizona Institutional Review Boards. Permission to conduct research in Indonesia was granted by the Indonesian Institute of Sciences, and informed consent was obtained from all project participants.

We thank Kevin Keyes, Daniel Teberg and Meryanne Tumonggor for assistance in analysing genetic data in the Hammer Laboratory at the University of Arizona. We also thank Jonathan Friedlaender (Temple University) for providing samples from New Britain, as well as for helpful discussion regarding their provenance; Brian Hallmark (University of Arizona) for GIS support; and Lisa Kent (University of Arizona) for data entry. Indonesian samples were obtained by J.S.L. and H.S., and by Golfiani Malik, Wuryantari Setiadi, Loa Helena Suryadi and Meryanne Tumonggor of the Eijkman Institute for Molecular Biology, Jakarta, Indonesia, with the assistance of Indonesian Public Health clinic staff. This research was supported by grants from the National Science Foundation, the James McDonnell Foundation Robustness Programme at the Santa Fe Institute and the Eijkman Institute, Jakarta, Indonesia.

Author contributions: M.P.C. designed research; J.S.L. and H.S. collected samples; M.P.C. and T.M.K. performed research; M.P.C. analysed data; and M.P.C. and M.F.H. wrote the paper.

The authors declare no conflicts of interest.

REFERENCES

Barker, G. 2005 The archaeology of foraging and farming at Niah Cave, Sarawak. *Asian Perspect.* **44**, 90–106. (doi:10.1353/asi.2005.0004)

Bellwood, P. 1978 *Man's conquest of the Pacific: the prehistory of Southeast Asia and Oceania*. Auckland, New Zealand: Collins.

Bellwood, P. 2005 *The first farmers: the origins of agricultural societies*. Oxford, UK: Blackwell Publishing.

Bellwood, P. 2007 *Prehistory of the Indo-Malaysian Archipelago*. Canberra, Australia: ANU E Press, Australian National University.

Bertorelle, G. & Excoffier, L. 1998 Inferring admixture proportions from molecular data. *Mol. Biol. Evol.* **15**, 1298–1311.

Brothwell, D. R. 1960 Upper Pleistocene human skull from Niah caves, Sarawak. *Sarawak Mus. J.* **15–16**, 323–349.

Chakraborty, R., Kamboh, M. I., Nwankwo, M. & Ferrell, R. E. 1992 Caucasian genes in American Blacks: new data. *Am. J. Hum. Genet.* **50**, 145–155.

Chikhi, L., Bruford, M. W. & Beaumont, M. A. 2001 Estimation of admixture proportions: a likelihood-based approach using Markov Chain Monte Carlo. *Genetics* **158**, 1347–1362.

Cox, M. P. 2005 Indonesian mitochondrial DNA and its opposition to a Pleistocene era origin of proto-Polynesians in Island Southeast Asia. *Hum. Biol.* **77**, 179–188. (doi:10.1353/hub.2005.0037)

Cox, M. P. 2006 Extreme patterns of variance in small populations: placing limits on human Y-chromosome diversity through time in the Vanuatu Archipelago. *Ann. Hum. Genet.* **71**, 390–406. (doi:10.1111/j.1469-1809.2006.00327.x)

Cox, M. P. 2008 The genetic environment of Melanesia: clines, clusters and contact. In *Population genetics research progress* (ed. V. T. Koven), pp. 45–83. New York, NY: Nova Science Publishers.

Cox, M. P., Woerner, A. E., Wall, J. D. & Hammer, M. F. 2008 Intergenic DNA sequences from the human X chromosome reveal high rates of global gene flow. *BMC Genet.* **9**, e76.

Denham, T. 2005 Envisaging early agriculture in the highlands of New Guinea: landscapes, plants and practices. *World Archaeol.* **37**, 290–306. (doi:10.1080/00438240500095447)

Denham, T. & Donohue, M. 2009 Pre-Austronesian dispersal of banana cultivars West from New Guinea: linguistic relics from eastern Indonesia. *Archaeol. Oceania* **44**, 18–28.

Denham, T. P., Haberle, S. G., Lentfer, C., Fullager, R., Field, J., Therin, M., Porch, N. & Winsborough, B. 2003 Origins of agriculture at Kuk Swamp in the Highlands of New Guinea. *Science* **301**, 189–193. (doi:10.1126/science.1085255)

Dewar, R. E. 2003 Rainfall variability and subsistence systems in Southeast Asia and the western Pacific. *Curr. Anthropol.* **44**, 369–388. (doi:10.1086/368348)

Ellegren, H. 2009 The different levels of genetic diversity in sex chromosomes and autosomes. *Trends Genet.* **25**, 278–284. (doi:10.1016/j.tig.2009.04.005)

Friedlaender, J. S., Friedlaender, F. R., Hodgson, J. A., Stoltz, M., Koki, G., Horvat, G., Zhadanov, S., Schurr, T. G. & Merriwether, D. A. 2007 Melanesian mtDNA complexity. *PLoS ONE* **2**, e248. (doi:10.1371/journal.pone.0000248)

Friedlaender, J. S. *et al.* 2008 The genetic structure of Pacific Islanders. *PLoS Genet.* **4**, e19. (doi:10.1371/journal.pgen.0040019)

Hage, P. 1999 Reconstructing ancestral Oceanic society. *Asian Perspect.* **38**, 200–277.

Hage, P. & Marck, J. 2003 Matrilineality and the Melanesian origin of Polynesian Y chromosomes. *Curr. Anthropol.* **44**, S121–S127. (doi:10.1086/379272)

Hedrick, P. W. 2007 Sex: differences in mutation, recombination, selection, gene flow, and genetic drift. *Evolution* **61**, 2750–2771. (doi:10.1111/j.1558-5646.2007.00250.x)

Hellenthal, G., Auton, A. & Falush, D. 2008 Inferring human colonization history using a copying model. *PLoS Genet.* **4**, e1000078. (doi:10.1371/journal.pgen.1000078)

Hill, C. *et al.* 2007 A mitochondrial stratigraphy for Island Southeast Asia. *Am. J. Hum. Genet.* **80**, 29–43. (doi:10.1086/510412)

Howells, W. W. 1973 *The Pacific islanders*. New York, NY: Charles Scribner's Sons.

Jakobsson, M. *et al.* 2008 Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998–1003. (doi:10.1038/nature06742)

Jordan, F. M., Gray, R. D., Greenhill, S. J. & Mace, R. 2009 Matrilocal residence is ancestral in Austronesian societies. *Proc. R. Soc. B* **276**, 1957–1964. (doi:10.1098/rspb.2009.0088)

Karafet, T. M. *et al.* 2005 Balinese Y-chromosome perspective on the peopling of Indonesia: genetic contributions from pre-Neolithic hunter–gatherers, Austronesian farmers, and Indian traders. *Hum. Biol.* **77**, 93–114. (doi:10.1353/hub.2005.0030)

Kayser, M., Lao, O., Saar, K., Brauer, S., Wang, X., Nürnberg, P., Trent, R. J. & Stoneking, M. 2008 Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *Am. J. Hum. Genet.* **82**, 194–198. (doi:10.1016/j.ajhg.2007.09.010)

Kimura, R., Ohashi, J., Matsumura, Y., Nakazawa, M., Inaoka, T., Ohtsuka, R., Osawa, M. & Tokunaga, K. 2008 Gene flow and natural selection in oceanic human populations inferred from genome-wide SNP typing. *Mol. Biol. Evol.* **25**, 1750–1761. (doi:10.1093/molbev/msn128)

Krigbaum, J. & Datan, I. 1999 The deep skull of Niah. *Borneo* **5**, 13–17.

Lansing, J. S. *et al.* 2007 Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc. Natl Acad. Sci. USA* **104**, 16 022–16 026.

Lansing, J. S., Watkins, J. C., Hallmark, B., Cox, M. P., Karafet, T. M., Sudoyo, H. & Hammer, M. F. 2008 Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations. *Proc. Natl Acad. Sci. USA* **105**, 11 645–11 650.

Li, J. Z. *et al.* 2008 Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104. (doi:10.1126/science.1153717)

Mona, S., Tommaseo-Ponzetta, M., Brauer, S., Sudoyo, H., Marzuki, S. & Kayser, M. 2007 Patterns of Y-chromosome diversity intersect with the trans-New Guinea hypothesis. *Mol. Biol. Evol.* **24**, 2546–2555. (doi:10.1093/molbev/msm187)

Mona, S. *et al.* 2009 Genetic admixture history of eastern Indonesia as revealed by Y-chromosome and mitochondrial DNA analysis. *Mol. Biol. Evol.* **26**, 1865–1877. (doi:10.1093/molbev/msp097)

Mulvaney, D. J. & Kamminga, J. 1999 *Prehistory of Australia*. Washington, DC: Smithsonian Institution Press.

Novembre, J. & Di Rienzo, A. 2009 Spatial patterns of variation due to natural selection in humans. *Nat. Rev. Genet.* **10**, 745–755. (doi:10.1038/nrg2632)

O'Connell, J. F. & Allen, J. 2004 Dating the colonization of Sahul (Pleistocene Australia–New Guinea): a review of recent research. *J. Archaeol. Sci.* **31**, 835–853. (doi:10.1016/j.jas.2003.11.005)

Serre, D. & Paabo, S. 2004 Evidence for gradients of human genetic diversity within and among continents. *Genome Res.* **14**, 1679–1685. (doi:10.1101/gr.2529604)

Soares, P. *et al.* 2008 Climate change and postglacial human dispersals in Southeast Asia. *Mol. Biol. Evol.* **25**, 1209–1218. (doi:10.1093/molbev/msn068)

Spriggs, M. 2000 Out of Asia: the spread of southeast Asian Pleistocene and Neolithic maritime cultures in Island Southeast Asia and the western Pacific. *Mod. Quat. Res. SE Asia* **16**, 51–75.

Wall, J. D., Cox, M. P., Mendez, F. L., Woerner, A., Severson, T. & Hammer, M. F. 2008 A novel DNA sequence database for analyzing human demographic history. *Genome Res.* **18**, 1354–1361. (doi:10.1101/gr.075630.107)

Wallace, A. R. 1869 *The Malay Archipelago: the land of the orang-utan, and the bird of paradise*. London, UK: Macmillan and Company

Wang, J. 2003 Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* **164**, 747–765.

Wijsman, E. M. & Cavalli-Sforza, L. L. 1984 Migration and genetic population structure with special reference to humans. *Ann. Rev. Ecol. Syst.* **15**, 279–301. (doi:10.1146/annurev.es.15.110184.001431)

9

ACCURACY OF MOLECULAR DATING WITH THE RHO STATISTIC: DEVIATIONS FROM COALESCENT EXPECTATIONS UNDER A RANGE OF DEMOGRAPHIC MODELS

Cox, M.P. 2008. Accuracy of molecular dating with the rho statistic: Deviations from coalescent expectations under a range of demographic models. *Human Biology* 80:335-357.
http://www.jstor.org/stable/41466549

ABSTRACT The ρ statistic is commonly used to infer chronological dates for molecular lineages, especially from mitochondrial DNA sequences obtained in anthropological contexts. Since this approach was described 12 years ago, it has been applied to estimate molecular dates in more than 200 studies, including some published in top-tier journals. However, this method has not been well evaluated, and the accuracy of dates obtained from the ρ statistic remains unknown, especially for genetic data collected from populations with complex demographic histories. Here, molecular dates inferred from ρ are compared against coalescent expectations from a range of demographic models. This exercise reveals considerable inaccuracy. Molecular dates based on ρ have a slight downward bias with large asymmetric variance and commonly exhibit substantial type I error rates, where the true age of a lineage falls outside the 95% confidence bounds derived from the variance of ρ. Furthermore, demography proves to be a strong confounding factor in estimating molecular dates accurately, especially for populations in which bottlenecks, founder events, and size changes have played important historical roles. Therefore considerable caution should be applied to inferences made from molecular dates based on the ρ statistic, many of which may be misleading and warrant considerable skepticism.

Accuracy of Molecular Dating with the Rho Statistic: Deviations from Coalescent Expectations Under a Range of Demographic Models

MURRAY P. COX[1,2]

Abstract The ρ statistic is commonly used to infer chronological dates for molecular lineages, especially from mitochondrial DNA sequences obtained in anthropological contexts. Since this approach was described 12 years ago, it has been applied to estimate molecular dates in more than 200 studies, including some published in top-tier journals. However, this method has not been well evaluated, and the accuracy of dates obtained from the ρ statistic remains unknown, especially for genetic data collected from populations with complex demographic histories. Here, molecular dates inferred from ρ are compared against coalescent expectations from a range of demographic models. This exercise reveals considerable inaccuracy. Molecular dates based on ρ have a slight downward bias with large asymmetric variance and commonly exhibit substantial type I error rates, where the true age of a lineage falls outside the 95% confidence bounds derived from the variance of ρ. Furthermore, demography proves to be a strong confounding factor in estimating molecular dates accurately, especially for populations in which bottlenecks, founder events, and size changes have played important historical roles. Therefore considerable caution should be applied to inferences made from molecular dates based on the ρ statistic, many of which may be misleading and warrant considerable skepticism.

A common aim in many genetic studies is determining chronological ages from sequence data. Routine questions include estimating the time to the most recent common ancestor (TMRCA) (Cox et al. 2008a) or the age of a particular subclade within a gene tree (Cox 2005). A range of dating methods can be applied toward these goals, which typically utilize diversity within the taxonomic group of interest. Apart from simple analytical estimators (Tang et al. 2002), various simulation techniques founded in coalescent theory are also available (Griffiths and Tavaré 1994a; Tavaré et al. 1997). One common method—especially for mitochondrial DNA (mtDNA) data sets from anthropological contexts—uses the summary statistic rho (or ρ). Forster et al. (1996) first proposed this approach to molecular dating,

[1]Arizona Research Laboratories, University of Arizona, AZ 85721.
[2]Santa Fe Institute, Santa Fe, NM 87501.

Human Biology, August 2008, v. 80, no. 4, pp. 335–357.

KEY WORDS: GENETIC DATING, MOLECULAR CHRONOLOGY, ρ STATISTIC, MITOCHONDRIAL DNA.

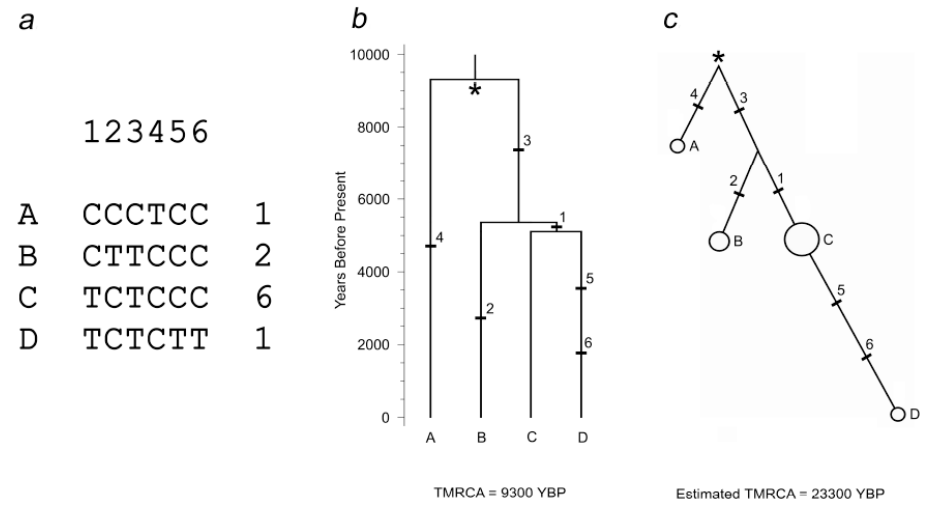

Figure 1. An example mtDNA control region data set. A 500-bp sequence was simulated under the standard coalescent for a sample of 10 individuals from a constant-size population with $N_0 = 10^4$. (a) Polymorphism table of six segregating sites (1–6) with derived C → T mutations producing four novel haplotypes (A–D). The number of individuals carrying each haplotype is shown to the right. (b) Gene tree topology showing haplotypes and known divergence times as generated by the coalescent. The time to the most recent common ancestor (TMRCA), indicated by an asterisk, is 9,300 years before present (YBP). (c) Median-joining network with topology identical to that in part b. The empirical distribution of polymorphisms from each unique lineage l is 1, 2, 2, 2, 2, 2, 2, 2, 2, 4. The statistic ρ is simply the mean number of polymorphisms weighted by the number of individuals carrying each lineage; that is, $\rho = 2.1$ (standard deviation = 0.92). The ρ statistic can be scaled by the mutation rate and sequence length to generate an estimate of the gene tree TMRCA (23,300 years; 95% confidence interval [3,250–43,400]).

which has since been applied in more than 200 studies. Increasing citation rates (Web of Science, June 2008) suggest that the ρ statistic is commonly viewed as an established and robust method for estimating molecular dates. Indeed, genetic dates inferred from the ρ statistic have appeared in at least one top-tier journal (Olivieri et al. 2006). It is therefore surprising that the accuracy of molecular dates inferred from the ρ statistic has not been rigorously evaluated, especially under a broad range of demographic models. Such a validation exercise is undertaken here.

The ρ statistic is defined as the average number of polymorphisms l observed along m unique lineages (weighted by sample size) stemming from a given ancestral node in a resolved genealogical tree (Figure 1). The ancestral node is often defined as the most recent common ancestor (MRCA) of a set of sequences (i.e., the sample MRCA) but can be any internal node chosen at will. Assuming that the mutation rate is known for the genetic locus under study, it is claimed that a simple equation translates the ρ statistic into an estimate of the molecular age

of the ancestral node in real chronological time. Confidence intervals on this date can be derived from the expected variance of ρ. This dating approach is relatively simple, but its accuracy remains unknown. However, it seems unlikely that the method is routinely accurate, particularly given the variance of genetic data sets generated under even the simplest demographic models. It is therefore reasonable to question the accuracy of molecular dates inferred by applying this method to empirical data sets with more complex demographic histories.

Here, I use coalescent modeling to determine the accuracy of molecular dates estimated from the ρ statistic. First, I examine the bias and variance of point estimates of dates obtained for a simple constant-size population. Next, I consider different methods of constructing confidence intervals and determine their error rates. Finally, I investigate the accuracy of molecular dates obtained under a range of more complex demographic models (e.g., those including population growth, bottlenecks, and structure). I conclude that genetic dates inferred from the ρ statistic are often slightly downward biased with large asymmetric variance and that they commonly exhibit substantial type I error rates, where the true date for more than 5% of data sets lies outside the 95% confidence bounds estimated from the variance of ρ. Furthermore, the widely held perception that the accuracy of ρ dating is independent of demography is false; type I error rates commonly depend on the demographic history of the study population.

Methods

Data Sets. Mitochondrial DNA control region data sets were simulated using the *n*-coalescent of Kingman (1982), as implemented in the program ms (Hudson 2002). Simulations were conditioned on a sequence length of 500 base pairs (bp), a female intergeneration interval of 26 years (Fenner 2005), and a transition mutation rate of 1.8×10^{-7}/bp/yr in accordance with Forster et al. (1996) and Saillard et al. (2000). As is true for real mtDNA data sets, simulated data were modeled without recombination, and for simplicity, homoplasy, infrequent transversions, and natural selection were also disregarded. The effects of deviating from this basic setup are addressed later.

Demographic Models. Molecular dating with the ρ statistic was performed under a series of demographic models (Table 1), which are intended to represent important aspects of real human demographic histories (i.e., simulations reflect realistic parameter values for a wide range of human population groups). All data points represent results from 10^4 coalescent data sets. First, a single Wright-Fisher deme of constant effective size ($N_0 = 10^3$) with a large sample size ($N_S = 100$) was modeled. Second, a single Wright-Fisher deme of constant effective size ($N_0 = 10^3$) was modeled with a range of sample sizes (N_S = 5–500). Third, a single Wright-Fisher deme with variable constant effective size ($N_0 = 10^3$–10^4) was modeled with $N_S = 100$. Fourth, an instantaneous bottleneck that ended 1,000

Table 1. Demographic Models Used to Explore the ρ Statistic

Demographic Model	N_S	N_A	N_0	N_B	T_B	T_G	T_S
Constant size	100		10^3				
Sampling error	5–500		10^3				
Variable constant size	100		10–10^4				
Bottleneck	100	10^3	10^3	1–10^3	0–10^4		
Growth	100	10–10^4	10–10^4			10^4	
Structure	100		(500, 500)				0–10^4

N_S, sample size.
N_A, ancestral population size.
N_0, modern population size.
N_B, population size during a bottleneck.
T_B, bottleneck duration.
T_G, time of onset of population growth.
T_S, time of onset of population structure.
All population sizes are defined as effective numbers of individuals; all times are defined in terms of years, not generations.

years ago was modeled for a population with ancestral and modern effective size $N_A = N_0 = 10^3$ and a sample size of $N_S = 100$. The state space of bottleneck models with durations of 0–10^4 years and effective sizes during the bottleneck equal to ~0–100% of the ancestral size N_A was explored, and the type I error rate surface was determined for a square grid containing 220 simulation points. Fifth, exponential population growth beginning 10^4 years ago from a range of ancestral effective sizes ($N_A = 10$–10^4) to a range of modern effective sizes ($N_0 = 10$–10^4) was modeled with $N_S = 100$. The exponential growth rate a is related to N_A and N_0 by

$$a = \frac{\ln(N_0/N_A)}{t_{\text{gen}}}. \tag{1}$$

The type I error rate surface was explored as before using a triangular grid containing 406 data points such that $N_0 \geq N_A$. Sixth, population structure was modeled for two Wright-Fisher demes with constant effective size ($N_0 = 500$). Demes were subdivided at a given time in the past ($t = 0$–10^4 years), migration rates m were set to 0 following subdivision, and $N_S = 100$ was sampled equally from the two populations at $t = 0$. This model can be seen as representing a case of cryptic subdivision.

Summary Statistics. The summary statistic ρ was defined by Forster et al. (1996) as the average number of mutations l along m unique haplotypes sampled from n individuals. Each lineage stems from a defined ancestral node given a resolved gene tree. This statistic has the expectation

$$E(\rho) = \frac{\sum_{i}^{m} n_i l_i}{n} \tag{2}$$

with variance

$$\sigma^2(\rho) = \frac{\sum_{i}^{m} n_i^2 l_i}{n^2}. \tag{3}$$

Forster et al. (1996) suggest that multiplication of the ρ statistic with a known mutation rate scalar yields an unbiased estimator of molecular age for the given ancestral node in real chronological time. This method assumes that mutations along the gene tree fit a molecular clock. For a known sequence length b in base pairs and a mutation rate per base pair μ, the mutation rate scalar s is

$$s = (\mu b)^{-1}. \tag{4}$$

Two-tailed confidence bounds can be placed on the ρ statistic (and hence on its estimator of chronological age) in two ways. The first and most widely used approach is to assume that the density of l is approximated by a normal distribution with mean and variance of ρ. Alternative confidence intervals (detailed later) can be calculated directly from the observed distribution of values of l. This approach, unlike the normal approximation, makes no assumptions about how polymorphisms are distributed among sampled lineages stemming from the ancestral node.

The ρ statistic is commonly calculated with the program Network (Bandelt and Forster 1995). However, the graphical interface of Network is poorly suited to analyzing large numbers of simulated data sets, and ρ was calculated here using custom command-line software (code available on request). This program has been validated against Bandelt and Forster's original application.

Genealogical Nodes. The ancestral node is defined here as the MRCA of a set of sampled individuals (i.e., the sample MRCA) simply because all genealogies have an MRCA regardless of their gene tree topology. This approach to validating molecular dating estimators is standard practice (cf. Basu and Majumdar 2003). Because the node representing the sample MRCA is not inherently special (with minor exceptions discussed later), results obtained from this validation exercise are broadly generalizable to other nodes in the genealogy. The TMRCA was extracted directly from coalescent simulations, in units of N_0 generations, using custom software (code available on request).

Accuracy of Molecular Dates. The precision of inferred genetic dates was determined by comparing estimated TMRCAs, calculated from the ρ statistic, with TMRCAs known directly from coalescent theory. Type I errors occur when

the known coalescent TMRCA falls outside the 95% confidence bounds of the TMRCA estimated from ρ. This should not occur by chance for more than 5% of data sets. To judge the accuracy of ρ-based genetic dates, I explored a range of type I error values (α) and used two methods of calculating confidence bounds. Simulated data sets with no segregating sites were excluded from error rate calculations. Such data sets necessarily have $\rho = s^2(\rho) = 0$, and hence associated TMRCA estimates also have a mean and variance of 0. True coalescent TMRCAs are necessarily nonzero (despite the absence of observed segregating sites) and would contribute to the type I error rate if these data sets were included. Because it is unusual for a researcher to estimate the TMRCA of an empirical data set with no observed variation, simulated data sets of this nature were similarly disregarded. Note, however, that excluding these data sets means that the following simulation results must be considered best-case scenarios from the perspective of type I error rates (i.e., this manipulation minimizes observed error).

Results

Here, I test the accuracy of genetic dates inferred from the ρ statistic (Forster et al. 1996; Saillard et al. 2000), first for a single constant-size population and subsequently for a range of more complex demographic models. Simulations were conditioned on optimal parameterizations (e.g., assuming a known mutation rate, no homoplasy) to determine best-case scenarios for the accuracy of ρ-based molecular dating. The effects of deviating from these optimal conditions are discussed later.

The basic test approach is relatively simple (Figure 1). A genetic data set is simulated under a given demographic model using coalescent theory. The genealogy of this data set is known absolutely, and therefore its TMRCA is also known exactly, without bias or variance. Subsequently, the ρ statistic is calculated for the same data set, and its confidence interval is either determined using the traditional approach (i.e., assuming a normal approximation for l) or determined directly from the empirical quantiles of l (see later). Because the sequence length and mutation rate are explicitly defined in the simulation model, the mutation rate scalar is also known exactly, without bias or variance. The product of the ρ statistic (or its confidence intervals) with the mutation rate scalar yields an estimate of the TMRCA (or its confidence intervals). The known coalescent TMRCA is compared with the ρ-based estimator of TMRCA and its confidence intervals. Because all other simulation parameters are known absolutely, any bias, variance, or error in the estimated TMRCA necessarily reflects inaccuracy associated entirely with the ρ-based dating method.

By way of example, Figure 1 shows a representative mtDNA control region data set drawn randomly from a small constant-size population generated using coalescent theory. From the table of polymorphisms (Figure 1a), a gene tree (Figure 1b) and network (Figure 1c) can be constructed. An asterisk indicates the MRCA on both the tree and the network, and from the coalescent the TMRCA is known

without error to be 9,300 years. Observe that an average of 2.1 polymorphisms (standard deviation of 0.92 polymorphism) occurs along each unique lineage l from the MRCA (the average is weighted by the number of individuals sampled with each lineage). Given a sequence length of 500 bp and a mutation rate of 1.8×10^{-7}/bp/yr, the expectation is to observe 1 mutation every 11,111 years; this is the mutation rate scalar described by Eq. (4). The product of the ρ statistic and the mutation rate scalar estimates the TMRCA of this data set at 23,300 years (95% confidence interval [3,250–43,400]). Here, the TMRCA estimated from ρ is more than twice as large as the true TMRCA, and although the 95% confidence interval inferred from the variance of ρ incorporates the true age (i.e., 9,300 years), the breadth of these confidence bounds is considerable. As will be seen, this is a relatively positive outcome from inferring molecular dates with the ρ statistic.

In the following sections, I examine several key characteristics of ρ-based molecular dating, including the accuracy of TMRCA estimates from a single panmictic population of constant size, the distribution of the variable l (and its effect on the calculation of confidence intervals), the breadth of 95% confidence intervals, and the distribution of type I error rates. Finally, I explore how type I error rates change with the demographic model under study, including effects of sampling error, population bottlenecks, growth, and structure.

Accuracy of TMRCA Estimates. Figure 2 compares known coalescent TMRCAs with estimated TMRCAs based on the ρ statistic for a constant-size population with $N_0 = 10^3$ and $N_S = 100$ (model 1). For this particular demographic model, molecular dates inferred from the ρ statistic tend to be slightly downward biased (solid line) relative to the true coalescent TMRCAs (Figure 2a). TMRCAs estimated from the ρ statistic were younger on average than the true values by 5,500 years. However, inferred dates also had considerable variance (95% confidence interval [−29,300–22,300]) (Figure 2b); observing inferred dates 1.6 times smaller or larger than the true age is not statistically unlikely. Although molecular dates based on the ρ statistic are significantly correlated with true values ($r = 0.71$, $P \ll 0.001$), this association accounts for only 50% of the observed variance. As a general rule, molecular dates based on the ρ statistic can underestimate or overestimate the true age of lineages by a substantial margin.

Calculation of Confidence Intervals. Confidence bounds are usually placed on molecular dates inferred with the ρ statistic using the expected variance of ρ [Eq. (3)]. Confidence intervals generated in this way are symmetric about ρ (and hence are symmetric about the inferred molecular date). Importantly, however, this approach assumes that the number of polymorphisms l occurring along m unique lineages stemming from a given ancestral node is drawn from a normal distribution. However, the distribution of l is not normal (Shapiro-Wilks test, all $P \ll 0.001$) (Figure 3). Instead, for large sample sizes, the empirical distributions are better approximated by Poisson densities with λ equal to the scaled population mutation rate θ. Because θ is not often known, this finding has little

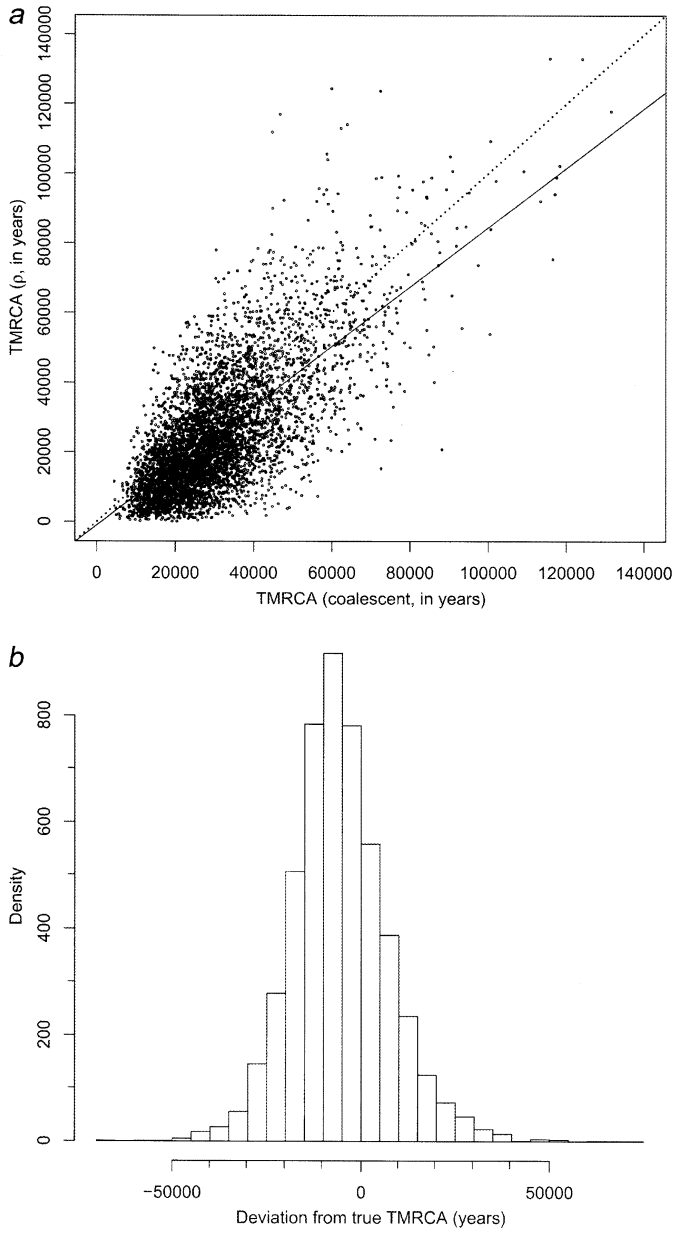

Figure 2. Accuracy of TMRCA estimates. (a) Comparison of known coalescent TMRCAs with mean estimates from the ρ statistic for a constant-size population with $N_0 = 10^3$. Estimates from the ρ statistic should be distributed along the dotted line but tend to be slightly less than expected. (b) For this particular demography, TMRCAs estimated from the ρ statistic average 5,500 years younger than their true values (95% confidence interval [−29,300–22,300]). Note, however, that the variance about the mean is considerable, and true dates can be much smaller or much larger than estimates inferred from ρ.

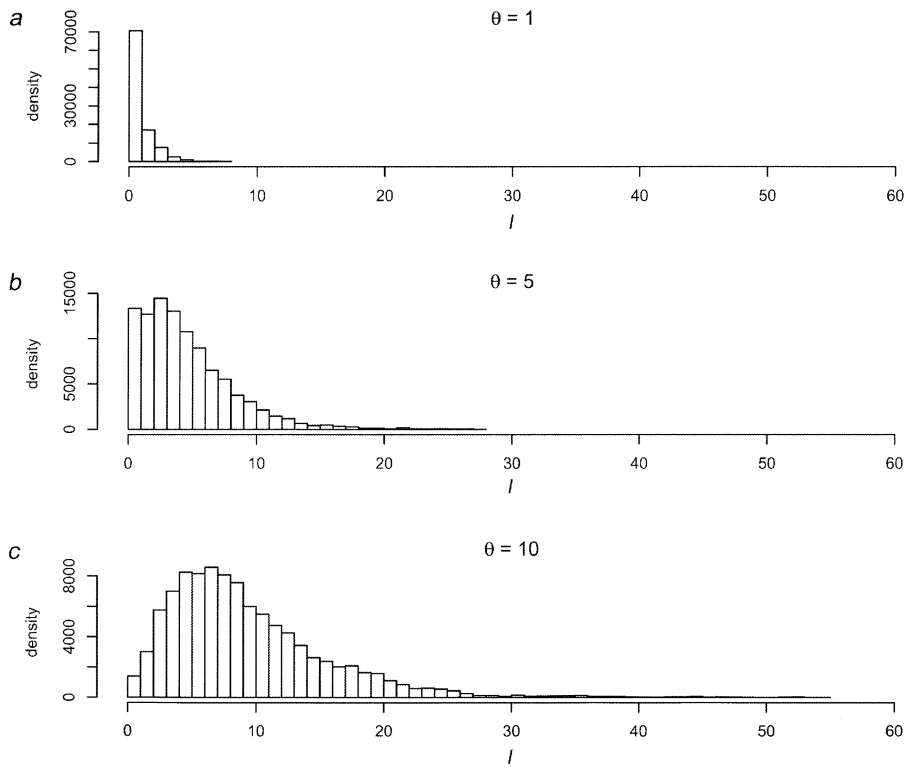

Figure 3. Distributions of the summary statistic l, the number of polymorphisms observed per unique lineage, weighted by the number of individuals carrying that lineage for $N_S = 100$. Distributions of l are plotted for (a) θ $(= N_e\mu) = 1$, (b) $\theta = 5$, and (c) $\theta = 10$. Densities are not normally distributed (Shapiro-Wilks test, all $P \ll 0.001$).

practical application; placing bounds on an unknown θ is nontrivial, given an empirical distribution of random variables drawn from an unknown distribution, $X \sim \text{Pois}(\lambda = \theta)$. However, this analysis does suggest that confidence intervals generated under the assumption of normality may be inaccurate, and as an alternative, quantile-based confidence intervals were also determined directly from the empirical distribution of l per simulated data set. Because the densities of l are often skewed (see Figure 3), quantile-based confidence intervals are generally asymmetric about ρ (and hence are asymmetric about the inferred molecular date). Note that this quantile approach makes no theoretical assumptions about the distribution of l, but other factors likely affect its accuracy, including the sample size N_S, the number of m unique lineages, and the number of observed segregating sites S. These last two factors depend highly on the underlying demography. Furthermore, confidence intervals generated by assuming a normal approximation are

both continuous and boundless $(-\infty, \infty)$. Conversely, confidence intervals determined empirically from the quantiles of l are discrete [i.e., they increase in integer units, the number of segregating sites observed along any given branch], and in practice they have relatively small, defined bounds [0, max(l)]. This explains some of the properties of quantile-based confidence intervals observed in the following sections (e.g., Figure 4), where both methods of constructing confidence intervals are evaluated more thoroughly.

Breadth of Confidence Intervals. Figure 4 shows the breadth of 95% confidence intervals, ranked by size. These intervals were drawn from a constant-size population with $N_0 = 10^3$ and $N_S = 100$ (model 1). Under this particular model (but see later for other models), both confidence intervals show similar patterns whether generated assuming a normal approximation (solid line) or determined directly from the quantiles of l (dashed line). Note that, as described earlier, normal approximation confidence intervals are continuous, whereas quantile-based confidence intervals increase in discrete integer units (Figure 4a). Conversely, all curves are continuous when the observed breadths of confidence intervals are displayed as ratios of the true TMRCA, a continuous variable (Figure 4b). For this particular demographic model, the median breadth of confidence intervals is about 1 times the true TMRCA (or here, about 28,800 years). However, observing confidence intervals as wide as 2 times the true TMRCA is not statistically uncommon (upper and lower dotted lines). Furthermore, confidence intervals are asymmetric about the true TMRCA; placement of the true TMRCA within these bounds is slightly upward biased. (Note that this is a separate issue from whether confidence intervals are symmetric or asymmetric about ρ.) As a general rule, the breadths of confidence intervals surrounding molecular dates generated from the ρ statistic are large. Consequently, it will often prove difficult to distinguish testable hypotheses that hinge on even moderately large differences between two chronological dates (e.g., whether the age of one clade significantly predates the age of another).

Accuracy of Confidence Intervals. Even disregarding the absolute size of confidence intervals, the general accuracy of these bounds is also unknown. Confidence intervals generated using a normal approximation for the variance of ρ may have significant error because l is not a random normal variable. Similarly, sample sizes, the number of unique lineages, and the number of observed segregating sites may all introduce error into confidence intervals inferred directly from the quantiles of l. Therefore both approaches may be expected a priori to exhibit high failure rates. Type I error rates were determined by simulating coalescent data sets, calculating confidence intervals for ρ-based molecular dates, and determining how often these bounds encompass the true TMRCA. Type I error rates are routinely large; Figure 5 shows how observed error rates change with defined error rates for a constant-size population with $N_0 = 10^3$ and $N_S = 100$ (model 1). The vertical dotted line shows where the type I error rate α is set to 0.05. Here, just by chance, 5% of the true TMRCA values should be observed falling outside the 95% confidence

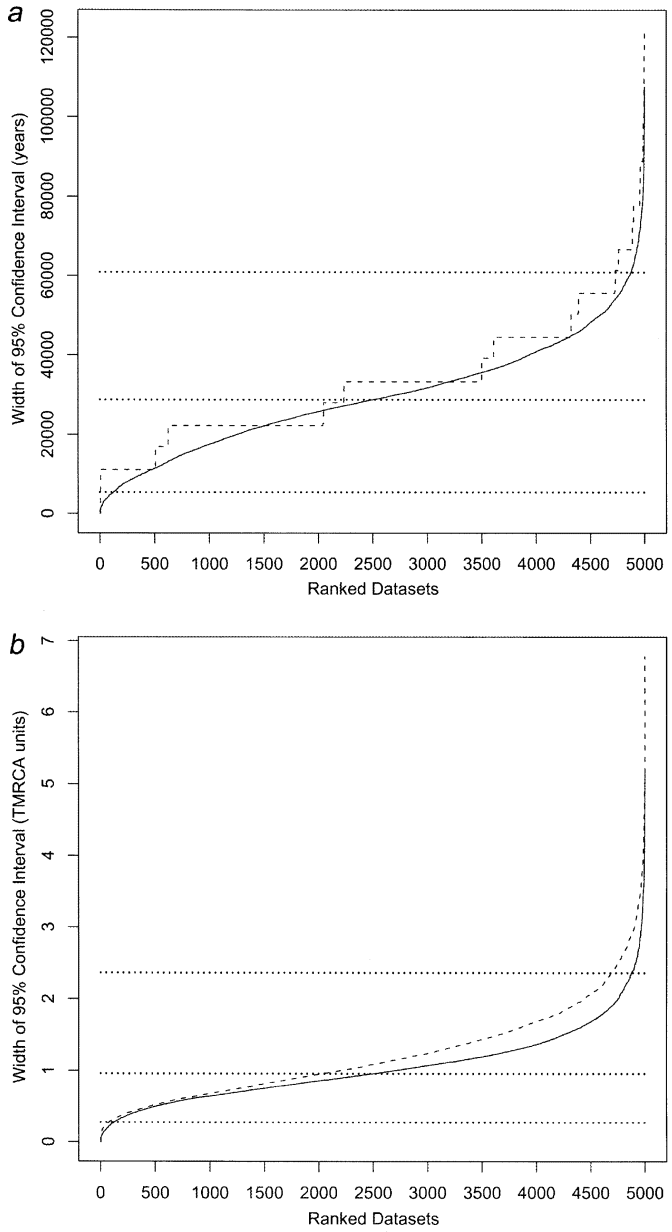

Figure 4. Breadth of 95% confidence intervals. Ninety-five percent confidence intervals ranked by size were calculated using the continuous normal approximation (solid line) or the discrete quantiles of l (dashed line) for 5,000 coalescent data sets. The breadth of the 95% confidence interval for the median data set (central horizontal dotted lines) is (a) 28,800 years or (b) about 1 times the true TMRCA. The range of sizes of normal approximation confidence intervals ($\alpha = 0.05$) is shown by the upper and lower horizontal dotted lines. Values were calculated for data sets drawn from a constant-size population ($N_0 = 10^3$) with a sample size of $N_S = 100$.

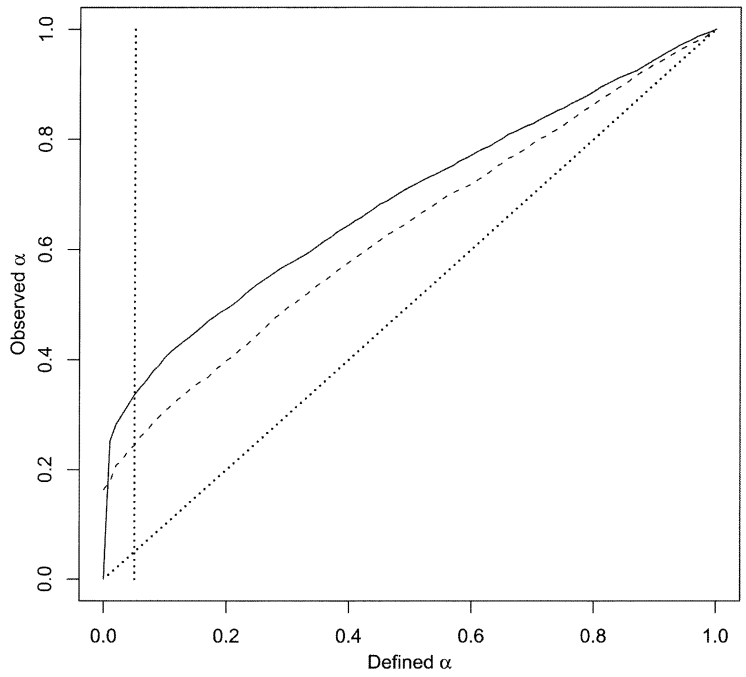

Figure 5. Type I error curves representing the accuracy of genetic dating using the ρ statistic. The diagonal dotted line indicates the expectation for a good test statistic (defined and observed α values are identical). Lines above this boundary indicate where confidence intervals exclude the true TMRCA more frequently than suggested by the defined α. For instance, when α is set to have an error rate of 5% (vertical dotted line), confidence intervals assuming a normal approximation (solid line) have an actual error rate of 34%, and confidence intervals generated from the quantiles of l (dashed line) have an actual error rate of 24%. Note that the error rate observed with quantile-based confidence intervals never falls to 0 even when $\alpha = 0$. The quantiles of l yield confidence intervals with discrete upper bounds (see text) that the true TMRCA commonly exceeds. Conversely, confidence intervals generated using the normal approximation are boundless and cannot be exceeded by any real number. Error rates were calculated from 5,000 coalescent data sets drawn from a constant-size population ($N_0 = 10^3$) with a sample size of $N_S = 100$.

intervals generated from ρ. This level of accuracy is expected for a good statistic, whose behavior on this plot would be denoted by the diagonal dotted line. However, both sets of confidence intervals have much higher error rates when α is set at 5%: 34% when using confidence intervals that assume a normal approximation (solid line) and 24% when using quantile-based confidence intervals (dashed line). Put differently, 34% and 24% of data sets, respectively, generated under this demographic model have a true TMRCA that is not encompassed by the 95%

confidence intervals calculated from the ρ statistic, even though such confidence intervals are routinely large in absolute terms. (Note that the true TMRCA is frequently, but not always, older.) Confidence intervals generated empirically from the distribution of l perform better than confidence intervals that assume a normal approximation, except for $\alpha \ll 0.01$. However, such a preference is marginal. As a general rule, confidence intervals calculated from the ρ statistic do not contain the true TMRCA for a substantial proportion of data sets, and the method therefore exhibits high type I error rates.

Sampling Error Effects. To determine the effects of sampling error, I modeled a single Wright-Fisher deme with constant effective size ($N_0 = 10^3$), with the sample size varying from 5 to 500 individuals (model 2). The effects of sampling error are relatively weak. However, observed type I error rates do vary between simulation runs for small sample sizes ($N_S < 25$), and the resulting molecular dates may therefore be unreliable. Rates are especially elevated when quantile-based confidence intervals are used. (This dependency on N_S was suggested earlier for quantile-based confidence intervals.) Error rates stabilize once $N_S \geq 25$, albeit at the high levels observed previously for simulations with $N_S = 100$ (i.e., 28% and 34% for quantile-based and normal approximation confidence intervals, respectively). Although genetic data simulated under many demographic models exhibit high error rates, these seem to be driven by factors other than sampling error for moderate to large N_S. As a general rule, sampling error does not elevate the observed error rate of ρ-based molecular dates for commonly encountered sample sizes ($N_S \simeq 25$–50).

Population Size Effects. Figure 6 shows error rates of molecular dates calculated from the ρ statistic for populations with different constant effective sizes ($N_0 = 10$–10^4) (model 3). As noted, molecular dates calculated for a constant-size population with $N_0 = 10^3$ have a 34% error rate when using confidence intervals that assume a normal approximation and a 24% error rate when using quantile-based confidence intervals (see Figure 5). Here, the error rate is seen to increase with the effective size of the population. For $N_0 = 10^4$ (i.e., the global effective size of modern humans; Haigh and Maynard Smith 1972), confidence intervals generated with the normal approximation have an error rate approaching 75%, and confidence intervals inferred directly from the quantiles of l have an error rate of 44%. As a general rule, confidence intervals do not contain the true TMRCA for a substantial proportion of data sets, and this error increases with the effective size of the population under study.

For completeness, note that the graph in Figure 6 (variable N_0 with fixed mutation rate μ) is identical to the graph for fixed N_0 with variable mutation rate μ. This is because θ ($= N_e\mu$) is a linear combination of these two factors. Because the mutation rate is considered a known quantity in empirical studies that use ρ-based dating, the effects of variable mutation rates are not explicitly explored here.

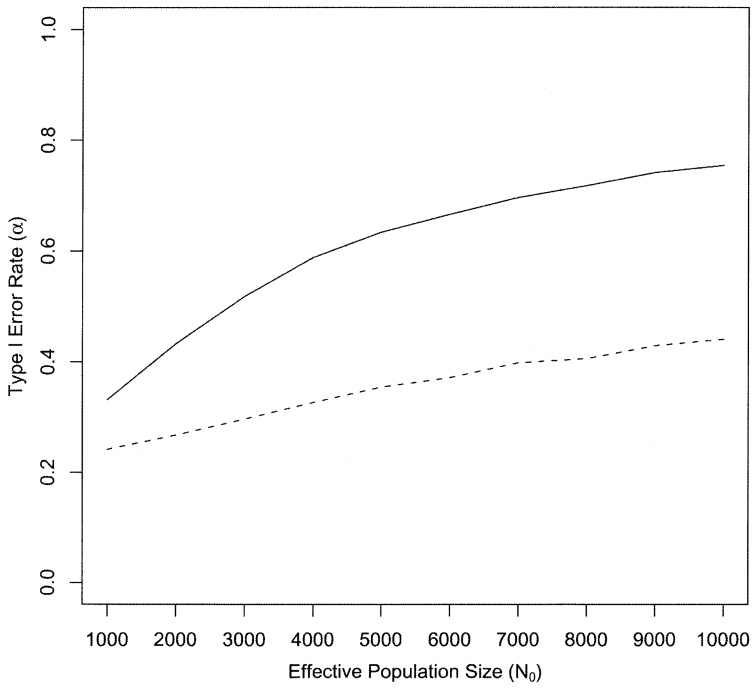

Figure 6. Type I error rates for constant-size populations with variable N_0. Error rates were determined from 95% confidence intervals calculated assuming a normal approximation (solid line) and the empirical quantiles of l (dashed line).

It is worth emphasizing, however, that mutation rates and type I error rates also exhibit a dependent relationship. This necessarily means that attempts to estimate mutation rates using observed ρ and a known genetic date would be subject to the same errors discussed throughout this report.

Population Bottleneck Effects. Figure 7 shows the error rate surface for a series of demographic bottlenecks of variable extent and duration (model 4). As illustrated in Figure 6, the two methods of calculating confidence intervals can react quite differently to the same demographic conditions, but both are still error prone. Ninety-five percent confidence intervals generated using the normal approximation have higher error rates following extreme bottlenecks (i.e., bottlenecks in which the effective size decreases to less than about 20% of the population's initial size), whereas error rates improve in this same situation when confidence intervals are inferred directly from the quantiles of l. Note, however, that observed error rates are still substantially above their expected value ($\alpha = 0.05$). As a general rule, molecular dates have high error rates that vary with the extent and duration of population bottlenecks.

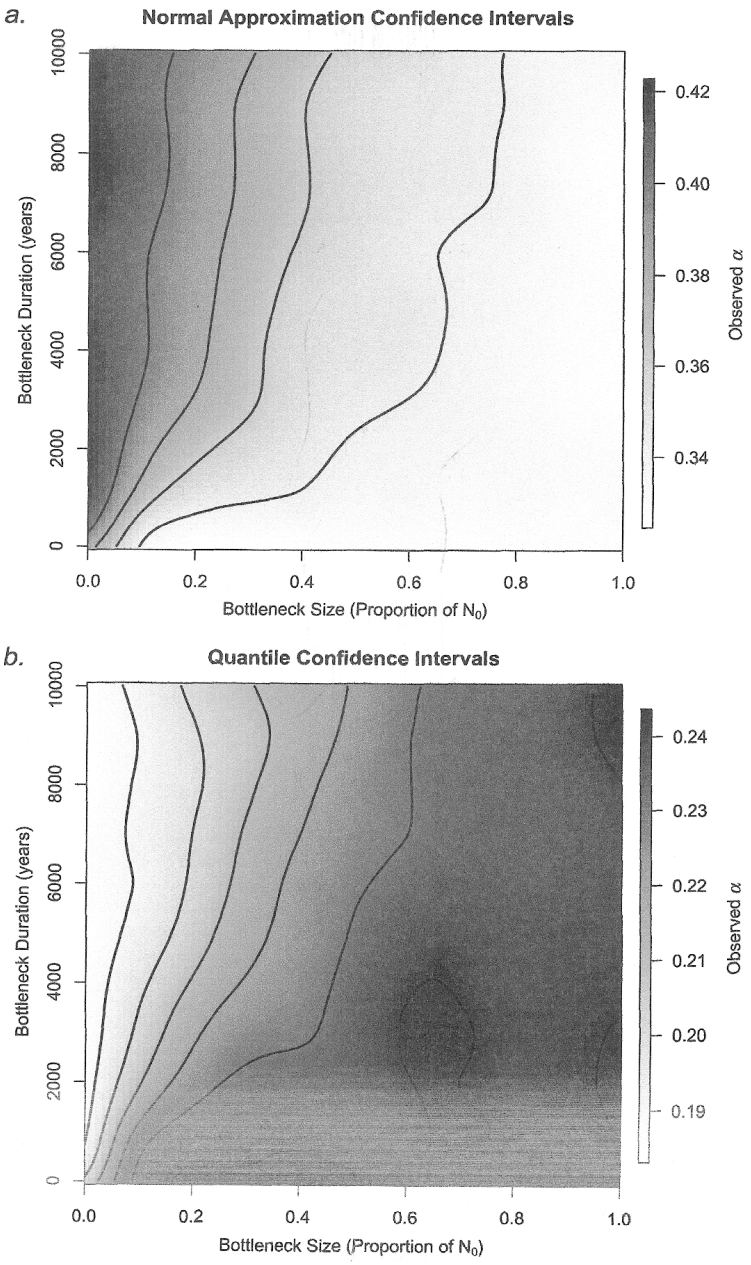

Figure 7. Type I error surface for a series of bottleneck models. Observed type I error rates are shown for demographic models with bottlenecks of variable extent (x-axis) and duration (y-axis) that ended 10^3 years ago. Error rates were determined using 95% confidence intervals calculated (a) assuming a normal approximation and (b) the empirical quantiles of l.

Population Growth Effects. Figure 8 shows the error rate surface for a series of population models with variable growth rates (model 5); error rates along the diagonal (i.e., instances of no population growth) independently illustrate the trend shown in Figure 6. Again, the two methods of calculating confidence intervals react differently to the same demographic conditions, although both are still error prone. Long-term effective population sizes, approximated here by the harmonic mean of effective population sizes per generation (Wright 1938), are correlated strongly with error rates, regardless of whether confidence intervals are generated using a normal approximation ($r = 0.47$, $P \ll 0.001$) or are quantile-based ($r = 0.69$, $P \ll 0.001$). Nevertheless, these account for only 22% and 47% of the observed variance, respectively. Conversely, growth rates may or may not have an effect on observed error rates. Growth rates (a) are correlated with type I error rates obtained using normal approximation confidence intervals ($r = 0.56$, $P \ll 0.001$) and explain 31% of the observed variance (cf. Basu and Majumdar 2003), but there is no correlation observed when quantile-based confidence intervals are used ($r = 0.050$, $P = 0.31$). For strong population growth, observed error rates decrease slightly when quantile-based confidence intervals are used, but this effect is not observed with confidence intervals that assume a normal approximation. It is worth emphasizing, however, that error rates remain too high for most biologically realistic rates of population growth (two- to tenfold growth) (Voight et al. 2005; Wall and Przeworski 2000). As a general rule, ρ-based molecular dates exhibit high error rates when calculated on genetic data sets drawn from growing populations.

Population Structure Effects. To determine the effects of demic structure, I modeled a population split for two Wright-Fisher demes with constant effective size ($N_0 = 500$) and absolute subdivision ($m = 0$) beginning at a specified time in the past ($t = 0$–10^4 years) (model 6). Population substructure has a relatively weak effect on observed error rates, which for practical intents, are identical to those of a constant-size population without substructure. As a general rule, error rates of molecular dating with the ρ statistic are unaffected by simple population subdivision, but they show the same high rates of error as those for populations with constant size. This finding holds true for the island model simulated here, but results may differ under more biologically realistic scenarios (e.g., under an isolation-with-migration model).

Discussion

The accuracy of genetic dates inferred from the ρ statistic is explored here for a single population of constant size as well as for a range of more complex demographic models. Key findings are that molecular dates estimated from the ρ statistic are often slightly downward biased with large asymmetric variance and that they commonly exhibit substantial type I error rates, where the true TMRCA falls outside the confidence bounds inferred from the variance of ρ. This dating

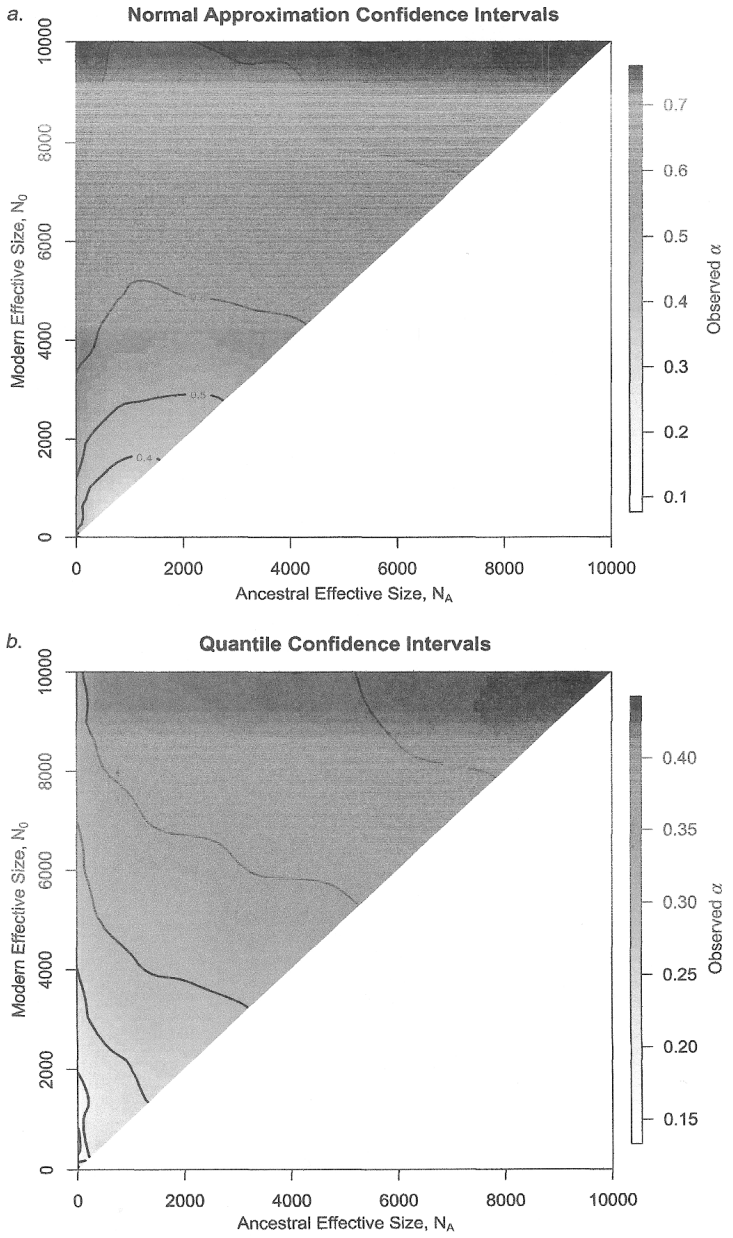

Figure 8. Type I error surface for a series of growth models. Observed type I error rates are shown for demographic models with growth from variable ancestral effective size (x-axis) to variable modern effective size (y-axis) over 10^4 years. Diagonals represent a series of demographic models with constant effective size (i.e., equivalent to but calculated independently of those in Figure 6). Error rates were determined using 95% confidence intervals calculated (a) assuming a normal approximation and (b) the empirical quantiles of l.

method is often considered model-free; that is, the demographic history of the study population is usually treated as unimportant for inferring ρ-based molecular dates. In practice, however, demography can be a strong confounding influence on molecular dates generated using this method. For instance, type I error rates increase with effective size (constant or growing) and extreme bottlenecks (or equivalently, constricted founder events). Only sampling error and simple island-model population structure do not seem to alter the observed error rates of molecular dates generated using ρ.

Importantly, several practical confounding factors were explicitly disregarded in these simulations, which can therefore be considered best-case scenarios. First, mutation rates are never known with absolute certainty. Forster et al. (1996) and Saillard et al. (2000) advocate a transition mutation rate of 1.8×10^{-7}/bp/yr. However, this mutation rate is approximately fourfold lower than other estimates, such as 6.6×10^{-7}/bp/yr (Sigurðardóttir et al. 2000), 6.1×10^{-7}/bp/yr (Howell et al. 2003), and 7.6×10^{-7}/bp/yr (Santos et al. 2005). For the 275-bp mtDNA region considered by Forster et al. (1996) and Saillard et al. (2000), these mutation rates would be equivalent to 1 mutation every 5,510, 5,960, or 4,790 years (cf. the usual mutation scalar of 1 transition every 20,180 years). However, there is considerable debate over the meaning (and relative utility) of mutation rates calculated from pedigree and phylogenetic studies (Santos et al. 2005). Because the molecular date estimator is a simple product of ρ and the mutation rate scalar, this uncertainty will be reflected in molecular dates inferred for real mtDNA assemblages. Second, molecular dating with the ρ statistic assumes constancy in the mutation rate (i.e., a molecular clock). In reality, mutation rates can change over time (Ho and Larson 2006; Pulquério and Nichols 2007). Third, mutation rates are not spatially uniform. Instead, rates vary by position in the mtDNA control region (Hasegawa et al. 1993; Malyarchuk and Derenko 2001; Rosset 2007). Fourth, high mutation rates and rate heterogeneity elevate levels of recurrent mutation (homoplasy) (Basu and Majumdar 2003) and multiple observed character states (heteroplasmy) (Bendall et al. 1996). These in turn may introduce uncertainty, in the form of reticulation, into the topology of the genealogical network. Missing data can similarly degrade the accuracy of inferred networks (Joly et al. 2007). Because the ρ statistic can be calculated only on a fully resolved topology (i.e., a gene tree), an approximation must otherwise be obtained by averaging ρ across all possible trees consistent with a given network topology containing reticulation. Uncertainty increases for molecular dates generated from ρ as resolution of the underlying gene tree decreases. Fifth, the choice of locus can affect observed sequence diversity and, consequently, associated dating estimates. Not all loci will exhibit similar coalescent times, even for genetic regions in complete linkage disequilibrium. For instance, estimated TMRCAs inferred from different portions of the mtDNA genome are at least somewhat variable (Non et al. 2007). Sixth, selection on genes will mediate patterns of polymorphism in noncoding regions that are linked by means of genetic hitchhiking (Kivisild et al. 2006). Given the important metabolic roles played by many mitochondrially encoded

proteins, this factor may violate the key assumption of selective neutrality used in the coalescent simulations performed here. At least some of these factors will add additional uncertainty, perhaps substantially so, to the bias, variance, and error rates of molecular dating with the ρ statistic as described here.

Much of this difficulty is not unexpected, and it is worth emphasizing that molecular date estimates will always be accompanied by substantial uncertainty. The variance of the age of a node on a gene tree is large relative to its mean and does not decrease appreciably with increased sample size (Griffiths and Tavaré 1994b; Tang et al. 2002). For instance, the population MRCA of a locus can be sampled from a panmictic deme with probability $P(\text{TMRCA} \mid n) = (n-1)/(n+1)$ (Kliman and Hey 1993; Saunders et al. 1984); that is, deep nodes in a genealogy are sampled with high probability even with small sample sizes. If the length of a sequence available from a locus is constrained (as it is with mtDNA), the number of segregating sites that occur on deeper branches of the tree will also be limited. Sampling additional individuals cannot improve this estimate. Increased sampling may allow for improved resolution of younger clades on a genealogical tree, but with the proviso that the number of polymorphisms occurring in younger clades will likely be smaller, which may in turn inflate the variance of young molecular age estimates. In practice, genetic dates will always have large confidence bounds; it is clearly desirable, however, that these should contain the true TMRCA given the limits of α.

If molecular dates are a necessary research outcome, there are alternatives to the ρ statistic. Every approach has its own limitations, including computational expense and model assumptions that may not hold for the biological system under study. Also, note that not all these methods have been explored as thoroughly as the ρ statistic is here. However, they will still prove preferable to the ρ statistic in many circumstances. The three main approaches for calculating ages of particular nodes on a genealogical tree are point estimators, approximate coalescent methods, and full coalescent methods. Point estimators apply simple analytical equations and do not require complex computation; one such method is detailed by Tang et al. (2002). Conversely, approximate coalescent methods determine the likelihood of summaries of the data by simulation under a prescribed set of demographic models; an example of this approach is the method of Tavaré et al. (1997). Full coalescent methods determine the likelihood of the full data set, also by simulation under a prescribed set of demographic models; one example is the method described by Griffiths and Tavaré (1994a). This particular approach, implemented in the program GeneTree, is an especially popular option (Griffiths 2007). Bayesian methods, such as that implemented in Beast, are another alternative (Drummond and Rambaut 2007). Of course, the end goal of having dated genetic lineages is often to understand the temporal relationships of the populations that carry them. In such cases, calculating population divergence times may prove to be the more direct approach, and full coalescent methods implemented under a fairly general demographic model are currently available to infer this parameter (Hey and Nielsen 2004, 2007; Nielsen and Wakeley 2001).

If the ρ statistic must be used for molecular age estimates, inferred dates should at the least be reported as mean values with 95% confidence intervals. This is usual statistical practice, but it has frequently been disregarded in many population genetic studies where a mean date with standard deviations, or worse still, just a mean date, is more commonly the only value reported. This practice can easily mislead readers into believing that published molecular dates have more certainty than is actually attributable to them. Preferably, application of the ρ statistic for molecular dating should be accompanied by independent age estimates generated using other methods. The approaches discussed here are reasonable alternatives and have been validated, and at least some have been applied extensively in the wider genetics literature. Finally, existing publications that report molecular dates calculated from ρ should be evaluated somewhat critically. Their conclusions may not prove to be robust in light of the more complex analyses that are currently being undertaken on large multilocus genomic data sets (Garrigan et al. 2007), and errors in molecular dating may underpin a sizable proportion of these discrepancies.

By means of example, consider a case study in which molecular dating with the ρ statistic has influenced anthropological interpretations of human prehistory. In 1998 the ρ statistic was used to date a specific mtDNA lineage, the Polynesian motif, for a small population sample ($N_S = 6$) from eastern Indonesia (Richards et al. 1998). This lineage occurs at high frequency in the greater Pacific (Cox et al. 2008b), where it has been tentatively associated with the spread of Neolithic migrants, ultimately from mainland Southeast Asia (Bellwood 2005). The stepwise spread of Neolithic migrants and technologies appears to have occurred relatively quickly, and archeological signals of the Neolithic are observed only in eastern Indonesia from about 3,500 years ago (Spriggs 1989, 2003). However, on the basis of ρ dating, Richards and colleagues placed the MRCA of the Polynesian motif in eastern Indonesia at 17,000 years before present (95% confidence interval [5,500–34,500]). These confidence bounds do not encompass the earliest archeological signals of the Neolithic in eastern Indonesia, and therefore an association between the Polynesian motif and the expansion of Neolithic agriculturalists has been, and continues to be, questioned (Hill et al. 2007; Oppenheimer and Richards 2001, 2002; Oppenheimer 2003). However, there are several concerns about the validity of this molecular date. First, the sample size ($N_S = 6$) is insufficiently large to avoid considerable sampling error, a point noted previously by Cox (2005). Second, molecular dates inferred from the ρ statistic can vary substantially from the true age, and therefore the Polynesian motif may not be as old as 17,000 years. (It may, of course, be older.) Third, there are sizable error rates associated with molecular dates generated from ρ, and the 95% confidence interval reported by Richards and colleagues probably does not encompass the true age of the Polynesian motif lineage. Such error may be especially likely if the Polynesian motif is indeed associated with a Neolithic expansion; population growth appreciably elevates type I error rates. Consequently, the molecular date inferred for the Polynesian motif using the ρ statistic seems to have considerable uncertainty and does

not appear to convincingly disavow an association between the Polynesian motif and the expansion of Neolithic agriculturalists. The original date, however, persists in the literature more than a decade after it first appeared in print.

In conclusion, it is difficult to advocate for continued application of the ρ statistic to infer molecular dates, and considerable caution should be exercised in drawing historical conclusions from such inferences. Coalescent simulations show that the ρ statistic produces molecular dates that are often slightly downward biased with large asymmetric variance and that commonly exhibit substantial type I error rates. When 95% confidence intervals are used, observed type I error rates far exceeding 5% are not uncommon under all the demographic models considered here. Furthermore, the widely held perception that the accuracy of ρ-based molecular dates is not dependent on demography proves false. Demography is a strong contributing factor to the uncertainty of molecular dates generated using this method, especially for populations in which bottlenecks, founder events, and population size changes have been important historical processes. This, of course, encompasses most human population groups.

Acknowledgments I thank S. Kobourov (University of Arizona) for access to the dispersed computing grid maintained by the Department of Computer Science; and H. Norton, F. Mendez, and A. Russell (University of Arizona) for helpful discussions.

Received 18 April 2008; revision received 27 May 2008.

Literature Cited

Bandelt, H.-J., and P. Forster. 1995. *Network: Phylogenetic Network Software.* Available at http://www.fluxus-engineering.com/netwinfo.htm

Basu, A., and P. Majumdar. 2003. A comparison of two popular statistical methods for estimating the time to most recent common ancestor (TMRCA) from a sample of DNA sequences. *J. Genet.* 82:7–12.

Bellwood, P. 2005. *The First Farmers: The Origins of Agricultural Societies*. Oxford: Blackwell.

Bendall, K. E., V. A. Macaulay, J. R. Baker et al. 1996. Heteroplasmic point mutations in the human mtDNA control region. *Am. J. Hum. Genet.* 59:1276–1287.

Cox, M. P. 2005. Indonesian mitochondrial DNA and its opposition to a Pleistocene era origin of proto-Polynesians in Island Southeast Asia. *Hum. Biol.* 77:179–188.

Cox, M. P., F. L. Mendez, T. M. Karafet et al. 2008a. Testing for archaic hominin admixture on the X chromosome: Model likelihoods for the modern human *RRM2P4* region from summaries of genealogical topology under the structured coalescent. *Genetics* 178:427–437.

Cox, M. P., A. J. Redd, T. M. Karafet et al. 2008b. A Polynesian motif on the Y chromosome: Population structure in remote Oceania. *Hum. Biol.* 79:525–535.

Drummond, A. J., and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7:214.

Fenner, J. N. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* 128:415–423.

Forster, P., R. Harding, A. Torroni et al. 1996. Origin and evolution of Native American mtDNA variation: A reappraisal. *Am. J. Hum. Genet.* 59:935–945.

Garrigan, D., S. B. Kingan, M. Metni Pilkington et al. 2007. Inferring human population sizes, divergence times, and rates of gene flow from mitochondrial, X, and Y chromosome resequencing data. *Genetics* 177:2195–2207.

Griffiths, R. C. 2007. *Genetree v. 9.0.* Available at http://www.stats.ox.ac.uk/ griff/software.html

Griffiths, R. C., and S. Tavaré. 1994a. Ancestral inference in population genetics. *Stat. Sci.* 9:307–319.

Griffiths, R. C., and S. Tavaré. 1994b. Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* 46:131–159.

Haigh, J., and J. Maynard Smith. 1972. Population size and protein variation in man. *Genet. Res. Cambridge* 19:73–89.

Hasegawa, M., A. Di Rienzo, T. D. Kocher et al. 1993. Toward a more accurate time scale for the human mitochondrial DNA tree. *J. Mol. Evol.* 37:347–354.

Hey, J., and R. Nielsen. 2004. Multilocus methods for estimating population sizes, migration rates, and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167:747–760.

Hey, J., and R. Nielsen. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl. Acad. Sci. USA* 104:2785–2790.

Hill, C., P. Soares, M. Mormina et al. 2007. A mitochondrial stratigraphy for island Southeast Asia. *Am. J. Hum. Genet.* 80:29–43.

Ho, S. Y., and G. Larson. 2006. Molecular clocks: When times are a-changin'. *Tr. Genet.* 22:79–83.

Howell, N., C. B. Smejkal, D. A. Mackey et al. 2003. The pedigree rate of sequence divergence in the human mitochondrial genome: There is a difference between phylogenetic and pedigree rates. *Am. J. Hum. Genet.* 72:659–670.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.

Joly, S., M. I. Stevens, and B. J. Van Vuuren. 2007. Haplotype networks can be misleading in the presence of missing data. *Syst. Biol.* 56:857–862.

Kingman, J. F. C. 1982. *On the Genealogy of Large Populations*. Sheffield, U.K.: Applied Probability Trust.

Kivisild, T., P. Shen, D. P. Wall et al. 2006. The role of selection in the evolution of human mitochondrial genomes. *Genetics* 172:373–387.

Kliman, R. M., and J. Hey. 1993. DNA sequence variation at the *period* locus within and among species at the *Drosophila melanogaster* complex. *Genetics* 133:375–387.

Malyarchuk, B. A., and M. V. Derenko. 2001. Variation of human mitochondrial DNA: Distribution of hot spots in hypervariable segment I of the major noncoding region. *Russ. J. Genet.* 37:823–832.

Nielsen, R., and J. Wakeley. 2001. Distinguishing migration from isolation: A Markov chain Monte Carlo approach. *Genetics* 158:885–896.

Non, A. L., A. Kitchen, and C. J. Mulligan. 2007. Identification of the most informative regions of the mitochondrial genome for phylogenetic and coalescent analyses. *Mol. Phylogenet. Evol.* 44:1164–1171.

Olivieri, A., A. Achilli, M. Pala et al. 2006. The mtDNA legacy of the Levantine early Upper Paleolithic in Africa. *Science* 314:1767–1770.

Oppenheimer, S. 2003. Austronesian spread into Southeast Asia and Oceania: Where from and when? In *Pacific Archaeology: Assessments and Prospects*, C. Sand (ed). Noumea, New Caledonia: Le Cahiers de l'Archéologie en Nouvelle-Calédonie, 55–70.

Oppenheimer, S., and M. Richards. 2001. Fast trains, slow boats, and the ancestry of the Polynesian Islanders. *Sci. Prog.* 84:157–181.

Oppenheimer, S., and M. Richards. 2002. Polynesians: Devolved Taiwanese rice farmers or Wallacean maritime traders with fishing, foraging, and horticultural skills. In *Examining the Farming/Language Dispersal Hypothesis*, P. Bellwood and C. Renfrew (eds.). Cambridge, U.K.: McDonald Institute for Archaeological Research, 287–297.

Pulquério, M. J., and R. A. Nichols. 2007. Dates from the molecular clock: How wrong can we be? *Tr. Ecol. Evol.* 22:180–184.

Richards, M., S. Oppenheimer, and B. Sykes. 1998. MtDNA suggests Polynesian origins in eastern Indonesia. *Am. J. Hum. Genet.* 63:1234–1236.

Rosset, S. 2007. Efficient inference on known phylogenetic trees using Poisson regression. *Bioinformatics* 23:142–147.

Saillard, J., P. Forster, N. Lynnerup et al. 2000. MtDNA variation among Greenland Eskimos: The edge of the Beringian expansion. *Am. J. Hum. Genet.* 67:718–726.

Santos, C., R. Montiel, B. Sierra et al. 2005. Understanding differences between phylogenetic and pedigree-derived mtDNA mutation rate: A model using families from the Azores Islands (Portugal). *Mol. Biol. Evol.* 22:1490–1505.

Saunders, I. W., S. Tavaré, and G. A. Watterson. 1984. On the genealogy of nested subsamples from a haploid population. *Adv. Appl. Prob.* 16:471–491.

Sigurðardóttir, S., A. Helgason, J. R. Gulcher et al. 2000. The mutation rate in the human mtDNA control region. *Am. J. Hum. Genet.* 66:1599–1609.

Spriggs, M. 1989. The dating of the Island Southeast Asian Neolithic: An attempt at chronometric hygiene and linguistic correlation. *Antiquity* 63:587–613.

Spriggs, M. 2003. Chronology of the Neolithic transition in Island Southeast Asia and the Western Pacific: A view from 2003. *Rev. Archeol.* 24:57–80.

Tang, H., D. O. Siegmund, P. Shen et al. 2002. Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. *Genetics* 161:447–459.

Tavaré, S., D. Balding, R. C. Griffiths et al. 1997. Inferring coalescence times from DNA sequence data. *Genetics* 145:505–518.

Voight, B. F., A. M. Adams, L. A. Frisse et al. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl. Acad. Sci. USA* 102:18,508–18,513.

Wall, J. D., and M. Przeworski. 2000. When did the human population size start increasing? *Genetics* 155:1865–1874.

Wright, S. 1938. Size of a population and breeding structure in relation to evolution. *Science* 87:430–431.

10

TESTING FOR ARCHAIC HOMININ ADMIXTURE ON THE X-CHROMOSOME: MODEL LIKELIHOODS FOR THE MODERN HUMAN *RRM2P4* REGION FROM SUMMARIES OF GENEALOGICAL TOPOLOGY UNDER THE STRUCTURED COALESCENT

Cox, M.P., F.L. Mendez, T.M. Karafet, M. Metni Pilkington, S.B. Kingan, G. Destro-Bisol, B.I. Strassmann, and M.F. Hammer. 2008. Testing for archaic hominin admixture on the X-chromosome: Model likelihoods for the modern human *RRM2P4* region from summaries of genealogical topology under the structured coalescent. *Genetics* 178:427-437.
http://doi.org/10.1534/genetics.107.080432

ABSTRACT A 2.4-kb stretch within the *RRM2P4* region of the X chromosome, previously sequenced in a sample of 41 globally distributed humans, displayed both an ancient time to the most recent common ancestor (*e.g.*, a TMRCA of ~2 million years) and a basal clade composed entirely of Asian sequences. This pattern was interpreted to reflect a history of introgressive hybridization from archaic hominins (most likely Asian *Homo erectus*) into the anatomically modern human genome. Here, we address this hypothesis by resequencing the 2.4-kb *RRM2P4* region in 131 African and 122 non-African individuals and by extending the length of sequence in a window of 16.5 kb encompassing the *RRM2P4* pseudogene in a subset of 90 individuals. We find that both the ancient TMRCA and the skew in non-African representation in one of the basal clades are essentially limited to the central 2.4-kb region. We define a new summary statistic called the minimum clade proportion (p_{mc}), which quantifies the proportion of individuals from a specified geographic region in each of the two basal clades of a binary gene tree, and then employ coalescent simulations to assess the likelihood of the observed central *RRM2P4* genealogy under two alternative views of human evolutionary history: recent African replacement (RAR) and archaic admixture (AA). A molecular-clock-based TMRCA estimate of 2.33 million years is a statistical outlier under the RAR model; however, the large variance associated with this estimate makes it difficult to distinguish the predictions of the human origins models tested here. The p_{mc} summary statistic, which has improved power with larger samples of chromosomes, yields values that are significantly unlikely under the RAR model and fit expectations

better under a range of archaic admixture scenarios.

SUPPLEMENTARY MATERIALS

http://www.genetics.org/content/178/1/427/suppl/DC1

DOI: 10.1534/genetics.107.080432

Testing for Archaic Hominin Admixture on the X Chromosome: Model Likelihoods for the Modern Human *RRM2P4* Region From Summaries of Genealogical Topology Under the Structured Coalescent

Murray P. Cox,* Fernando L. Mendez,† Tatiana M. Karafet,* Maya Metni Pilkington,‡ Sarah B. Kingan,* Giovanni Destro-Bisol,§ Beverly I. Strassmann and Michael F. Hammer†,‡,**,[1]**

**ARL Division of Biotechnology, †Department of Ecology and Evolutionary Biology and ‡Department of Anthropology, University of Arizona, Tucson, Arizona 85721, §Department of Animal and Human Biology, University of Rome "La Sapienza," and Istituto Italiano di Antropologia, 00185 Rome, Italy and **Department of Anthropology, University of Michigan, Ann Arbor, Michigan 48109*

Manuscript received August 13, 2007
Accepted for publication November 11, 2007

ABSTRACT

A 2.4-kb stretch within the *RRM2P4* region of the X chromosome, previously sequenced in a sample of 41 globally distributed humans, displayed both an ancient time to the most recent common ancestor (*e.g.*, a TMRCA of ~2 million years) and a basal clade composed entirely of Asian sequences. This pattern was interpreted to reflect a history of introgressive hybridization from archaic hominins (most likely Asian *Homo erectus*) into the anatomically modern human genome. Here, we address this hypothesis by resequencing the 2.4-kb *RRM2P4* region in 131 African and 122 non-African individuals and by extending the length of sequence in a window of 16.5 kb encompassing the *RRM2P4* pseudogene in a subset of 90 individuals. We find that both the ancient TMRCA and the skew in non-African representation in one of the basal clades are essentially limited to the central 2.4-kb region. We define a new summary statistic called the minimum clade proportion (p_{mc}), which quantifies the proportion of individuals from a specified geographic region in each of the two basal clades of a binary gene tree, and then employ coalescent simulations to assess the likelihood of the observed central *RRM2P4* genealogy under two alternative views of human evolutionary history: recent African replacement (RAR) and archaic admixture (AA). A molecular-clock-based TMRCA estimate of 2.33 million years is a statistical outlier under the RAR model; however, the large variance associated with this estimate makes it difficult to distinguish the predictions of the human origins models tested here. The p_{mc} summary statistic, which has improved power with larger samples of chromosomes, yields values that are significantly unlikely under the RAR model and fit expectations better under a range of archaic admixture scenarios.

FOSSIL, archaeological, and genetic data all lend support to the hypothesis that *Homo sapiens* originated in Africa (McBrearty and Brooks 2000; McDougall *et al.* 2005; Garrigan and Hammer 2006). With the acceptance of the role of Africa in our species' origin, there is now increasing interest in the question of how the ancestral population that gave rise to anatomically modern humans (AMH) was structured. Did AMH emerge from a single, isolated African deme or from a subdivided ancestral population with gene flow among subpopulations? A related question is whether the expanding AMH population completely replaced or interbred with then contemporaneous archaic populations such as Neanderthals and *H. erectus* (Eswaran 2002; Templeton 2002; Garrigan *et al.* 2005a,b; Plagnol and Wall 2006; Wall and Hammer 2006). Early studies of nonrecombining regions such as mtDNA and the Y chromosome were consistent with the hypothesis of a single origin followed by complete replacement, sometimes referred to as the recent African replacement (RAR) model. While many of the more recently published DNA sequencing studies of X-linked and autosomal loci are also concordant with this RAR model, a growing number are not (Evans *et al.* 2006; Garrigan and Hammer 2006).

Garrigan *et al.* (2005b) published one of the first studies to posit recent admixture between AMH and an archaic human population. A resequencing study of 2.4 kb of the ribonucleotide reductase M2 pseudogene 4 (*RRM2P4*) in a sample of 41 globally diverse humans identified an unusual pattern of nucleotide polymorphism compared with most of the human genome. The reconstructed gene tree revealed two clades of allelic sequences that were estimated to have diverged ~2

[1]*Corresponding author:* ARL Division of Biotechnology, Life Sciences South, University of Arizona, Tucson, AZ 85721.
E-mail: mfh@u.arizona.edu

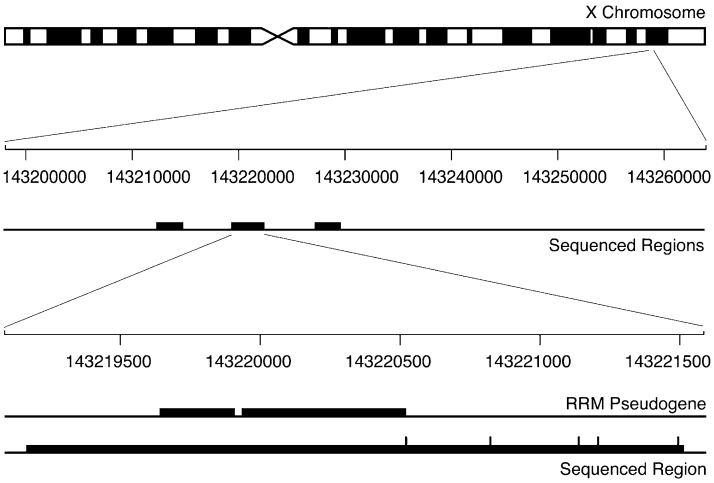

FIGURE 1.—Location of the *RRM2P4* locus on the X chromosome (top bar), positions of the three sequenced regions (middle bar), and placement of the *RRM2P4* pseudogene fragments relative to the *RRM2P4* central region (bottom bars). Vertical bars in the central region indicate SNPs defining the two basal clades. All coordinates correspond to the human genome UCSC March 2006 build.

million years ago (MYA). One clade with very little sequence variation was specific to Asians, while the other more diverse clade of *RRM2P4* sequences resembled a pattern typical of human variation and was globally distributed. By genotyping a diagnostic SNP in a large sample of humans, the divergent "Asian clade" was shown to be frequent in Southeast Asians and nearly absent in sub-Saharan Africans. The greater genealogical depth in Asia led to the hypothesis that *RRM2P4* is a genomic remnant of introgressive hybridization from an Asian archaic population (*H. erectus*) into AMH groups expanding from Africa. However, despite the increasing availability of genomic data, the uniqueness of such deep times to the most recent common ancestor (TMRCA) and of gene trees with non-African basal clades remains unclear. Importantly, we still do not understand the likelihood of such genealogies under either the RAR model or models involving admixture with archaic humans.

Garrigan *et al.*'s original *RRM2P4* study was limited by a paucity of sequence data and primarily qualitative analyses. Here, we extend earlier research by sequencing the 2.4-kb *RRM2P4* region in 131 African and 122 non-African individuals. We also determine the extent of sequence that exhibits the unusual pattern of polymorphism by selectively resequencing three DNA fragments totaling 5.6 kb within a 16.5-kb window flanking *RRM2P4*. We infer the likelihood of the observed *RRM2P4* genealogy using a suite of summary statistics and Monte Carlo coalescent simulations under the RAR model and a range of archaic admixture (AA) models (NORDBORG 2000; PLAGNOL and WALL 2006). The RAR models are parameterized by approximate Bayesian computation (ABC) conditioned on resequence data from an additional data set of 19 neutral, unlinked X-chromosomal loci. This resequence data set represents 12 Mb from 19 regions of the X chromosome, which are unlinked from genes (and each other) by medium to high rates of recombination (≥1.0 cM/Mb). These loci, which represent selectively neutral X-chromosomal diversity, will be published elsewhere. However, using these data, we show here that simple RAR models often produce TMRCA values that are similar to that of *RRM2P4*, while genealogies with skewing of basal clade membership toward non-Africans remain statistical outliers.

SUBJECTS AND METHODS

Regions sequenced: Resequence data for the *RRM2P4* locus were generated within a "trio" design (Figure 1), which is an economical approach to jointly ascertain detailed polymorphism patterns and larger-scale linkage-disequilibrium profiles (GARRIGAN *et al.* 2005b). We sequenced three genomic segments of 1725 bp [University of California, Santa Cruz (UCSC) March 2006 genome coordinates 143,212,138–143,213,863], 2341 bp (coordinates 143,219,154–143,221,495), and 1601 bp (coordinates 143,227,057–143,228,658), which were separated by unsequenced regions of 5291 and 5562 bp, respectively. Our central subregion is the same as that described by GARRIGAN *et al.* (2005b).

Sampling: The *RRM2P4* central region was sequenced in 131 African and 122 non-African individuals (panel A). These samples were chosen without prior information about *RRM2P4* lineage status. DNA samples representing Mandenka from Senegal ($n = 16$), Biaka Pygmies from the Central African Republic ($n = 16$), Khoisan from Namibia ($n = 9$), French Basque ($n = 16$), Han Chinese ($n = 16$), and Nasioi from Bougainville ($n = 16$) were purchased from the Centre d'Etude du Polymorphisme Humain (CANN *et al.* 2002). Samples of Baka Pygmies from Cameroon ($n = 23$) were provided by Giovanni Destro-Bisol and the Dogon from Mali ($n = 32$) were provided by Beverly Strassmann. Samples from the Dinka of southern Sudan ($n = 21$)

were collected in Tucson, Arizona, with informed consent. Samples from three Siberian populations, the Selkups ($n = 32$), Forest Nentsi ($n = 28$), and Tundra Nentsi ($n = 3$), were described previously (Karafet *et al.* 2002). Non-population-based samples ($n = 25$) from the Y chromosome consortium cell lines (Y Chromosome Consortium 2002) were also included in this panel. Resequence data for the full trio were generated in a second panel of 42 African and 48 non-African individuals (panel B) from three African groups (Khoisan, Mandenka, and Biaka) and three non-African groups (French Basque, Han, and Nasioi) (samples as described above). All sampling protocols were approved by the Human Subjects Committee at the University of Arizona and by the institutions of all collaborators who provided DNA samples.

Recombination analysis and genetree dating: Rates of linkage disequilibrium across the sequence were determined using LDhat (McVean and Spencer 2006). The *RRM2P4* central region shows only limited signs of recombination, and a most parsimonious tree was reconstructed by breaking low-frequency reticulations. The TMRCA of the tree and the age of its polymorphisms were estimated with Genetree (Griffiths 2007). Genetree employs a full maximum-likelihood method that is based on the standard coalescent (Kingman 1982) and assumes an infinite-sites mutational model. Likelihood surfaces for the population mutation rate, θ, and the population growth rate, β, were generated under a panmictic single-deme model and an island model of population structure. TMRCA values were inferred from maximum-likelihood parameterizations under both models.

Demographic models: Summaries of the *RRM2P4* central region were compared with values obtained from simulations under a structured coalescent (Nordborg 1997; Hudson 2002) to determine the likelihood of the observed genealogy. We employed a framework for human demography similar to that developed by Plagnol and Wall (2006), but modified (see below) to yield the number of segregating sites, *S*, consistent with an independent data set of 19 X chromosome loci sequenced in the same individuals (panel B). We present results from two demographic models: a two-deme RAR model and an AA scenario similar to Nordborg's (2000) isolation and admixture model (Figure 2). Alternative RAR models with varying levels of recent population subdivision (*e.g.*, one-deme and six-deme models) did not differ significantly from the two-deme RAR model reported here (our unpublished data). These models are not meant to represent the true history of human populations, but they do let us explore the effects of expansion and replacement *vs.* archaic admixture on patterns of genomic variation. Coalescent dates (scaled by $3N_e$ generations) were translated to chronological time using a 28-year mean intergeneration interval (Fenner 2005). The use of a lower estimate of the intergeneration interval (*e.g.*, a 25-year mean intergeneration interval) does not alter our conclusions (our unpublished data).

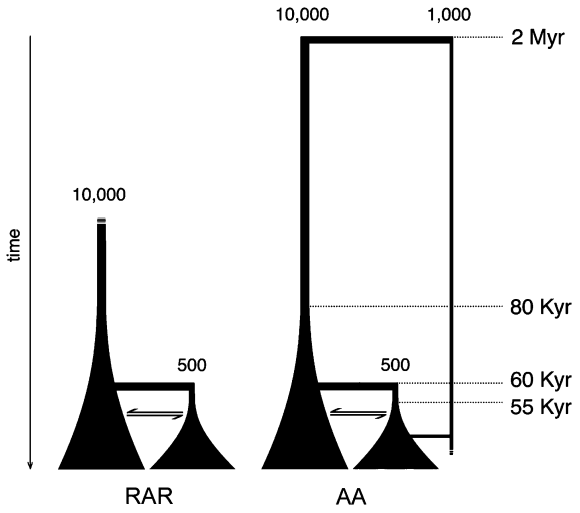

Figure 2.—Schematics of two demographic models. The recent African replacement (RAR) model grows exponentially from a 10^4 effective population size at 80 KYA, and 500 individuals found the non-African population at 60 KYA, experience a constant-sized bottleneck for 5 KYA, and grow exponentially from 55 KYA. The ancient admixture (AA) model includes introgression from an ancestral hominin source with constant population size of 10^3; the population leading to that of anatomically modern humans and ancestral hominins diverged 2 MYA. Modern effective sizes and the intercontinental migration rate were inferred by approximate Bayesian computation.

The RAR model depicts two panmictic, exponentially growing Wright–Fisher demes representing African and non-African populations. Growth begins in the African deme 80,000 years ago from a single ancestral population ($N_e = 10^4$) and continues until it reaches its current effective size. A small group ($N_e = 500$) splits from the African deme 60,000 years ago to form the non-African deme. This subgroup experiences a bottleneck for 5000 years before expanding exponentially to its current effective size. Modern effective sizes and the intercontinental migration rate were parameterized by approximate Bayesian computation (see below).

The AA model incorporates instantaneous admixture from an ancestral hominin source into the RAR model described above. An ancestral hominin population with constant effective population size of 10^3 splits from the ancestors of modern humans at 2 MYA. We considered a range of admixture rates (0–5%) from ancestral hominins to modern humans in Asia, coupled with a series of admixture times (10–55 KY before present). We could not infer an optimized AA model because the paucity of candidate loci is not sufficient both for model training and for subsequent statistical testing.

Model fitting via approximate Bayesian computation: We parameterized the RAR model to reflect what is known about deep human demography and, consequently, to produce simulated data sets that mimic real genomic data sets. Parameter inference rapidly becomes computationally intractable at high-dimensional state spaces, such as those associated with complex demographic models. Therefore, we fixed some demographic parameters that have been inferred elsewhere, *e.g.*, the time of onset of population growth and non-African bottleneck size (Plagnol and Wall 2006, Figure 2). However, a lower-dimensional state space of modern population size, N_0, and the intercontinental migration rate per generation, m, were inferred explicitly by ABC (Beaumont *et al.* 2002).

Essentially, we generated 10^4 coalescent simulations for each of 10^5 different sets of demographic parameters, $\Theta = \{N_0, m\}$, that were drawn randomly from two uniform distributions, $N_0 \in U[10^4, 10^5]$ and $m \in U[10^{-11}, 10^{-8}]$. Each set of coalescent simulations was compared with resequence data from 19 noncoding regions on the X chromosome (*i.e.*, panel B, 42 African and 48 non-African individuals; data to be published elsewhere). We estimated θ ($= 3N_0\mu$) for each demographic parameter set using the average mutation rate of these 19 loci, 8.3×10^{-10}/bp/year (range 4.8×10^{-10}–1.6×10^{-9}), inferred from sequence divergence (assuming a human/chimpanzee divergence time of 6 MYA). Furthermore, we chose to condition our ABC inference on the mean number of segregating sites, S, that was observed across the additional data set of 19 X chromosome loci ($\bar{S} = 26$). This statistic was chosen because it varies primarily with the mutation rate (a known parameter) and tree depth (a parameter we wished to infer). We selected the 0.01% of random demographic parameter sets, Θ, whose coalescent data sets produced an average number of segregating sites closest to the observed value ($\bar{S} = 26$). Mean values for N_0 and m were drawn from this subset of demographic parameter sets. These values represent best-fit estimates for the modern effective size, N_0, and migration rate, m, of the real-world demography underlying our observed 19 X chromosome neutral-locus genomic data set.

Test statistics: We calculated the approximate likelihoods of two summary statistics: the TMRCA and the minimum proportion of Africans in one of the two basal clades. Values were inferred from 10^5 replicates under all demographic models. We note that these tests are conservative, because both summaries are determined directly from coalescent genealogies. Resolution of the underlying genealogy is constrained by S for real data sets (*cf.* Nordborg 2000). The distribution of TMRCA values was extracted from the output of Hudson's (2002) ms using custom software (code available on request).

Here, we also define a new summary statistic, the minimum clade proportion (p_{mc}), which characterizes the minimum quantity of individuals from a specified geographic region or ethnic group (such as Africans) in each of the two basal clades of a binary gene tree (Figure 3). (This statistic is described fully in the appendix.) The p_{mc} statistic has an intuitive interpretation. Consider a binary tree of African and non-African sequences (Figure 3); clades (C_1, C_2) can be defined as the two basal branches that diverge from the coalescent of all sampled individuals. Here, the number of African chromosome copies is ($k_1 = 1$, $k_2 = 6$) and the total number of chromosome copies is ($n_1 = 4$, $n_2 = 10$). From Equation A1, the minimum clade proportion is the lesser of ($k_1/n_1 = \frac{1}{4}$, $k_2/n_2 = \frac{6}{10}$); here, $p_{\mathrm{mc}} = \frac{1}{4}$. The p_{mc} statistic was calculated directly from coalescent genealogies using custom software (code available on request).

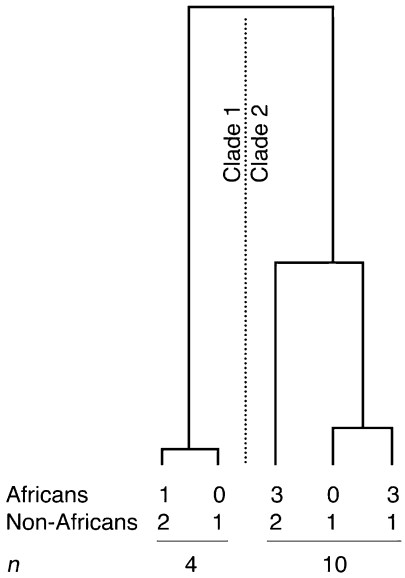

Figure 3.—Schematic of the minimum clade-proportion statistic, p_{mc}. See text for details.

RESULTS

Patterns of DNA sequence variation within and around *RRM2P4*: We sequenced 2.4 kb of the central *RRM2P4* segment, which encompasses the processed pseudogene (Figure 1), in 131 African and 122 non-African individuals. A total of 22 segregating sites were identified, producing 23 unique haplotypes (supplemental Table 1 at http://www.genetics.org/supplemental/). Similar to the previous study of Garrigan *et al.* (2005b), levels of nucleotide diversity are statistically higher in non-Africans ($\theta_\pi = 0.136$) than in Africans ($\theta_\pi = 0.0768$; $P \ll 0.001$; Table 1). Also, the central subregion still shows minimal evidence of recombination despite a sixfold increase in the size of the data set (supplemental Figure 1 at http://www.genetics.org/supplemental/).

TABLE 1

Estimates of summary statistics for the *RRM2P4* central and flanking regions

Region	Deme	n	Length	S	θ_w/bp (%)	Lower 5% quantile	θ_π/bp (%)	95% confidence interval
5′	Global	83	1727	5	0.058	—	0.068	(0.052, 0.081)
	African	37	1727	5	0.069	—	0.084	(0.059, 0.098)
	Non-African	46	1727	3	0.040	—	0.056	(0.034, 0.073)
Central	Global	253	2320	22	0.155	0.120	0.110	(0.093, 0.126)
	African	131	2320	16	0.126	0.071	0.077	(0.066, 0.087)
	Non-African	122	2320	15	0.120	0.104	0.136	(0.105, 0.163)
3′	Global	83	1612	7	0.088	—	0.061	(0.037, 0.081)
	African	37	1612	4	0.060	—	0.081	(0.044, 0.108)
	Non-African	46	1612	6	0.085	—	0.041	(0.014, 0.067)

Statistical confidence was determined from 10^4 bootstrap replicates. SNP diversity, θ_W, is significantly lower in the flanking regions than in the central region ($\alpha < 0.05$). African θ_π-values do not differ significantly among regions ($\alpha > 0.05$), but non-African θ_π- and θ_W-values are significantly higher in the central region. Within sequenced regions, non-Africans have significantly higher θ_π relative to Africans in the central region only ($P \ll 0.001$, underlined). This effect is not observed in the 5′ ($P = 0.966$) or 3′ regions ($P = 0.957$).

As before, this low level of recombination permitted partial reconstruction of a single nonreticulating gene tree for the central region (Figure 4). This genealogy has the same two unusual characteristics described by Garrigan *et al.* (2005b): a deep TMRCA and a basal clade composed almost entirely of Asian sequences. For example, of the 253 individuals resequenced for the central region, 21 are in the leftmost divergent clade in Figure 4 ("clade A"), and 20 of these are from East Asia and Oceania. A single Dogon from Mali was the only African member of this clade. This haplotype carries a SNP (1639) that occurs in both basal clades. An origin of this haplotype through recombination rather than homoplasy is more likely given the low mutation rate of the *RRM2P4* locus together with its moderate rate of recombination. Because the parental clade A form was found only in Asia and this haplotype is shared with a Melanesian, we suggest that this recombinant lineage may have originated in Asia and migrated recently to Africa.

We used both molecular-clock and coalescent approaches to estimate the TMCRA of the *RRM2P4* central subregion gene tree. Outgroup comparisons reveal an average of 22 nucleotide substitutions between all human and chimpanzee central *RRM2P4* region sequences. Given an average of 8.53 nucleotide differences observed between the two human *RRM2P4* lineages (*i.e.*, the average number of mutations between sequences across the base of the human gene tree), we estimate that the two deepest human clades diverged ~2.33 MYA (assuming a 6-MYA human–chimpanzee divergence time). We also inferred the TMRCA of the *RRM2P4* central subregion using a full maximum-likelihood method under both a panmictic and an island model. These models yielded TMRCA values of 1.24 and 2.88 MYA, which bracket the molecular-clock date (data not shown).

To determine the length of sequence within and around the *RRM2P4* locus that shows the unusual genealogical features, we sequenced two additional fragments of 1725 and 1601 bp that flank the central region (Figure 1) in a subset of African and non-African individuals (panel B). Levels of nucleotide diversity in African *vs.* non-African populations were more similar to average patterns for the genome, with African values ($\theta_\pi = 0.0836$ and 0.0805) greater than non-African values ($\theta_\pi = 0.0556$ and 0.0412) in the 5′- and 3′-flanking regions, respectively (Table 1). Recombination rates are low (0.43 cM/Mb) for the 5′ and central *RRM2P4* subregions, but are substantially elevated (16 cM/Mb) between the central and 3′ subregions (supplemental Figure 1 at http://www.genetics.org/supplemental/). This recombination hotspot effectively unlinks the central and 3′ subregions, which therefore have largely independent evolutionary histories. A network representing sequences of the entire 16.4-kb region illustrates the decoupling of these two genomic regions (supplemental Figure 2 at http://www.genetics.org/supplemental/). The unusual genealogy is less apparent in the 5′ fragment despite some linkage disequilibrium with the central region. (Recombination rates in these regions are about one-third the X chromosome average; supplemental Figure 1.) However, only five segregating sites were identified in the 5′-flanking region, a reduction of 63% over the central region diversity (Table 1), and no derived polymorphisms were identified in the 5′ region on chromosomes carrying central region clade A haplotypes (supplemental Table 2 at http://www.genetics.org/supplemental/). Because the unusual pattern of polymorphism is most apparent in the central region, we focus further analyses solely on this portion of the sequenced region.

RAR model parameters: The RAR model was parameterized by ABC. Because the model cannot be parameterized

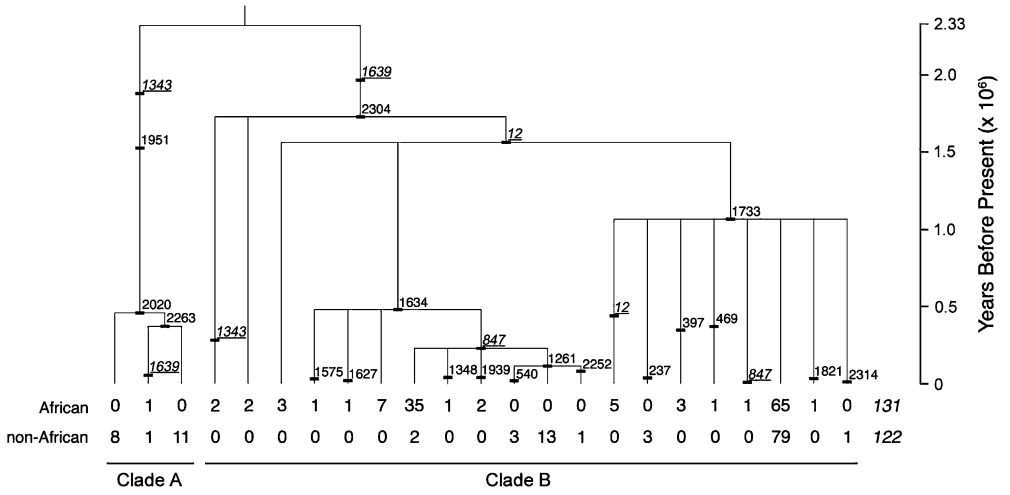

FIGURE 4.—Time-scaled gene tree of the *RRM2P4* central region. Polymorphisms are labeled with the nomenclature of GARRIGAN *et al.* (2005b) and are equivalent to that publication's Figure 1. Italicized and underlined polymorphisms indicate probable reticulation points. The proportions of African and non-African individuals carrying each lineage are indicated beneath the tree. Note the overrepresentation of non-Africans in the leftmost basal clade ("clade A"). Molecular-clock estimates of the TMRCA and mutation ages are shown in millions of years on the vertical axis.

on the test locus (here, *RRM2P4*), we conditioned the model on a separate training data set of 19 independent X chromosome noncoding regions. These resequence data are too extensive for detailed description here and are the subject of a separate publication. However, basic summaries of these 19 loci relevant to the current analysis are presented in supplemental Table 3 at http://www.genetics.org/supplemental/. The two-deme RAR model was parameterized by generating 10^5 random N_0 and m values (see full description in SUBJECTS AND METHODS) and accepting the 0.01% that best matched the number of segregating sites observed in 19 independent X chromosome noncoding loci (supplemental Figure 3 at http://www.genetics.org/supplemental/). The optimal parameters were inferred as a modern effective size, N_0, of 12,300 (range 12,000–12,500) and an intercontinental migration rate per generation, m, of 3.62×10^{-9} (range 6.75×10^{-10}–8.24×10^{-9}). Using parameter values at the extremes of these ranges had little effect on the following statistical analyses (our unpublished data). Importantly, the parameterized RAR model produces simulated data sets that yield summaries (such as growth rates and effective sizes) that are consistent with other demographic inferences (*e.g.*, from Genetree). Although there is strictly no way to assess the true history of our samples, this best-fit RAR model is a reasonable first approximation for human demographic history, as reconstructed from the X chromosome.

Assessing the uniqueness of the *RRM2P4* genealogy: The uniqueness of the *RRM2P4* central region was estimated using a simulation-based summary-likelihood approach under the RAR and AA models (Figure 2). Both the distribution of TMRCA values and the proportion of African individuals in one of the two basal clades were inferred under both models. The mean TMRCA under the two-deme model was 1.15 MYA, with dates exceeding 2.12 MYA being statistical outliers ($\alpha = 0.05$). While the maximum-likelihood method assuming panmixia returned a mean TMRCA value similar to that produced under the RAR model [*i.e.*, $P(\text{TMRCA} > 1.24 \times 10^6 \mid \text{RAR}) = 0.350$], both the molecular clock [$P(\text{TMRCA} > 2.33 \times 10^6 \mid \text{RAR}) = 0.030$] and the maximum-likelihood method assuming an island model [$P(\text{TMRCA} > 2.88 \times 10^6 \mid \text{RAR}) = 0.008$] yielded TMRCA values that are unusually old. Under the AA model, the TMRCA distribution is shifted deeper in time relative to RAR models, and genealogies exceeding 3 MYA would be expected under our archaic admixture scenario (Figure 5, a and b). The molecular-clock-based estimate of the true TMRCA is not an outlier under the AA model as specified here [$P(\text{TMRCA} > 2.33 \times 10^6 \mid \text{AA}) = 0.140$].

The geographical distribution of lineages on the central *RRM2P4* genealogy is also skewed: one of the two basal clades (Figure 4, clade A) is found infrequently in Africans (0.048) but commonly in non-Africans (0.952). This pattern can also be compared with the empirical distribution of p_{mc}, as determined from the additional data set of 19 X chromosome loci, among which the smallest minimum African clade proportion is 0.214. The likelihood of observing the same, or a more extreme, proportion of Africans (*i.e.*, $\frac{1}{21}$) in a basal clade was statistically significant under the RAR model [$P(p_{\text{mc}} \leq 0.048 \mid \text{RAR}) = 0.031$], but not so under the AA model [$P(p_{\text{mc}} \leq 0.048 \mid \text{AA}) = 0.24$]. To

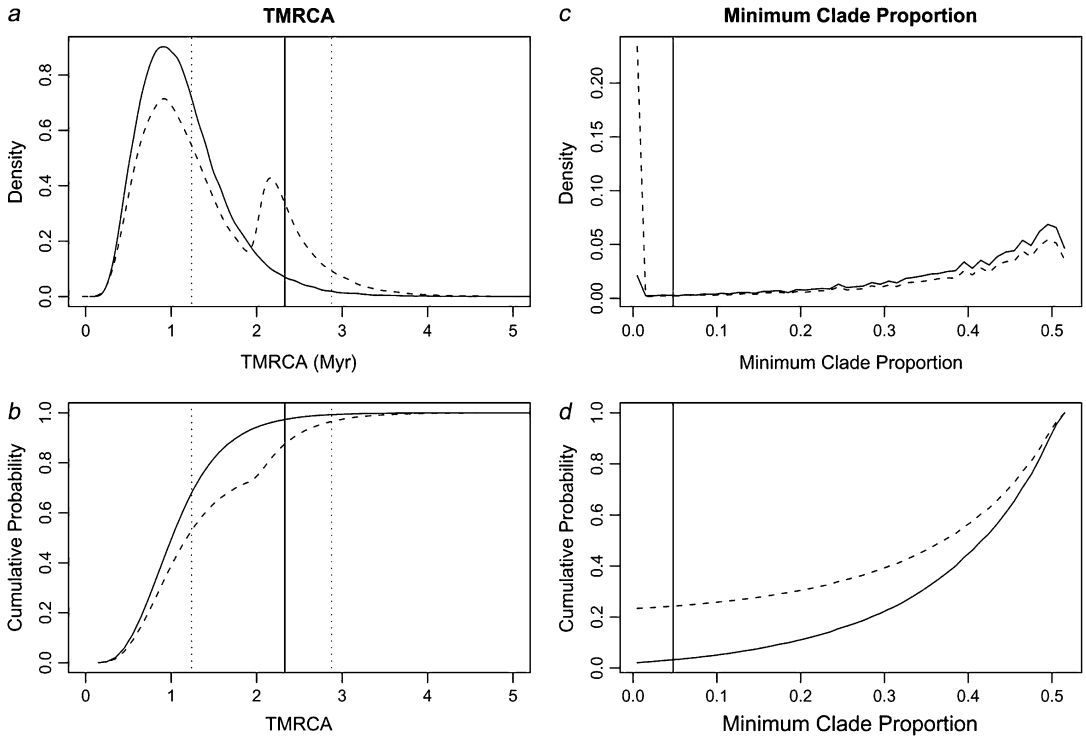

FIGURE 5.—Distribution and cumulative probability of (a and b) TMRCAs and (c and d) minimum clade proportions under the optimal two-deme RAR model (solid curve) and the corresponding model with 5% introgression from ancestral hominins at 50 KYA (dashed curve). The molecular-clock date is shown by a solid vertical line in a and b; the panmictic and island model dates are shown by dotted vertical lines (to the left and right, respectively). The admixture peak would increase as admixture occurred more frequently and shift right with deeper structure between admixing demes. The observed p_{mc} is illustrated by a solid vertical line in c and d.

further explore the effects of archaic admixture on the minimum clade proportion, we took the optimized RAR model and incorporated variable rates of admixture (0.5–5%) occurring at variable times (10–55 KY before present) and ran 10^5 simulations at each point on a 10×10 grid of parameter space (Figure 6). Basal clades dominated by non-Africans are observed more often on average as archaic admixture becomes more recent and more frequent. Higher rates of admixture make it more likely that one of the two basal clades derives entirely from the descendants of Asian ancestral hominins ($p_{mc} \rightarrow 0$), and more recent admixture makes it less likely that migrants will carry admixed lineages into African populations (also, $p_{mc} \rightarrow 0$).

DISCUSSION

GARRIGAN *et al.* (2005b) described a 2.4-kb region on the X chromosome with unusual genealogical structure in a sample of 41 humans: a deep TMRCA and a basal clade composed entirely of Asian ($n = 3$) sequences. This differs from most genealogies observed to date, where African individuals dominate at least one of the two basal clades (LABUDA *et al.* 2000; TAKAHATA *et al.* 2001; SATTA and TAKAHATA 2004). GARRIGAN *et al.* (2005b) also genotyped a single diagnostic SNP to test for the presence of clade A in a larger number of samples ($n = 570$ from 17 globally distributed populations). They discovered a decreasing frequency gradient centered on southern China (where the clade A is present >50%) and extremely low frequencies of the "Asian" divergent lineage in Europe, the Middle East, and Africa. (See Figure 1 in GARRIGAN *et al.* 2005b.) To explain the prevalence of basal *RRM2P4* lineages in East Asia they favored a model of recent admixture between divergent AMH and *H. erectus* populations; although they could not rule out founder effects leading to the loss of one of the two divergent lineages in Africa. Here, we resequence this locus in a much larger sample of humans, extend the length of the sequenced region, and test unusual aspects of the genealogy statistically, using a model-based coalescent simulation framework.

We generated resequencing data 5′ and 3′ of the central 2.4-kb region in a panel of 90 individuals to see whether the pattern originally described by GARRIGAN *et al.* (2005b) extended farther along the X chromosome. We found that a strong recombination hotspot almost completely decouples the central and 3′ regions,

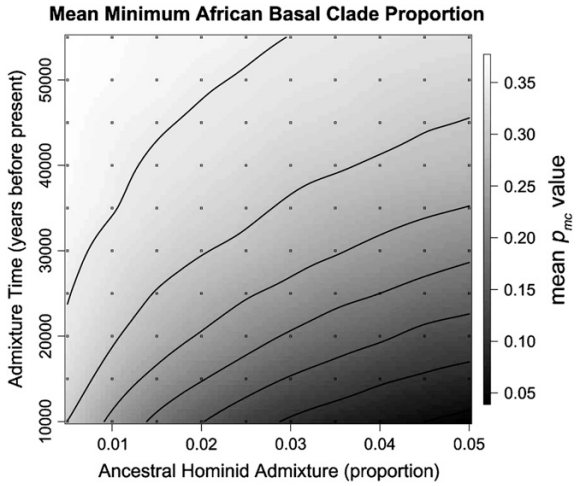

FIGURE 6.—Effect of admixture parameters on the minimum clade proportion. Fewer Africans are found in one basal clade as archaic Asian admixture occurs more frequently and more recently. The likelihood surface is generated by interpolation from mean p_{mc} values of 10^5 simulations at each point in a 10×10 grid (dots) covering the parameter space of admixture time and admixture proportion.

and despite linkage disequilibrium with the 5′ end of the sequence, the pattern of higher Asian diversity for the central region was not found for the other two regions. While it is possible that this reduced diversity may simply reflect the stochastic nature of the mutational process given the relatively short length of sequence examined (~2 kb), we suggest that the lineage history of the central region has been at least partly decoupled from those of the 5′ and 3′ regions through recombination. BLAST results show that the *RRM2P4* processed pseudogene has sequence conservation to the rhesus macaque (*Macaca mulatta*). Therefore, the high diversity of the central region does not result from increased mutation at the time of pseudogene insertion because human variation traces back only to the Homo lineage (*i.e.*, the TMRCA of human haplotypes is much more recent than the insertion time). Central region diversity is also unlikely to reflect paralogous gene conversion because the *RRM2P4* sequence has no close matches to other regions in the human genome. Finally, increased central region diversity is unlikely to reflect linkage to a gene under long-term balancing selection. The nearest 5′ gene is *SPANX-N2*, which encodes a protein of uncharacterized function. *SPANX-N2* is 580 kb, or ~1.8 cM, distant from the *RRM2P4* pseudogene, and the intervening region contains at least three hotspots of ~16 cM/Mb (each similar to the hotspot 3′ of the *RRM2P4* locus). While we cannot exclude the possibility that an unknown 5′ functional variant may be linked to the *RRM2P4* locus, the disparity in diversity between the *RRM2P4* 5′ and central regions makes linkage with a third selected locus located farther 5′ unlikely. In sum, it appears that polymorphisms that clearly define the unusual *RRM2P4* genealogy are found only within the stretch of noncoding DNA associated with the pseudogene (Figure 1) and that these sites are unlikely to be affected by recent positive selection.

Our sixfold larger database of central region sequences does not result in a substantial change in the topology of the gene tree described by GARRIGAN *et al.* (2005b) or in the partitioning of its two deepest branches. However, we did observe several new lineages at low frequency and discovered some novel evidence of recombination. Of the 21 individuals identified with the less frequent basal lineage, only one was from Africa. This Dogon individual from Mali was the same African individual identified as carrying a clade A lineage on the basis of the SNP-based genotyping assay in GARRIGAN *et al.* (2005b). Our extended sequence database identified only a single new polymorphism in clade A and 11 new polymorphisms in the more diverse globally distributed clade B (compare Figure 1 of GARRIGAN *et al.* 2005b with our Figure 4). This brings the total number of haplotypes in the divergent Asian clade A to 3, compared with a total of 20 haplotypes in the other basal clade. Interestingly, the new clade A haplotype was identified only in two individuals: the aforementioned Dogon individual and an individual from Bougainville Island, Melanesia. This is consistent with the possibility that recent migration carried this rare lineage between continents. Although we cannot exclude the possibility that unsampled African populations carry this clade at higher frequency than we observe here, our geographical coverage of the African continent is still quite extensive.

Our simulations show that the 2.33-million-year molecular-clock-based TMRCA estimate is a statistical outlier under the RAR model (Figure 5, a and b), but not under an AA model with ancestral structure dating from 2 MYA. However, maximum-likelihood estimates of the TMRCA inferred under different models of human population structure span a wide range of times. The large variance in TMRCA estimates is also expected because the short length of the central subregion sequence limits the number of sites segregating between the two basal clades. Unfortunately, the variance of TMRCA is always large relative to the mean and does not decrease appreciably with increased sample size (GRIFFITHS and TAVARÉ 1994; TANG *et al.* 2002; BASU and MAJUMDAR 2003). Indeed, as we increased the sample size from 41 (GARRIGAN *et al.* 2005b) to 253, we found no new polymorphic sites on the basal branches of the gene tree, which suggests that we have sampled sufficiently to observe the basal node of the genealogy for the entire global population. There is only a remote probability that we have not observed the deepest split of the *RRM2P4* tree in our data set of 253 individuals ($P \approx 0.0079$) (SAUNDERS *et al.* 1984; KLIMAN and HEY 1993). Our TMRCA estimates, while consistent with the AA

model specified here, are unlikely to be improved by sampling additional individuals.

We also considered the observation of Garrigan *et al.* (2005b) that Asian samples are overrepresented in one of the two basal clades of the *RRM2P4* tree. This pattern continued to hold even after increasing the size of our DNA sequence data set. To assess how unusual this aspect of the *RRM2P4* genealogy is under alternative models of human evolutionary history, we defined a new summary statistic, p_{mc}, which quantifies the skew in the proportion of individuals from two populations among the two basal clades of a gene tree. This summary statistic is applicable to *RRM2P4* because the central region is essentially tree-like. We observed $p_{mc} = 0.048$ for the central region of *RRM2P4*, which is significantly unlikely under the RAR model ($P = 0.031$, Figure 5, c and d). To examine the sensitivity of the p_{mc} statistic under a range of archaic admixture parameters, we simulated coalescent genealogies and varied both the admixture proportion and the timing of introgression (Figure 6). Although genealogies with small p_{mc} values are more common as the admixture proportion increases (*i.e.*, up to ~5%) and introgression begins more recently (*i.e.*, as recently as ~10 KYA), the *RRM2P4* genealogy is not a significant outlier under any of these AA model parameterizations. This further supports our conclusion that the *RRM2P4* genealogy fits expectations better under a scenario of archaic admixture.

While there are limitations with both TMRCA and p_{mc} for distinguishing predictions of the RAR and AA models, they do represent independent summaries of the data and, thus, complement one another. As already mentioned, the power of these two test statistics depends on different aspects of sampling. Variance in the estimate of the TMRCA is improved by longer sequences of the region with tree-like ancestry (in the case of *RRM2P4* this is limited by the small central region and flanking recombination), but only slightly by increasing the sample size. On the other hand, estimates of p_{mc} can be improved by increasing the sample size, because the variance of p_{mc} decreases approximately as the inverse of the sample size (analyses not shown). If the p_{mc} is genuinely an outlier under the RAR model, increasing the number of individuals sampled increases the power to reject RAR. Indeed, when we use the SNP data of Garrigan *et al.* (2005b), which included 177 Africans and 393 non-Africans, we reject the RAR model with greater confidence ($p_{mc} = 0.0189$, $P = 0.014$).

Further evidence in support of an archaic admixture model awaits analysis of additional loci exhibiting genealogical properties similar to the central *RRM2P4* region. Several candidates have already been identified (Harding *et al.* 1997; Ziętkiewicz *et al.* 2003; Stefansson *et al.* 2005; Shimada *et al.* 2007), but most lack rigorous statistical analyses under a range of demographic models, including ancient admixture alternatives. The frequency at which we expect to find introgressed regions depends largely on the amount of admixture between the two archaic populations (Wall 2000). Moreover, unless admixture was recent and involved highly divergent populations, the power to detect archaic admixture is low (Nordborg 2000). In the case of *RRM2P4*, divergence may have started at the time of separation of *H. ergaster*/*H. erectus* populations in Africa ~2 MYA (Anton and Swisher 2004). Yet, the length of the divergent sequence is short, possibly as a result of the nearby recombination hotspot or because admixture occurred in the more distant past and recombination has subsequently broken down the admixed chromosome. In any case, identifying longer sequences with greater divergence would allow for more sophisticated tests of archaic admixture. For example, Wall (2000) suggested a number of summary statistics that are based on both the level of divergence between two clades and the amount of recombination between them.

Recent population structure is another factor that may affect the probability of sampling a locus with a genealogy showing signs of archaic admixture. As pointed out by Nordborg (2000), population structure may actually increase the power to detect archaic admixture if we sample sufficiently among demes, because we would expect the introgressed alleles to still be present in the area of the world where admixture took place. In the case of *RRM2P4*, individuals carrying the less frequent divergent lineage are concentrated in East Asia, suggesting that admixture may have occurred at the Asian end of the global distribution of human populations. On the other hand, it is important to point out that current population structure is unlikely to reflect ancient patterns directly. Following a demic expansion, what was once subdivision between two African populations may now appear as structure between African and non-African populations. For loci with more ancient TMRCA, there is an increase in power to detect archaic admixture even if it occurred at more ancient times (Nordborg 2000). This means that for *RRM2P4*, which has an ancient TMRCA, we cannot be confident about where archaic admixture may have occurred geographically. In this regard, it is interesting to note that a growing number of loci have been discovered with two deeply divergent lineages where both the major and the minor types are present only in African populations (Barreiro *et al.* 2005; Garrigan *et al.* 2005a; Hayakawa *et al.* 2006). This supports models in which anatomically modern humans descend from a structured ancestral African population (Garrigan and Hammer 2006). We find some support that elevated admixture among highly divergent African subpopulations just prior to the recent African expansion could explain the pattern of polymorphism at *RRM2P4* (supplemental Figure 4 at http://www.genetics.org/supplemental/), but note that this model has little power to explain why *RRM2P4* clade A lineages are geographically restricted to East Asia today. For now, this locus represents

a genealogical history that is most consistent with recent admixture from an archaic hominin population in Asia.

We thank Zahra Mobasher (University of Arizona) for excellent technical assistance and David Morales (University of Arizona) for helpful discussion. This research forms part of the HOMINID project, a genomic resequence study funded by National Science Foundation grant BCS-0423670.

LITERATURE CITED

Anton, S. C., and C. C. Swisher, 2004 Early dispersal of Homo from Africa. Annu. Rev. Anthropol. **33:** 271–296.

Barreiro, L. B., E. Patin, O. Neyrolles, H. M. Cann, B. Gicquel *et al.*, 2005 The heritage of pathogen pressures and ancient demography in the human innate-immunity CD209/CD209L region. Am. J. Hum. Genet. **77:** 869–886.

Basu, A., and P. Majumdar, 2003 A comparison of two popular statistical methods for estimating the time to most recent common ancestor (TMRCA) from a sample of DNA sequences. J. Genet. **82:** 7–12.

Beaumont, M. A., W. Zhang and D. J. Balding, 2002 Approximate Bayesian computation in population genetics. Genetics **162:** 2025–2035.

Cann, H. M., C. de Toma, L. Cazes, M. F. Legrand, V. Morel *et al.*, 2002 A human genome diversity cell line panel. Science **296:** 261–262.

Eswaran, V., 2002 A diffusion wave out of Africa: the mechanism of the modern human revolution? Curr. Anthropol. **43:** 749–774.

Evans, P. D., N. Mekel-Bobrov, E. J. Vallender, R. R. Hudson and B. T. Lahn, 2006 Evidence that the adaptive allele of the brain size gene *microcephalin* introgressed into *Homo sapiens* from an archaic *Homo* lineage. Proc. Natl. Acad. Sci. USA **103:** 18178–18183.

Fenner, J. N., 2005 Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am. J. Phys. Anthropol. **128:** 415–423.

Garrigan, D., and M. F. Hammer, 2006 Reconstructing human origins in the genomic era. Nat. Rev. Genet. **7:** 669–680.

Garrigan, D., Z. Mobasher, S. B. Kingan, J. A. Wilder and M. F. Hammer, 2005a Deep haplotype divergence and long-range linkage disequilibrium at Xp21.1 provide evidence that humans descend from a structured ancestral population. Genetics **170:** 1849–1856.

Garrigan, D., Z. Mobasher, T. Severson, J. A. Wilder and M. F. Hammer, 2005b Evidence for archaic Asian ancestry on the human X chromosome. Mol. Biol. Evol. **22:** 189–192.

Griffiths, R. C., 2007 Genetree v. 9.0. http://www.stats.ox.ac.uk/~griff/software.html.

Griffiths, R. C., and S. Tavaré, 1994 Simulating probability distributions in the coalescent. Theor. Popul. Biol. **46:** 131–159.

Harding, R. M., S. M. Fullerton, R. C. Griffiths, J. Bond, M. J. Cox *et al.*, 1997 Archaic African *and* Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. **60:** 772–789.

Hayakawa, T., I. Aki, A. Varki, Y. Satta and N. Takahata, 2006 Fixation of the human-specific CMP-*N*-acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution. Genetics **172:** 1139–1146.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18:** 337–338.

Karafet, T. M., L. P. Osipova, M. A. Gubina, O. L. Posukh, S. L. Zegura *et al.*, 2002 High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. Hum. Biol. **74:** 761–789.

Kingman, J. F. C. (Editor), 1982 *On the Genealogy of Large Populations.* Applied Probability Trust, Sheffield, UK.

Kliman, R. M., and J. Hey, 1993 DNA sequence variation at the *period* locus within and among species at the *Drosophila melangaster* complex. Genetics **133:** 375–387.

Labuda, D., E. Ziętkiewicz and V. Yotova, 2000 Archaic lineages in the history of modern humans. Genetics **156:** 799–808.

McBrearty, S., and A. S. Brooks, 2000 The revolution that wasn't: a new interpretation of the origin of modern human behavior. J. Hum. Evol. **39:** 453–563.

McDougall, I., F. H. Brown and J. G. Fleagle, 2005 Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature **433:** 733–736.

McVean, G., and C. C. Spencer, 2006 Scanning the human genome for signals of selection. Curr. Opin. Genet. Dev. **16:** 624–629.

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics **146:** 1501–1514.

Nordborg, M., 2000 On detecting ancient admixture, pp. 123–136 in *Genes, Fossils and Behaviour: An Integrated Approach to Human Evolution,* edited by P. Donnelly. IOS Press, Amsterdam.

Plagnol, V., and J. D. Wall, 2006 Possible ancestral structure in human populations. PLoS Genet. **2:** e105.

Satta, Y., and N. Takahata, 2004 The distribution of the ancestral haplotype in finite stepping-stone models with population expansion. Mol. Ecol. **13:** 877–886.

Saunders, I. W., S. Tavaré and G. A. Watterson, 1984 On the genealogy of nested subsamples from a haploid population. Adv. Appl. Probab. **16:** 471–491.

Shimada, M. K., K. Panchapakesan, S. Tishkoff, A. Q. Nato, Jr. and J. Hey, 2007 Divergent haplotypes and human history as revealed in a worldwide survey of X-linked DNA sequence variation. Mol. Biol. Evol. **24:** 687–698.

Stefansson, H., A. Helgason, G. Thorleifsson, V. Steinthorsdottir, G. Masson *et al.*, 2005 A common inversion under selection in Europeans. Nat. Genet. **37:** 129–137.

Takahata, N., S. H. Lee and Y. Satta, 2001 Testing multiregionality of modern human origins. Mol. Biol. Evol. **18:** 172–183.

Tang, H., D. O. Siegmund, P. Shen, P. J. Oefner and M. W. Feldman, 2002 Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. Genetics **161:** 447–459.

Templeton, A., 2002 Out of Africa again and again. Nature **416:** 45–51.

Wall, J. D., 2000 Detecting ancient admixture in humans using sequence polymorphism data. Genetics **154:** 1271–1279.

Wall, J. D., and M. F. Hammer, 2006 Archaic admixture in the human genome. Curr. Opin. Genet. Dev. **16:** 606–610.

Y Chromosome Consortium, 2002 A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res. **12:** 339–348.

Ziętkiewicz, E., V. Yotova, D. Gehl, T. Wambach, I. Arrieta *et al.*, 2003 Haplotypes in the Dystrophin DNA segment point to a mosaic origin of modern human diversity. Am. J. Hum. Genet. **73:** 994–1015.

Communicating editor: N. Takahata

APPENDIX

Here, we define a new summary statistic, the minimum clade proportion (p_{mc}). In its simplest form, this statistic characterizes the proportion of individuals from a specified group (*e.g.*, Africans) in each of the two basal clades of a binary gene tree. The minimum clade proportion can thus be defined as

$$p_{\mathrm{mc}} = \min\left(\frac{k_1}{n_1}, \frac{k_2}{n_2}\right), \tag{A1}$$

where $n_{1,2}$ and $k_{1,2}$ are, respectively, the total number of individuals and the number of chromosome copies from the specified group in basal clades 1 and 2. Clades 1 and 2 are defined arbitrarily because, *a priori*, branch rotation has no effect on the topology of a gene tree. We also note that the p_{mc} statistic is not itself a proportion, although the quantity (k_i/n_i) is. Strictly, $p_{\mathrm{mc}} \in [0, 1)$; *i.e.*, the minimum clade proportion does not include unity.

The variance of p_{mc} decreases approximately as the inverse of the sample size; *i.e.*, increasing the number of sampled individuals increases the power to reject the null model (our unpublished simulation results). Furthermore, the power of this test depends on sampling sizes as follows. Consider that we have N_1 individuals from the specified group, $N_1 + N_2$ total individuals, and an observed $p_{\mathrm{mc}} = \lambda$. Also, given an arbitrarily chosen clade, p_1 is that clade's frequency in the specific group, whereas p_2 is the frequency of that clade in the remaining individuals. For a given hypothesis H

$$P(p_{\mathrm{mc}} = \lambda \mid H, N_1, N_2) = \int_{p_1, p_2} P(p_{\mathrm{mc}} = \lambda \mid N_1, N_2, p_1, p_2) \cdot f(p_1, p_2 \mid H)\, dp_1\, dp_2, \tag{A2}$$

where $f(p_1, p_2 \mid H)$ is the probability density for the population frequencies given the hypothesis. Note that the first factor of the integrand in Equation A2 depends only on allele frequencies and sample sizes and not on the hypothesis, whereas the second factor depends only on the hypothesis. The conditioned probability distribution of p_{mc} therefore resembles the ratio of two nonindependent binomial distributions. Given p_1, p_2, N_1, and N_2 and taking $A \sim \mathrm{Binom}(p_1, N_1)$ and $B \sim \mathrm{Binom}\ (p_2, N_2)$, it follows that

$$p_{\mathrm{mc}} = \min\left(\frac{A}{A + B}, \frac{N_1 - A}{N_1 + N_2 - A - B}\right). \tag{A3}$$

If the sample size is sufficiently large and the frequencies p_1 and p_2 are not identical, the probability distribution of p_{mc} converges to the probability distribution of the ratio of random variables that has the smallest mean. As N_1 and N_2 increase, the values of A and B approach $N_1 p_1$ and $N_2 p_2$, respectively, whereas their standard errors grow on the order of $\sqrt{N_1}$ and $\sqrt{N_2}$. Consequently, it is perhaps not surprising that the variance of p_{mc} decreases at least as the inverse of the smallest sample size. In other words, for cases where population frequencies are unlikely under the null hypothesis, raising the sample size can increase the power to reject the model (*cf.* TMRCA).

We emphasize that the likelihood of an observed p_{mc} depends strongly on the demographic model underlying the null hypothesis. In practice, the probability that $p_{\mathrm{mc}} \leq \lambda$ must be determined by coalescent simulation.

Finally, we note that the p_{mc} statistic can be generalized to any subset of chromosome copies, K, and any number of clades, C (indexed by i), from a data set, N. Necessary conditions include $K \subset N$, $C_i \subset N$, and $0 < |K| \leq |N|$, where $|K|$ and $|N|$ are the cardinals of K and N, respectively. It follows that the number of chromosome copies from the specified group in clade C_i is $k_i = |(K \cap C_i)|$, and the total number of individuals in clade C_i is simply $n_i = |C_i|$. Consequently, the minimum clade proportion can be defined more generally for m clades as

$$p_{\mathrm{mc}} = \min\left(\frac{k_1}{n_1}, \ldots, \frac{k_m}{n_m}\right). \tag{A4}$$

Only coalescent simulations for which the relationship on the gene tree of the clades, m, is identical to that of the data should be used to determine the probability of this generalized p_{mc} statistic.

11

MALE DOMINANCE RARELY SKEWS THE FREQUENCY DISTRIBUTION OF Y CHROMOSOME HAPLOTYPES IN HUMAN POPULATIONS

Lansing, J.S., J.C. Watkins, B. Hallmark, **M.P. Cox**, T.M. Karafet, H. Sudoyo, and M.F. Hammer. 2008. Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations. *Proceedings of the National Academy of Sciences USA* 105:11645-11650.
http://doi.org/10.1073/pnas.0710158105

ABSTRACT A central tenet of evolutionary social science holds that behaviors, such as those associated with social dominance, produce fitness effects that are subject to cultural selection. However, evidence for such selection is inconclusive because it is based on short-term statistical associations between behavior and fertility. Here, we show that the evolutionary effects of dominance at the population level can be detected using noncoding regions of DNA. Highly variable polymorphisms on the nonrecombining portion of the Y chromosome can be used to trace lines of descent from a common male ancestor. Thus, it is possible to test for the persistence of differential fertility among patrilines. We examine haplotype distributions defined by 12 short tandem repeats in a sample of 1269 men from 41 Indonesian communities and test for departures from neutral mutation-drift equilibrium based on the Ewens sampling formula. Our tests reject the neutral model in only 5 communities. Analysis and simulations show that we have sufficient power to detect such departures under varying demographic conditions, including founder effects, bottlenecks, and migration, and at varying levels of social dominance. We conclude that patrilines seldom are dominant for more than a few generations, and thus traits or behaviors that are strictly paternally inherited are unlikely to be under strong cultural selection.

SUPPLEMENTARY MATERIALS
http://www.pnas.org/content/105/33/11645?tab=ds

Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations

J. Stephen Lansing*†, Joseph C. Watkins‡§, Brian Hallmark*, Murray P. Cox†¶, Tatiana M. Karafet¶, Herawati Sudoyo‖, and Michael F. Hammer¶

Departments of *Anthropology and ‡Mathematics, and ¶Division of Biotechnology, Biological Sciences West, University of Arizona, Tucson, AZ 85721; †Santa Fe Institute, Santa Fe, NM 87521; and ‖The Eijkman Institute for Molecular Biology, Diponegoro 69, Jakarta 10430, Indonesia

Edited by Simon A. Levin, Princeton University, Princeton, NJ, and approved June 3, 2008 (received for review October 25, 2007)

A central tenet of evolutionary social science holds that behaviors, such as those associated with social dominance, produce fitness effects that are subject to cultural selection. However, evidence for such selection is inconclusive because it is based on short-term statistical associations between behavior and fertility. Here, we show that the evolutionary effects of dominance at the population level can be detected using noncoding regions of DNA. Highly variable polymorphisms on the nonrecombining portion of the Y chromosome can be used to trace lines of descent from a common male ancestor. Thus, it is possible to test for the persistence of differential fertility among patrilines. We examine haplotype distributions defined by 12 short tandem repeats in a sample of 1269 men from 41 Indonesian communities and test for departures from neutral mutation-drift equilibrium based on the Ewens sampling formula. Our tests reject the neutral model in only 5 communities. Analysis and simulations show that we have sufficient power to detect such departures under varying demographic conditions, including founder effects, bottlenecks, and migration, and at varying levels of social dominance. We conclude that patrilines seldom are dominant for more than a few generations, and thus traits or behaviors that are strictly paternally inherited are unlikely to be under strong cultural selection.

cultural evolution | Indonesia | neutral theory | selection | Red Queen

Evolutionary social scientists analyze the fitness consequences of behavior, where the currency of fitness is reproductive success. Many studies have argued that reproductive skew biased toward dominant or high-ranking men is very common in human communities: "In more than one hundred well studied societies, clear formal reproductive rewards for men are associated with status: high-ranking men have the right to more wives" (1). Demographic statistics collected over short time scales support these claims (2). Although variation in male fitness is known to occur, an important unanswered question is whether such differences are heritable and persist long enough to have evolutionary consequences at the population level. In this report we show that it is possible to use genetic markers to test for the signature of heritable reproductive skew within population groups.

Several authors have modeled the effects of reproductive skew in different contexts (3, 4). Fig. 1 contrasts two models: the standard Wright-Fisher model in which all individuals produce offspring with equal probability (i.e., no reproductive skew) and a model of male dominance in which individuals whose parents had many offspring are more likely to have more children themselves. Investigations of genetic variation on the nonrecombining portion of the Y chromosome provide an opportunity to ask which of these models better characterizes a given population. Here, we genotype 12 short tandem repeats (Y-STRs) and a battery of SNPs on the Y chromosomes of 1269 men from 41 Indonesian communities. We then construct Y-chromosome haplotypes for each sample [supporting information (SI) Table S1]. Unlike human genetic studies that sample unrelated individuals from a broad geographical watershed and thus fail to capture community-level processes, we sample male residents of several indigenous and relatively isolated communities. These populations include neolocal Borneo hunter-gatherers and Central Javanese rice farmers, matrilocal horticulturalists on Flores, patrilocal Balinese wet-rice farmers, and patrilocal horticultural clans on Sumba, Nias, and Flores.

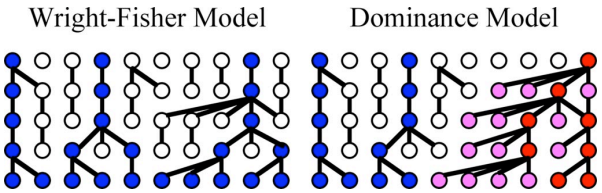

Fig. 1. Population models. Colored dots represent individuals who have descendants in the final generation. Red dots represent dominant individuals who are more likely to reproduce. Pink dots represent nondominant individuals having a dominant ancestor. Blue dots represent nondominant individuals having no dominant ancestor. Dominant individuals in 1 generation are chosen at random from the offspring of dominant individuals in the previous generation. (See *SI Text Population Models* for more details.)

To detect evidence of heritable reproductive skew, we use statistical procedures that are based on the Ewens sampling formula (5). In particular, we apply Slatkin's exact test of neutrality, which is based on the haplotype frequency distribution (6, 7). Communities experiencing reproductive skew among patrilines will tend to have haplotype frequency distributions that, over time, become unlikely under the neutral model, skewed toward an excess of common haplotypes (patrilines). Such skewed distributions would provide evidence for cultural selection, defined as the heritable nongenetic transmission of any kind of behavior that affects reproductive success (see ref. 8).

Results

Strikingly, we find that only 5 of 41 communities (12%) have haplotype frequency distributions that are unlikely under the neutral model (i.e., rejected the null hypothesis of neutrality). Fig. 2 shows the haplotype frequency distributions of 1 of these 5 communities (Fig. 2*B*), and a neutral deme (Fig. 2*A*). All 5

ANTHROPOLOGY

Author contributions: J.S.L. designed research; J.S.L., J.C.W., B.H., M.P.C., T.M.K., and H.S. performed research; J.S.L., J.C.W., B.H., M.P.C., T.M.K., H.S., and M.F.H. analyzed data; and J.S.L., J.C.W., B.H., M.P.C., T.M.K., and M.F.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

§To whom correspondence should be addressed. E-mail: jwatkins@math.arizona.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0710158105/DCSupplemental.

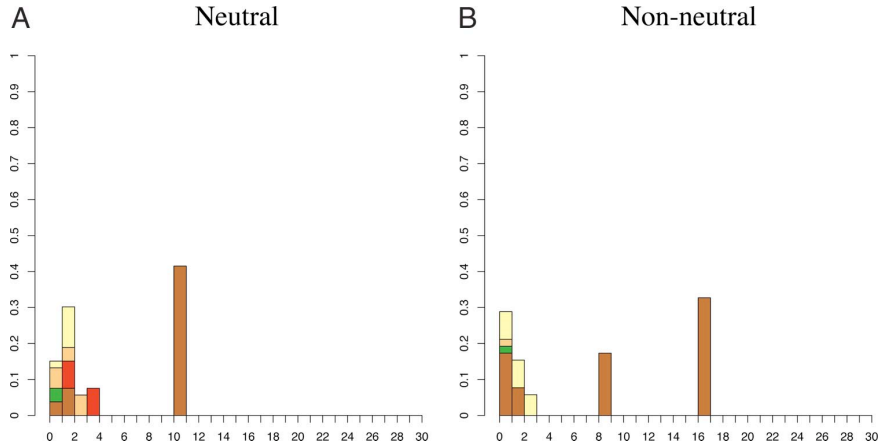

Fig. 2. Example haplotype distributions. The height of column *i* is the number of individuals whose haplotype is represented *i* times in the sample (normalized by sample size). (*A*) An example of a community haplotype distribution that does not show evidence of departure from neutrality based on the exact test: Sumba/Mamboro. (*B*) An example of a community showing departure from neutrality based on the exact test: Sumba/Wanokaka. Although nonneutral cases often appear to have mass shifted to the right, it generally is not possible to determine the outcome of the exact test by inspection. Colors indicate Y chromosome haplogroup membership (C, F, K, M, O, or S). All 41 distributions and the color key are given in Fig. S1. Additional haplogroup information is in Table S5.

cases that rejected neutrality show a similar pattern: 1 or more common haplotype(s) (i.e., a high-frequency patriline) paired with a large number of low-frequency haplotypes (Fig. S1). This contrasts with patterns produced under the neutral model with different population sizes. Observed haplotype frequency distributions typical of large communities (i.e., in which the genetic diversity of the sampled population is high) are characterized by many unique haplotypes and a few high-frequency lineages. Smaller communities with reduced neutral diversity tend to have fewer low-frequency haplotypes and fewer common patrilines.

Dominance Model and Power of the Exact Test. Given these findings, we wanted to know how often the Ewens exact test fails to reject neutrality when male dominance is in fact occurring. The Ewens test is nonparametric and is not based on any specific model. It is independent of both village effective population size and Y-STR mutation rate. We constructed an analytical model to determine the ability of this test to detect departures from neutral equilibrium as a result of heritable reproductive skew caused by social dominance. In this model, dominant males comprise a fraction δ of the total population and have a selective advantage σ over the nondominant males. Offspring in the next generation are chosen according to a haploid Wright-Fisher model with competition between the 2 classes, dominant and nondominant. To maintain the fraction of dominant males at δ for each subsequent generation, a proportion of sons of the dominant males is chosen at random to be dominant for the next iteration of the model. Thus, for this 2-parameter model, either $\delta = 0$ or $\sigma = 0$ yields neutral populations (see *SI Text Population Models* for additional details).

To calculate the power of the Ewens exact test under this model, we simulate nonneutral populations, sample them, and apply the exact test to observe departures from the expectation under neutrality. To condition the simulations on our Indonesian data, we simulate populations taking (*i*) samples of size 20 from smaller villages having effective population size 100 using 10 Y-STRs, and (*ii*) samples of size 35 from larger villages having effective population size 300 using 12 Y-STRs. Tables S2 and S3 show the results of these analyses for all parameters. For example, for a village of size 300 with only 4% of the men producing twice as many offspring ($\sigma = 1$), the power of our test is 0.39. Assuming this level of dominance, the probability that 5 (or fewer) of 41 villages rejects neutrality is ≈0.02%. Thus, even weak dominance would be detectable in a sample of 41 villages. Fig. 3 gives this probability for the power calculated from any model of selection.

Demographic Effects. Departures from neutrality can arise from both selective and demographic causes (e.g., changes in population size and structure). Consequently, inferring whether observed cases result from reproductive skew or from other factors is not straightforward. There are 2 possibilities. First, demographic processes might mask the effects of male dominance and make villages appear neutral when in fact they are experiencing heritable reproductive skew, leading to false negatives. Alternatively demographic processes might cause neutral populations to appear nonneutral (i.e., leading to false positives). In this section we discuss the effects of 2 classes of demographic process on the Ewens test: reductions in population size (including founder effects and bottlenecks) and migration.

In general, bottlenecks (or founder events) reduce genetic

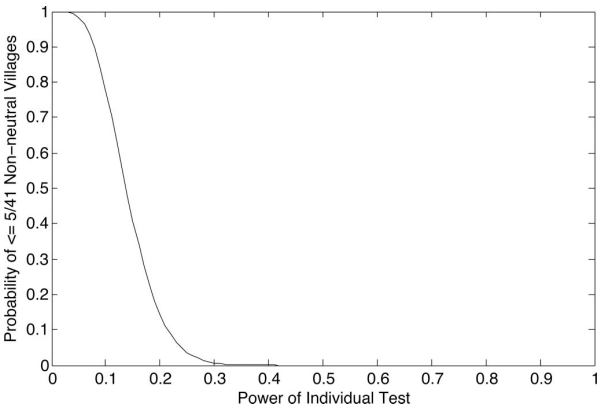

Fig. 3. Probability of seeing 5 or fewer (of 41) nonneutral villages for a given power of the individual tests to detect departure from neutrality. If the power of the individual tests is 0.2, the probability of observing ≤5 is ≈0.2. If the power of the individual tests is > 0.3, then this probability is vanishingly small.

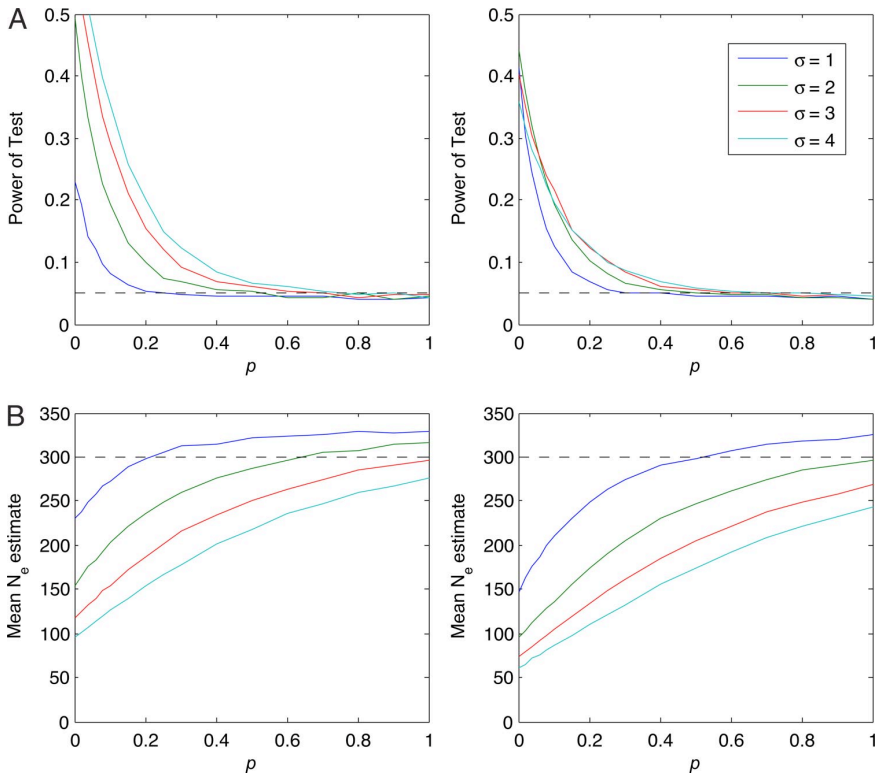

Fig. 4. Effects of adding nonheritable reproductive skew to a simulated population of size $n = 300$. The first column shows simulation results for $\delta = 0.02$, the second column for $\delta = 0.06$. (*A*) Reduction of power. As p increases, the power to detect skew is reduced. However, the length of time lineages are dominant also decreases, nullifying any long-term evolutionary advantage of dominance. (*B*) N_e reduction. Effective population size remains depressed for all values of p, providing an additional way to assess the strength of both heritable and nonheritable skew. Estimates eventually exceed the actual population size because the estimator is biased upwards.

variation (Fig. S2) by leading to the rapid loss of rare lineages. Eventually, if the bottleneck remains strong, more common lineages will be lost also. In the phase of population recovery after a bottleneck, new mutations tend to occur on different Y chromosomes. Thus, during the bottleneck phase we expect to find a reduction in the frequency of rare haplotypes, whereas in the recovery phase we expect to see an excess of rare haplotypes compared with the expectation for a population in neutral equilibrium. For false positives, the most extreme case is a neutral population undergoing a bottleneck in which all individuals possess the same haplotype. This is conceivable, for example, in the case of a founding event by a small kin group. Thus, neutral populations may test as nonneutral in the period immediately following a strong bottleneck (Fig. S3*A*) (see *Materials and Methods*).

In a population experiencing reproductive skew as a result of male dominance, a bottleneck will reduce genetic variation further. Because a reduction in haplotype diversity reduces the power of the exact test, dominance will be more difficult to detect during the recovery phase. Simulations indicate that populations return to the nonneutral condition in ~20 generations as long as dominance remains constant (Fig. S3*B*). This raises the question of whether we are underestimating the importance of male dominance as a result of low power to reject neutrality after a bottleneck. We believe this to be unlikely for several reasons. First, the mean village size, at 280 male household heads, is sufficiently large to indicate that frequent bottlenecks occurring in the past few generations are very unlikely (Table S4). Second, haplogroup diversity (i.e., the number and relative abundance of Y-chromosome lineages as defined by SNPs) is relatively high (Table S5). Finally, reductions in effective population size that one expects to accompany a bottleneck are not seen. The bottlenecks simulated here are quite extreme, with reductions in population size of 90% for 10 generations. Moreover, if bottlenecks do occur, they also independently will reduce the demographic effects of male dominance. For example, if a population is reduced from size 300 to 35 for 10 generations, then the chance that any individual has a descendant in a recovering population is $< 2\%$ (this becomes $< 1\%$ if the bottleneck lasts for 20 generations) (Fig. S2). Thus, it is unlikely that Y-chromosome lineages of dominant males would survive a severe population bottleneck.

Migration is another factor that affects patterns of diversity within populations. Increasing the number of new alleles entering the population each generation by migration is functionally equivalent to increasing the mutation rate, μ, in a Wright-Fisher model. Consequently, migration will not cause false positives because the exact test is not dependent on the mutation rate under this model. However, for villages experiencing reproductive skew, high migration reduces the power of the Ewens exact test. In our sample, the highest migration rates are expected in Balinese villages because they often are adjoining and speak the same language. We simulated an Isolation with Migration model to estimate migration rates (see *SI Text Estimation of Balinese Migration Rates*). To determine whether this could impact the power of our tests significantly, we simulated moderate and high migration ($m = 0.01$, 0.03/gen) as part of our power analyses (Tables S2 and S3). Even when sample sizes are small and

dominance is weak, the reduction in power with these levels of migration is minimal, amounting to only a few percent. For example, with 4% dominant males producing 3 times as many children as nondominant males, high migration reduces the power from 32% to 26% in small village simulations (Table S2).

Nonheritable Reproductive Skew. We also consider another model of reproductive skew in which the composition of the dominant group varies. The distinction between the heritable and nonheritable cases is important: any basis for cultural selection must involve some heritable behavior or trait that confers a reproductive advantage. In the context of social dominance, this distinguishes between situations in which, over multiple generations, males persistently pass dominance to their sons versus situations in which the membership of reproductively dominant groups is transient. To consider this case, we add a parameter p to the model of dominance. This parameter gives the fraction of the dominant class that enters from the nondominant class in a given generation. Consequently, for example, if $p = 0.05$, then a dominant lineage will span, on average, $20 = 1/0.05$ generations. If $p = 1$, then reproductive skew is uncorrelated between generations. In that limiting case, dominant individuals produce significantly more offspring than nondominants, but this tendency is attained randomly in each generation. Values of p significantly >0 could result in a "Red Queen" effect, in which even acute competition between patrilines does not translate into lasting dominance (8). The power analysis for this model shows a small reduction in power for the case $p = 0.05$. For $p = 0.15$, dominant lines persist on average only 6 or 7 generations, and, not surprisingly, we find a clear reduction in power for this case. The power of selection to structure the genetic composition of the community also is reduced as p increases and the composition of the dominant group becomes more fluid (Fig. 4A). Concomitant with this result, we also find in our simulations a clear reduction in the village effective population size, N_e, for small values of p and some reduction for all values (Fig. 4*B*). To look for reductions in N_e, we estimated the effective population sizes of sampled communities based on the number of haplotypes and compared these values with best-estimate demographic data made available by village records or the local Indonesian government. For the communities in which the neutrality test was not rejected, effective size estimates (compensating for migration) are similar to census demographic values (Table S4). In contrast, the nonneutral communities have N_e estimates that are much smaller than the number of male heads of households, in some cases dramatically smaller. Because we do not see a marked reduction in N_e for the villages that test neutral, and these villages maintain high levels of haplogroup diversity (Table S5), it appears that values of $p \gg 0$ are uncommon for the communities in our sample. This argues against significant reproductive skew, which would reduce both N_e and diversity quickly in small communities.

Cases That Depart from Neutral Expectations. Finally, we investigate possible explanations for departures from neutrality for each of the 5 nonneutral communities. Three communities come from Sumba, a remote island in which residence is patrilocal, descent is traced through the patriline, marriage is polygamous, and competition for status and resources among clans is endemic (9). Given these circumstances, it perhaps is remarkable that the remaining 5 Sumbanese communities we sampled failed to show statistically significant evidence of heritable reproductive skew. In Bali, where residence is patrilocal and there is competition among patrilineal descent groups, the only nonneutral community is South Batur. This village was part of the larger village of Batur, which fragmented in 1948 after a period of rivalry between factions that often were associated with descent groups. A very recent process of village fissioning led to resettlement of some households based on lineage affiliation (10). The remaining nonneutral community is a patrilineal and patrilocal community in central Flores. It once was the site of a minor princedom that became an administrative center during the Dutch colonial era and now serves as a district capital ("kecamatan"). Our sample includes a large proportion of civil servants born elsewhere, probably accounting for its diverse haplotype distribution.

In sum, most of the sampled populations do not show evidence of a departure from neutral stochastic equilibrium with respect to male lineages. If reproductive skew inherited between generations were a pervasive and ongoing process, we would expect to observe frequent rejections of the Ewens exact test of neutrality. We do not observe such rejection in 88% of our Indonesian communities. Even our nonneutral Indonesian communities may not necessarily reflect the action of male reproductive skew but rather the signature of a very recent or nonneutral founding group. For the reasons discussed earlier, recent bottlenecks are unlikely to be masking significant skew. We conclude that male reproductive skew is at best weak in most of our sampled population groups, despite their varied subsistence strategies and kinship practices.

Discussion

The implications of our results extend beyond the Indonesian cases described here to the broader question of the relationship between reproductive skew and the genetic structure of human communities. The genetic or cultural-evolutionary effects of dominance become apparent only by extending analyses from the inclusive fitness of individuals to population-level consequences over the course of multiple generations. Here, we take a population-genetic approach to infer selection by detecting departures from neutral drift-mutation equilibrium at the population level rather than from variation in the reproductive success of individuals (11–13). Our results indicate that dominance effects generally do not persist over multiple generations. The lack of evidence of reproductive skew in these communities means that heritable traits or behaviors that are passed paternally, be they genetic or cultural, are unlikely to be under strong selection.

The discovery that neutral processes can explain most haplotype frequency distributions in these communities parallels earlier results from the development of neutral theory in genetics and ecology. As Kimura (14) observed in his original article, the prevalent opinion in the 1960s held that almost all mutations are under selection. This opinion was slow to change. More recently, ecologists similarly have suggested that a neutral model, in which species in the same trophic level are functionally equivalent or neutral with respect to each other, might adequately explain species-abundance distributions in ecological communities (15–17). In anthropology, the recent availability of appropriately sampled community-level polymorphism data now enables us to distinguish both genetic and cultural selection from neutral demographic processes with surprising precision. In these Indonesian communities, male dominance seldom translates into increased fertility among descendents over evolutionary timescales.

Materials and Methods

Ewens Sampling Formula. Our statistical methods are based on the appropriateness of the Ewens sampling formula (5) as a null model for the process under study. This sampling formula applies to closed populations in situations that meet the following criteria for samples, genetic data, infinite alleles, and equilibrium:

1. Samples: The sample size n is small compared with the constant haploid population size, N.
2. Genetic Data: Each mutation is selectively neutral, taking place from 1 generation to the next with probability μ.

3. Infinite Alleles: Each mutation gives rise to a novel haplotype.
4. The population is in equilibrium.

The sampling formula gives the distribution of a sample taken from the assumed equilibrium frequency distribution or configuration of the population. A configuration is denoted by $\mathbf{b} = (b_1, b_2, \ldots)$ where b_i is the number of haplotypes represented i times in a sample of n individuals. Thus, possible sampling configurations satisfy $\Sigma_i i b_i = n$. The Ewens sampling formula states that this configuration has probability

$$P(\text{configuration } \mathbf{b}) = \frac{n!}{\theta(\theta+1)\ldots(\theta+\mathrm{n}-1)} \prod_i \left(\frac{\theta}{\mathrm{i}}\right)^{b_i} \frac{1}{b_i!}.$$

The choice of θ generally is approximately equal to $2N_e\mu$ where N_e is the effective population size.

We now examine the assumptions individually. Then we show how the Ewens sampling formula forms the basis for a test of neutrality and an estimator of θ based on the number of haplotypes. We apply the test and the estimator to the sample sites. Finally, we introduce a 2-parameter model to assess the impact of these departures from neutrality and the power of the test and compare our method with another approach.

1. Samples. In total, 1269 Y chromosomes were collected from 41 communities on 6 islands (Bali, Borneo, Flores, Java, Nias, Sumba) in Indonesia (Fig. S4). Sample sizes range from 16 to 54 individuals. All samples were collected from volunteer donors with written informed consent and appropriate permits from the Indonesian Government via the Eijkman Institute for Molecular Biology. The University of Arizona Human Subject Committee approved sampling protocols. In all cases, care was taken to exclude individuals related within the last 3 generations, as determined by detailed personal genealogies. This aspect of the sampling procedure effectively eliminates the effects of recent demographic events.

Samples were designed to be a small proportion of the village. The number of samples n in each community is given in Table S1. The assumption of constant population is a part of our null hypotheses. Departures from the constant population assumption will be considered in our power analysis.

2. Genetic data. The polymorphic sites from the nonrecombining part of human Y chromosome included a set of 74 previously published binary markers (18) and 4 additional polymorphisms: M208, M210, M346, and M356 (19–22). Binary markers were analyzed with a hierarchical strategy (23, 24) when additional sample genotyping was restricted to the appropriate downstream mutations along the haplogroup tree.

For the microsatellite analysis, 10 STRs (*DYS19*, *DYS388*, *DYS389I*, *DYS389II*, *DYS390*, *DYS391*, *DYS392*, *DYS393*, *DYS426*, and *DYS439*) were typed as described by Redd *et al.* (25). Some sites also were scored for *DYS438* and *DYS457* loci. Table S1 indicates the total number of STRs typed for each sampling location.

A haplotype is defined by its sequence of repeat numbers. Consequently, a haplotype mutates if at least 1 STR mutates. If we let M denote the number of STRs, and let ν_i denote the mutation rate for the ith Y-STR, then we can calculate the mutation rate for Y-STR haplotypes as

$$\begin{aligned}\mu &= P(\text{at least one STR mutates}) = 1 - P(\text{no STR mutates})\\ &= 1 - \prod_i P(i\text{th STR does not mutate})\\ &= 1 - \prod_i (1-\nu_i)\\ &\approx 1 - \left(1 - \frac{1}{M}\sum \nu_i\right)^M.\end{aligned}$$

Consequently, μ depends only on the average mutation rate, not its variability. Using the average STR mutation rate of 0.0021 per generation (26, 27) for the average above, we calculate:

$$\begin{aligned}10\text{ STRs: } \mu &= 1 - (1-0.0021)^{10} = 0.0208,\\ 12\text{ STRs: } \mu &= 1 - (1-0.0021)^{12} = 0.0249.\end{aligned}$$

3. Infinite alleles. Y-STRs mutate by either increasing or decreasing their repeat number. The measure of the failure of the infinite alleles assumption is the mean frequency of STR mutations that do not result in a novel haplotype. This possibility of 2 individuals being identical in state but not identical by descent is called ''homoplasy.''

A variety of STR mutation models have been studied (28–30). Mutation models that permit multiple repeat changes are less likely to mutate to an existing haplotype than mutation models that permit only a mutation changing the repeat number by 1. In addition, as we shall see, homoplasy affects tests for neutrality and estimates for θ only if it reduces the number of haplotypes in the sample. Irrespective of the model, the impact of homoplasy is to reduce the number of haplotypes and to increase the fraction of individuals that belong to common haplotypes.

4. The population is in equilibrium. Neutral models for the evolution of the configuration of individuals are Markov chains. For Markov chains that model this evolution, the distribution of the configuration of haplotypes converges to its equilibrium. The questions arise: What is the time needed for this convergence, and what is the nature of the convergence? Analytical results and simulations (data not shown here) confirm that these Markov chains possess a cutoff phenomenon. The cutoff occurs at approximately $N_e(\log\theta)/\theta$ generations. Here, we take a generation to be 31 years (31).

The analysis also shows that a neutrally evolving population has reached equilibrium if each individual in the population has experienced at least 1 mutation along the line of descent connecting that individual to a founding member of the population. If a community is neutrally evolving but the genetic data have not yet reached equilibrium, then the current population will have some signature of the founding population. Fig. S3*A* shows this phenomenon for a single neutral population that began with all individuals sharing the same haplotype.

Statistical Procedures. The Ewens sampling formula is a 1-parameter probability distribution function. It can be expressed as the product of two terms

$$\begin{aligned}&\frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_i \left(\frac{\theta}{i}\right)^{b_i} \frac{1}{b_i!}\\ &\qquad = n! \prod_i \left(\frac{1}{i}\right)^{b_i} \frac{1}{b_i!} \cdot \frac{\theta^K}{\theta(\theta+1)\cdots(\theta+n-1)}.\end{aligned}$$

For a fixed value of K, the first term gives the likelihood for a given configuration $\mathbf{b}$. The second shows that the number of haplotypes $K = \sum_i b_i$ is a sufficient statistic for the parameter θ. Note that the form of this product shows that, given the number of haplotypes, the distribution of haplotypes does not depend on the parameter θ and consequently does not depend on either the population size or the mutation rate.

Test for Neutrality. Based on the first term in the product of the Ewens sampling formula, Slatkin (6, 7) developed an exact test for significant departures from the neutral hypothesis. The exact test calculates the probability of all configurations with fixed sample size, n, and fixed haplotype number, K, and sums the probabilities of the configurations that are less likely than the observed configuration. That value, P_E, is reported in Table S1. The test is 2-tailed: values of $P_E < 0.025$ indicate a departure from neutrality in the direction of a distribution that is ''too even,'' often taken to indicate the presence of balancing selection. Values of $P_E > 0.975$ indicate departure in the direction of an ''overly uneven'' distribution, which can indicate either the presence of positive selection or demographic history. Table S1 shows the values of P_E for all of the sampled villages. Highlighted departures from neutrality have $P_E > 0.975$.

Power Analysis. To determine the power of the test to detect neutrality under the dominance model, populations were simulated under a range of parameters values for σ and δ under 2 choices for village population size and 2 levels of migration (Tables S2 and S3). For each set of parameters, 10,000 samples were simulated, the exact test was applied, and the number testing as nonneutral was counted. The estimated power of the test is the fraction testing as nonneutral. See *SI Text* for further details of the analysis

Estimate of θ. The maximum likelihood estimator of θ is found by solving (32):

$$K = \sum_{j=0}^{n-1} \frac{\theta}{\theta + j},$$

ANTHROPOLOGY

where n is the sample size. Using value of μ given earlier, we are able to estimate effective population size, N_e, using K. The values for K and the estimated values for θ for each community are given in Table S1. We compare these estimates with the number of male household heads known from census data in Table S4. Estimates for θ are biased upwards.

ACKNOWLEDGMENTS. We thank these individuals for their analysis of the genetic data at the Hammer Laboratory of the University of Arizona: Meryanne Tumonggor, Yuliana Ling, Daniel Teberg, Megan Dragony, Emily Landeen, Fernando Mendez, and Rachel Stewart. We also thank John Schoenfelder, who provided the sampling locations map, Jennifer Dunne, John Pepper, and two anonymous reviewers for comments, and the 1269 Indonesian men who made this study possible. Genetic samples were obtained by JSL and HS and by Golfiani Malik, Wuryantari Setiadi, Loa Helena Suryadi, and Meryanne Tumonggor of the Eijkman Institute, with the assistance of Indonesian Public Health clinic staff, following protocols for the protection of human subjects established by both the Eijkman Institute and the University of Arizona Institutional Review Boards. This work was supported by the National Science Foundation, the James McDonnell Foundation Robustness program at the Santa Fe Institute, and the Eijkman Institute for Molecular Biology, Jakarta, Indonesia. Permission to conduct research in Indonesia was granted by the Indonesian Institute of Sciences.

1. Clarke AL, Low BS (2001) Testing evolutionary hypotheses with demographic data. *Popul Dev Rev* 27:633–660.
2. Winterhalder B, Smith E (2000) Analyzing adaptive strategies: Human behavioral ecology at twenty-five. *Evol Anthropol* 9:51–72.
3. Cannings C (1974) The latent roots of certain Markov chains arising in genetics: A new approach, I. Haploid models. *Adv Appl Probab* 6:260–290.
4. Sibert A, Austerlitz F, Heyer E (2002) Wright-Fisher revisited: The case of fertility correlation. *Theor Popul Biol* 62:181–197.
5. Ewens WJ (1972) The sampling theory of selectively neutral alleles. *Theor Popul Biol* 3:87–112.
6. Slatkin M (1996) A correction to the exact test based on the Ewens sampling distribution. *Genet Res* 68:259–260.
7. Heyer E, Sibert A, Austerlitz F (2005) Cultural transmission of fitness: Genes take the fast lane. *Trends Genet* 21(4):234–239.
8. Van Valen L (1973) A new evolutionary law. *Evol Theor* 1:1–30.
9. Lansing JS, *et al.* (2007) Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc Natl Acad Sci USA* 104:16022–16026.
10. Lansing JS (2006) *Perfect Order*: *Recognizing Complexity in Bali* (Princeton Univ Press, Princeton, New Jersey).
11. Fisher RA (1930) *The Genetical Theory of Natural Selection* (Clarendon Press, Oxford, UK).
12. Hofbauer J, Sigmund K (1991) *The Theory of Evolution and Dynamical Systems* (Cambridge Univ Press, Cambridge, UK).
13. Nowak, M. A (2006) *Evolutionary Dynamics*: *Exploring the Equations of Life* (Harvard Univ Press, Cambridge, Massachusetts).
14. Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217:624–626.
15. Hubbell SP (2001) *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton Univ Press, Princeton, New Jersey).
16. Hu X-S, He F, Hubbell SP (2006) Neutral theory in macroecology and population genetics. *Oikos* 113:548–556.
17. Alonso D, Etienne R, McKane A (2006) The merits of neutral theory. *Trends Ecol Evol* 21:451–457.
18. Karafet TM, *et al.* (2005) Balinese Y-chromosome perspective on the peopling of Indonesia: Genetic contributions from pre-Neolithic hunter-gatherers, Austronesian farmers, and Indian traders. *Hum Biol* 77:93–114.
19. Kayser M, *et al.* (2003) Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *Am J Hum Genet* 72:281–302.
20. Kayser M, *et al.* (2006) Melanesian and Asian origins of Polynesians: MtDNA and Y-chromosome gradients across the Pacific. *Mol Biol Evol* 23:2234–2244.
21. Hudjashov G, *et al.* (2007) Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc Natl Acad Sci USA* 104:8726–8730.
22. Sengupta S, *et al.* (2005) Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian Pastoralists. *Am J Hum Genet* 78:202–221.
23. Underhill PA, *et al.* (2000) Y chromosome sequence variation and the history of human populations. *Nat Genet* 26:358–361.
24. Hammer MF, *et al.* (2001) Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol* 18:1189–1203.
25. Redd AJ, *et al.* (2002) Forensic value of 14 novel STRs on the human Y chromosome. *Forensic Sci Int* 130:97–111.
26. Dupuy BM, Stenersen M, Egeland T, Olaisen B (2004) Y-chromosomal microsatellite mutation rates: Differences in mutation rate between and within loci. *Hum Mutat* 23:117–124.
27. Gusmão L, *et al.* (2005) Mutation rates at Y chromosome specific microsatellites. *Hum Mutat* 26:520–528.
28. Di Rienzo A, *et al.* (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proc Natl Acad Sci USA* 91:3166–3170.
29. Sainudiin R, Durrett RT, Aquadro CF, Nielsen R (2004) Microsatellite mutation models: Insights from a comparison of humans and chimpanzees. *Genetics* 168:383–395.
30. Watkins JC (2007) Microsatellite evolution: Markov transition functions for a suite of models. *Theor Popul Biol* 71:147–159.
31. Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol* 128:415–423.
32. Tavaré, S., Zeitouni, O (2004) *Lectures on Probability Theory and Statistics. Ecole d'Etés de Probabilité de Saint-Flour XXXI* (Springer-Verlag, New York).

12

COEVOLUTION OF LANGUAGES AND GENES ON THE ISLAND OF SUMBA, EASTERN INDONESIA

Lansing, J.S., **M.P. Cox**, S.S. Downey, B.M. Gabler, B. Hallmark, T.M. Karafet, P. Norquest, J.W. Schoenfelder, H. Sudoyo, J.C. Watkins, and M.F. Hammer. 2007. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proceedings of the National Academy of Sciences USA* 104:16022-16026.
http://doi.org/10.1073/pnas.0704451104

ABSTRACT Numerous studies indicate strong associations between languages and genes among human populations at the global scale, but all broader scale genetic and linguistic patterns must arise from processes originating at the community level. We examine linguistic and genetic variation in a contact zone on the eastern Indonesian island of Sumba, where Neolithic Austronesian farming communities settled and began interacting with aboriginal foraging societies ≈3,500 years ago. Phylogenetic reconstruction based on a 200-word Swadesh list sampled from 29 localities supports the hypothesis that Sumbanese languages derive from a single ancestral Austronesian language. However, the proportion of cognates (words with a common origin) traceable to Proto-Austronesian (PAn) varies among language subgroups distributed across the island. Interestingly, a positive correlation was found between the percentage of Y chromosome lineages that derive from Austronesian (as opposed to aboriginal) ancestors and the retention of PAn cognates. We also find a striking correlation between the percentage of PAn cognates and geographic distance from the site where many Sumbanese believe their ancestors arrived on the island. These language-gene-geography correlations, unprecedented at such a fine scale, imply that historical patterns of social interaction between expanding farmers and resident hunter-gatherers largely explain community-level language evolution on Sumba. We propose a model to explain linguistic and demographic coevolution at fine spatial and temporal scales.

Coevolution of languages and genes on the island of Sumba, eastern Indonesia

J. Stephen Lansing[†‡§], Murray P. Cox[¶], Sean S. Downey[†], Brandon M. Gabler[†], Brian Hallmark[‖], Tatiana M. Karafet[¶], Peter Norquest[†], John W. Schoenfelder[†,††], Herawati Sudoyo[‡‡], Joseph C. Watkins[‖], and Michael F. Hammer[¶]

[†]Department of Anthropology, University of Arizona, 1009 East South Campus Drive, Tucson, AZ 85721; [¶]Division of Biotechnology, Biosciences West, University of Arizona, Tucson, AZ 85721; [‡]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87521; [‖]Department of Mathematics, University of Arizona, 617 North Santa Rita Avenue, Tucson, AZ 85721; [††]Cotsen Institute of Archaeology, University of California, 308 Charles E. Young Drive North, Los Angeles, CA 90095; and [‡‡]Eijkman Institute for Molecular Biology, Diponegoro 69, Jakarta 10430, Indonesia

Edited by Simon A. Levin, Princeton University, Princeton, NJ, and approved August 23, 2007 (received for review May 15, 2007)

Numerous studies indicate strong associations between languages and genes among human populations at the global scale, but all broader scale genetic and linguistic patterns must arise from processes originating at the community level. We examine linguistic and genetic variation in a contact zone on the eastern Indonesian island of Sumba, where Neolithic Austronesian farming communities settled and began interacting with aboriginal foraging societies ≈3,500 years ago. Phylogenetic reconstruction based on a 200-word Swadesh list sampled from 29 localities supports the hypothesis that Sumbanese languages derive from a single ancestral Austronesian language. However, the proportion of cognates (words with a common origin) traceable to Proto-Austronesian (PAn) varies among language subgroups distributed across the island. Interestingly, a positive correlation was found between the percentage of Y chromosome lineages that derive from Austronesian (as opposed to aboriginal) ancestors and the retention of PAn cognates. We also find a striking correlation between the percentage of PAn cognates and geographic distance from the site where many Sumbanese believe their ancestors arrived on the island. These language–gene–geography correlations, unprecedented at such a fine scale, imply that historical patterns of social interaction between expanding farmers and resident hunter-gatherers largely explain community-level language evolution on Sumba. We propose a model to explain linguistic and demographic coevolution at fine spatial and temporal scales.

Austronesian languages | cognate | contact zone | language evolution | Y chromosome haplogroups

Languages, like populations, change over time, but the rules governing language change are still not well understood. Because lexical and structural innovation, borrowing, and loss are difficult to observe and quantify over brief periods (1) and are impossible to witness over long periods, researchers are forced to undertake indirect approaches to infer the processes of language change. One such approach is to look for associations between linguistic and genetic classifications. Many well-known studies have identified associations between the languages and genes of human populations at continental and global geographic scales (2–6). A survey of these studies led Diamond and Bellwood (7) to hypothesize that many of these correlations are caused by the linked spread of prehistoric farmers and their languages outward from a number of widely dispersed agricultural homelands in Africa, the Near East/Europe, Asia, and the Americas. Under the simplest form of their hypothesis, genetic and linguistic variation evolves in parallel after the genes and languages of farmers replace those of hunter-gatherers in the path of expansion (Fig. 1*A*). According to Diamond and Bellwood (7), one of the best examples of the coevolution of language and genes was brought about by the Neolithic expansion of Austronesian-speaking farmers into previously uninhabited Polynesia and Micronesia.

Discrepancies between genetic and linguistic differentiation can arise through a number of processes (4, 8), perhaps the most important of which are genetic admixture (i.e., without language change) and language replacement (Fig. 1*A*) (7, 9, 10). These processes, which occur when migrating farmers meet resident hunter-gatherers face to face, likely characterize the expansion of Austronesian speakers into regions that were long occupied by indigenous populations in eastern Indonesia and New Guinea (7, 11, 12). However, most language–gene studies have sampled at a geographic scale, which is too coarse to permit any refined inference about the dynamics of language change in these contact zones. Information at a finer scale is essential to characterize the nature of contact relationships and infer mechanisms of linguistic and genetic transformation over recent temporal and fine spatial scales.

Toward this goal, we examine linguistic and genetic variation in a contact zone on the Indonesian island of Sumba. Broader regional studies support the initial settlement of Southeast Asia/Oceania by foraging societies by 40,000 to 45,000 BP (13). Languages of the geographically expansive Austronesian family occupy much of the Indonesian archipelago, except in far eastern Indonesia, where diverse and unrelated Papuan languages dominate. Recent syntheses place the Neolithic transition, considered to mark the arrival of Austronesian colonists in the vicinity of Sumba, at between 4,000 and 3,500 years ago (14, 15). At that time, small numbers of farmers speaking an Austronesian language likely came into contact with an indigenous population of foragers speaking aboriginal languages. Several circumstances favor Sumba as a site to investigate the relationship between population incursion and language change. Sumba is remote and culturally conservative, the last island in the archipelago where the majority adhered to a tribal or pagan religion at the close of the 20th century. Today, nearly all Sumbanese live in traditional farming villages composed of patrilocal clans. Contact between villages is limited, and population density is low. Perhaps the most telling indicator of the extent of contact between villages is the large number of languages now spoken on the island despite its small size (220×75 km^2). In this report, we reconstruct Sumbanese language relationships using a 200-word Swadish list, and we examine Y chromosome SNP and short tandem repeat (STR) diversity in a sample of Sumbanese villages. We propose a model of language–gene coevolution to explain the striking associations we observe among linguistic,

Author contributions: J.S.L. designed research; J.S.L., S.S.D., B.H., T.M.K., and H.S. performed research; S.S.D., B.M.G., B.H., T.M.K., P.N., J.W.S., and J.C.W. analyzed data; and J.S.L., M.P.C., B.H., P.N., J.C.W., and M.F.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Abbreviations: PAn, Proto-Austronesian; PS, Proto-Sumba; STR, short tandem repeat.

[§]To whom all correspondence should be addressed. E-mail: lansing@santafe.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0704451104

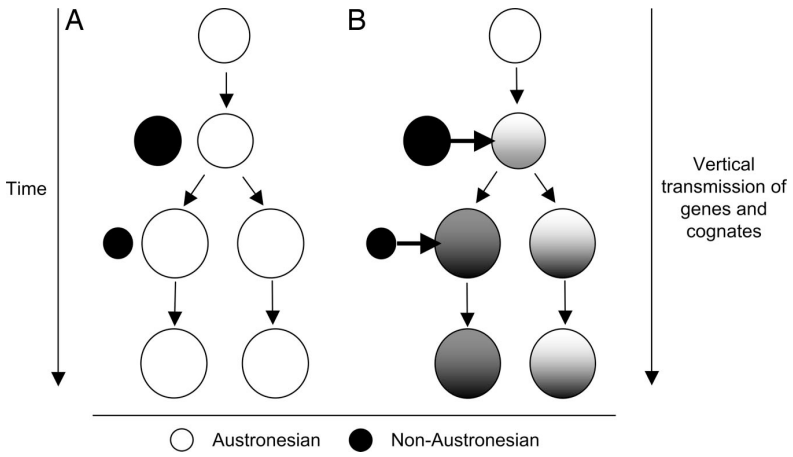

Fig. 1. Models for the evolution of languages and genes at two scales. Each circle represents a population of languages or genes, with parent populations shown on top and descendant populations shown below downward-facing arrows. Open circles, invading farming populations; filled circles, resident aboriginal populations. (*A*) Replacement models at larger geographic and temporal scales include two favored models in the literature: language replacement (i.e., the languages of an incoming population replace those of resident groups without gene flow) (10) and Diamond and Bellwood's (7) basic hypothesis (i.e., linguistic and genetic replacement by an incoming group with subsequent coevolution of descendant languages and genes). (*B*) An alternative model with codominant effects at smaller geographic and temporal scales involves both genetic admixture (e.g., demic diffusion) and the incursion of words that do not trace to PAn (horizontal arrows) in each descendant population after arrival of a founding Austronesian population (circle at center and top). A greater number of noncognates enters the population in the western part of Sumba where there are lower frequencies of Austronesian Y chromosome lineages (larger filled circles and thicker horizontal arrows) relative to the central part of Sumba.

genetic, and geographical data sets sampled at a fine geographic scale.

Results and Discussion

Linguistic Variation. We gathered twenty-nine 200-word Swadesh lists from diverse sites on the island (Fig. 2) and used traditional comparative linguistic approaches to identify cognates (i.e., words in two or more languages that can be traced to a common ancestor), sound correspondences, innovations, and loan words. On average, the Swadesh list for a Sumbanese language contains ≈70 cognates and 130 noncognates. In other words, ≈35% of the 200-word lexicon is directly descended from Proto-Austronesian

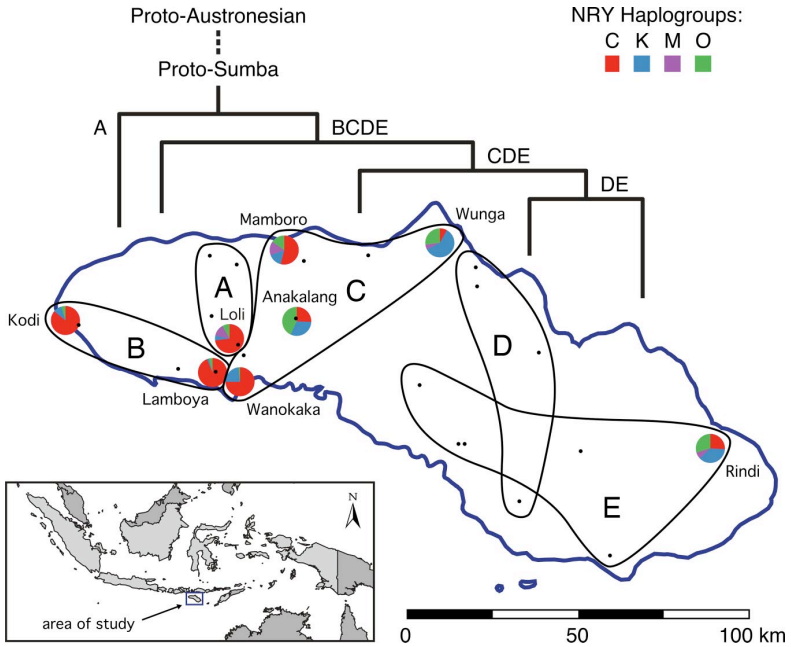

Fig. 2. Phylogenetic and geographic distributions of languages and Y chromosomes on Sumba. (*Upper*) Phylogenetic tree of Sumban language groups (A–E) (see *Materials and Methods*). (*Lower*) Map of Sumba showing geographic distribution of language groups (A–E) and Y chromosome haplogroups (C, K, M, and O). Pie charts represent frequencies of four Y chromosome haplogroups at eight locations sampled for both DNA and languages. Haplogroup O (green) is unevenly distributed, with lower frequencies in the western portion of the island. Small black dots indicate 20 additional language samples for which paired DNA samples were not available.

(PAn), whereas 65% are the result of innovations (i.e., word changes, losses, borrowings, etc.) that have occurred since the time of the common PAn ancestral language. Phylogenetic analysis was performed on the resulting database to reconstruct the relationships among these languages and to infer the scope and directionality of lexical borrowing and phonological change. The resulting tree clusters the languages of the 29 samples into five subgroups and strongly supports the hypothesis that Sumbanese languages derive from a single ancestral language, Proto-Sumba (PS), the identifiable lexicon of which is Austronesian in origin (Fig. 2). A sizable amount of the PAn vocabulary has been reconstructed (16, 17), and all of the languages in our sample retain PAn cognates. However, the proportion of these cognates in each language varies, and the lexicon of PS also contains many words that cannot be traced to PAn. Phonological and lexical variation between languages in our sample shows a clear geographic structure, with the main subgroups distributed across the island from west to east (Fig. 2). There is a pattern in the retention of PAn cognates, with the central languages being more conservative than those on the periphery, particularly the western languages. Sound change isoglosses (geographically bounded linguistic features) also are suggestive of higher heterogeneity in the western half of the island, where group A (and the northwest section of group C, apparently in heavy contact with group A) has been excluded by one of the most sweeping sound changes across the island, and where groups A and B (and again northwest C) have innovated another sound change that has not spread further east.

Y Chromosome Diversity. To investigate the paternal genetic relationships among Sumbanese villages, we obtained samples from 352 men inhabiting eight villages that also were sampled for languages and genotyped 71 SNPs. We identified 17 Y chromosome SNP haplogroups belonging to the C, K, M, and O lineages. These haplogroups have different histories: C-RPS4Y*, C-M38, K-M230, and M-P34 are believed to be associated with the colonization of eastern Indonesia by Paleolithic peoples. They are shared among eastern Indonesian, Papua New Guinean, and Melanesian populations and are absent or marginally present outside this region. An indigenous origin of K-M9* Y chromosomes also is suggested by its geographic distribution and the associated Y-STR diversity (data not shown). In contrast, haplogroup O likely entered this region much more recently (12, 18). Although there is controversy about the geographic origins of haplogroup O, there is general agreement that it is associated with the Late Holocene expansion of Austronesian-speaking farmers from southeast Asia to Indonesia and Oceania. Overall, only 16% of Sumbanese Y chromosomes belong to haplogroup O. The proportion varies among communities and in general exhibits a decreasing gradient from east to west, becoming rare in the western end of the island (Fig. 2).

Associations Among Linguistic, Genetic, and Geographic Distances. To test for language–gene associations, we performed matrix correlation tests (19) to evaluate the correspondence between linguistic distances determined from the reconstructed language tree and genetic distances among the eight sampled villages (Table 1). We observe a statistically significant positive correlation between linguistic and genetic distances ($r = 0.358$, $P = 0.023$). As pointed out by Nettle and Harriss (5), such a correlation could mean that languages and genes either coevolved from a single common ancestor (i.e., were shaped by the same events in population history) or evolved separately, but were conditioned by the same factors (i.e., were subjected to parallel but separate isolation by distance processes). One way to investigate this question is to control for geographic distance and test for a residual relationship between linguistic and genetic affiliations (5). Our finding of a slightly stronger correlation between linguistic and genetic variation when geography was held constant ($r = 0.474$, $P = 0.003$) provides evidence that languages and genetic variants on Sumba have actually evolved together. To verify that this association emerged within the time frame of the Austronesian expansion, we estimated the divergence time between the two most geographically distant communities in our sample, the villages of Rindi and Kodi, with an isolation-with-migration coalescent simulation model. The upper limit of the 95% confidence interval generated by the coalescent model is 4,875 years, which is consistent with the timing of the Austronesian expansion (20).

Table 1. Mantel test results for genetic, linguistic, and geographic distances

Correlation	*r*	*P*
Full correlation		
Genetics/geography	0.011	0.518
Genetics/language	0.358	0.023
Geography/language	0.673	0.000
Partial correlation		
Genetics/geography (language held constant)	−0.332	0.886
Genetics/language (geography held constant)	0.474	0.003
Language/geography (genetics held constant)	0.716	0.001

Genetic distance: Slatkin linearized $R_{ST}/1 - R_{ST}$. Language distance: ALINE distance (30).

We also found a strong positive correlation between linguistic and geographic distances, which becomes stronger when controlling for genetic variation ($r = 0.716$, $P = 0.001$), and no correlation between genetic and geographic distances (Table 1). The latter result is not surprising given the recent spread of farming on Sumba and the length of time typically needed for restricted gene flow between communities to produce a pattern of isolation by distance (21, 22). Moreover, because haplogroup O only accounts for ≈16% of Sumbanese Y chromosomes, we may have little power to detect a correlation between geographic and genetic differentiation associated with the expansion of Austronesian farmers. Focusing on the Austronesian component of genetic variation, we do find a positive correlation ($r = 0.627$, $P = 0.047$, by using a bootstrap approach; see *Materials and Methods*) between the percentages of haplogroup O and PAn cognates retained in the eight villages sampled for Y chromosomes (Fig. 3). This unprecedented correlation provides further

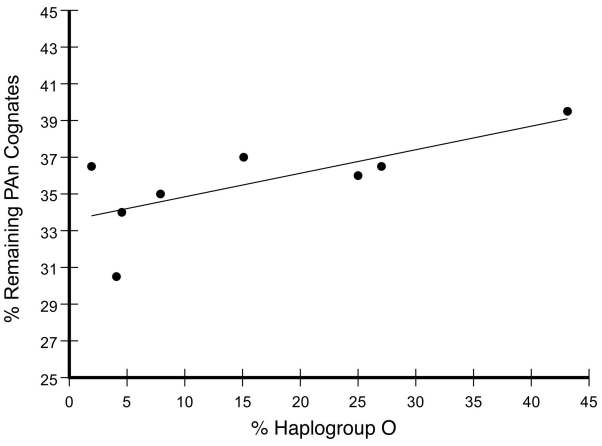

Fig. 3. Scatterplot of the percentage of PAn cognates versus the percentage of Austronesian Y chromosomes (haplogroup O) found at each location.

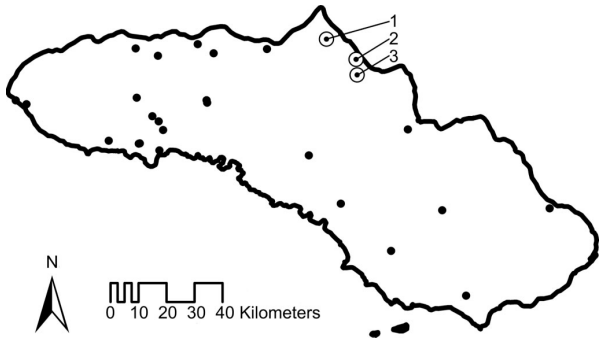

Fig. 4. Map showing approximate geographic locations of 29 language samples (black dots). Each of the three locations show a strongly significant correlation between their geographic distance from all other locations, and the percentage of PAn cognates retained are listed on the map (1, Wunga: *r*, −0.503; *P*, 0.006; 2, Rambangaru: *r*, −0.015; *P*, 0.005; 3, Kanatang: *r*, −0.507; *P*, 0.006).

evidence for the coevolution of linguistic and genetic variation on Sumba.

To investigate how the settlement history of the island may have influenced language evolution, we tested for correlations between the percentage of retained PAn cognates and geographic distances from a putative source population. We measured the geographic distances between each (putative source) population and all other populations and tested for a correlation with the percentage of PAn cognates retained. Only 3 of the 29 tested populations show a strongly significant correlation ($P < 0.01$), all of which are located on the central northern coast (Fig. 4). This result is concordant with oral history suggesting an origin near the village of Wunga (23) and implies that population history had a strong influence on patterns of language variation. Indeed, the language analysis (Fig. 2) is consistent with a splitting model in which a common founding population gave rise to daughter populations that subsequently diverged and produced several language subgroups. We hypothesize that, after settling near Wunga, the ancestral Austronesian population on Sumba expanded southward toward the center of the island. The first population split, which resulted in group A in the northwest, may have occurred before the Austronesians expanded south. By the time group B split and moved into the southwest, the main population must have expanded at least to the center of the island. This finding implies a later expansion to the east, with groups D and E splitting after the initial western expansion and probably after the main population (represented by group C) overtook the center of the island. Given the linguistic and genetic picture detailed previously, we hypothesize that this expansion must have involved a high degree of contact and intermarriage with the aboriginal population in successive stages, which would explain the generally low distribution of haplogroup O, its uneven distribution across the island, and the correlation with the percentage of retained Austronesian vocabulary in the subgroups.

Models of Language Change. This investigation of language and genetic variation on the island of Sumba produced results that are not easily explained by models of language evolution formulated on the basis of large-scale patterns of language variation (Fig. 1*A*). For example, simple models of language replacement without gene flow (e.g., elite dominance) or complete replacement of genes and languages (7) are not appropriate given the evidence for genetic admixture between Austronesian farmers and indigenous Papuan populations. Indeed, the frequency of indigenous Y chromosomes surpasses that of Austronesian Y chromosomes on Sumba (84% vs. 16%), with haplogroup O varying from 25–45% in the central and eastern parts of the island to <5% in the west. This distribution is consistent with a pattern of demic diffusion, whereby the incremental spread of farmers from their point of entry on the island was accompanied by frequent intermarriage with resident hunter-gatherers (9). It is unlikely that indigenous languages were fully replaced during the initial expansion of Austronesian on Sumba because we observe a high proportion of words (65%) that cannot be traced to PAn and loan words shared between different language groups that may have been absorbed from a now extinct indigenous source. Evidence for the latter hypothesis comes from the presence of non-Austronesian words (in particular, culturally significant words such as *husband*, *animal*, *dog*, and *sea*) in groups A and B (which do not form a subgroup) and their absence in subgroups C, D, and E to the east. Given the phylogenetic relationships in Fig. 2, this pattern is more easily explained by loans of these vocabulary items from a common non-Austronesian source, rather than by losses of ancestral Austronesian words in PS and later recovery in groups C, D, and E.

To account for these patterns of linguistic and genetic variation, we propose an alternative model of language evolution appropriate for the spatial scale of Sumba (Fig. 1*B*). In this model, intermarriage between expanding farmers and resident hunter-gatherers leads to progressively lower frequencies of haplogroup O Y chromosomes at increasing distances from the source population. What factors could lead to an association between Austronesian male lineages and the retention of PAn vocabulary across Sumba? Climate and population density data suggest that eastern Sumba remained sparsely populated during this expansion and new agricultural communities were relatively isolated. The north coast of East Sumba is the driest region in Indonesia, whereas West Sumba averages nearly three times more annual rainfall. This climatic variation is reflected in contemporary population densities: 28/km^2 in East Sumba and 97/ km^2 in West Sumba (24). We infer that, in preagricultural times, Sumba probably resembled aboriginal Australia, where human population density scaled with rainfall (25). In the wetter and more fertile region of West Sumba, expanding farmers likely came into contact with a larger indigenous population speaking non-Austronesian languages. This theory is attested to by lower frequencies of Austronesian lexical items (presumed to be due, at least in part, to loan words), as well as certain prominent phonological patterns in the west. As new farming villages proliferated in the populous west, the proportion of settlers of Austronesian descent would decrease, whereas the opportunities for linguistic contact would increase. Over time, these community-level processes gave rise to differential rates of language divergence/lexical borrowing and the association between languages and genes on Sumba. This scenario suggests a mechanism for language change: Rather than elite dominance, where a few individuals of an invading culture impose their language on a resident population, the extent of retention of PAn items is governed by the proportion of men in the population with Austronesian paternal ancestry. This codominant model also differs from the basic hypothesis of Diamond and Bellwood (7), in that a linguistic–genetic association evolves despite ongoing processes of demic diffusion and language shift. Whether the processes integrated in this model can explain patterns observed at continental scales remains an open question. However, a link can be postulated because large-scale patterns are contingent on processes occurring at local scales. This finding may be particularly true in the many cases of languages and genes spread by the recent dispersal of farmers (7, 26). More local-scale studies in contact zones with variable degrees of interaction among groups speaking different languages (e.g., Bantu and Khoisan in southern Africa, Indo-Iranian and Tibeto-Burman in south Asia, etc.) would be particularly helpful for determining the generality

of the model presented here. By incorporating lists of culturally appropriate words reflecting functional differences between farming and indigenous populations, future studies also may reveal more about the social dynamics favoring the retention of borrowed words. For now, the evidence provided here strongly suggests that language change in contact zones is scalar, language change can potentially vary in magnitude and character depending on factors inherent in the individual contact situation, and genetic analysis is a powerful tool that can be used to help formulate hypotheses of incipient language speciation.

Materials and Methods

Linguistic Classification. Lexical samples from 18 Sumbanese languages were obtained from lists collected and published by the National Language Center of the Indonesian National Department of Education (27) and cross-checked with word lists videotaped by J.S.L. and H.S at the sample locations. Videotaped recordings of 11 additional languages also were recorded at this time. These materials were organized and analyzed according to the principles of the traditional comparative method. Languages were first organized into rough groups according to shared lexical items. Where PAn or its descendent protoforms were known, these formed a backdrop with which to compare individual words, so that lexical innovations could be tagged as such and grouped together when shared. The second step was to organize these larger groups into subgroups based on shared phonological features. General reconstructions of lexical items were often possible at this point, and individual words could be compared with these reconstructions and phonological innovations noted. Languages that showed the same innovations were grouped together where appropriate. Finally, a search for loan words was conducted, the criteria being lexical and/or phonological innovations or retentions, which were unexpected within a certain subgroup. When potential loan words were identified, donor languages were sought out and, in many cases, identified based on geographic proximity to the borrowing language(s).

Y Chromosome Analysis. Two classes of markers on the Y chromosome, including 71 SNPs and 12 STRs, were genotyped as described elsewhere (18, 28).

Statistical Analyses. Geographic distances were calculated as a great circle distance based on GPS coordinates taken at the sample locations. Slatkin's linearized R_{ST} distances based on 12 Y chromosome microsatellites calculated with ARLEQUIN 3.0 (29) was used as the genetic distance. Quantitative measurement of distance between all pairs of languages was made by using ALINE distance based on the ALINE algorithm (30). This algorithm generates a score reflecting the phonetic similarity between words, which is converted to a distance by using a methodology currently in review (S.S.D., B.H, P.N., M.P.C, and J.S.L., unpublished data). The distance between two languages is calculated as the average distance between shared-meaning word pairs for those languages. To evaluate the correlation among linguistic, genetic, and geographic distances, we performed Mantel tests with ARLEQUIN 3.0 (27). A bootstrap analysis was used to estimate confidence intervals in the correlation between the percentages of PAn cognates and haplogroup O (see Fig. 3).

Swadesh word lists for Sumbanese languages were provided by the National Language Center of the Indonesian Department of Education. The Indonesian Institute of Science assisted with data collection. This work was supported by the National Science Foundation, the James McDonnell Foundation Robustness program at the Santa Fe Institute, and the Eijkman Institute for Molecular Biology.

1. Labov W (1994) *Principles of Linguistic Change*: *Internal Factors* (Blackwell, Oxford).
2. Barbujani G, Sokal RR (1990) *Proc Natl Acad Sci USA* 87:1816–1819.
3. Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J (1988) *Proc Natl Acad Sci USA* 85:6002–6006.
4. Chen JT, Sokal RR, Ruhlen M (1995) *Hum Biol* 67:595–612.
5. Nettle D, Harriss L (2003) *Hum Biol* 75:331–444.
6. Sokal RR (1988) *Proc Natl Acad Sci USA* 85:1722–1726.
7. Diamond J, Bellwood P (2003) *Science* 300:597–603.
8. Barbujani G (1997) *Am J Hum Genet* 61:1011–1014.
9. Ammerman AJ, Cavalli-Sforza LL (1984) *Neolithic Transition and the Genetics of Populations in Europe* (Princeton Univ Press, Princeton, NJ).
10. Renfrew C (1987) *Archaeology and Language* (Jonathan Cape, London).
11. Cox MP, Lahr MM (2006) *Am J Hum Biol* 18:35–50.
12. Kayser M, Brauer S, Cordaux R, Casto A, Lao O, Zhivotovsky LA, Moyse-Faurie C, Rutledge RB, Schiefenhoevel W, Gil D, *et al.* (2006) *Mol Biol Evol* 23:2234–2244.
13. O'Connell JF, Allen J (2004) *J Archaeol Sci* 31:835–853.
14. Bellwood P (1997) *Prehistory of the Indo-Malaysian Archipelago* (Univ of Hawaii Press, Honolulu).
15. Spriggs M (2003) *Rev Archaeol* 24:57–80.
16. Blust R (1995) *J World Prehist* 9:453–510.
17. Tryon DT (1995) *The Austronesian Languages* (Mouton de Gruyter, Berlin).
18. Karafet TM, Lansing JS, Redd AJ, Reznikova S, Watkins JC, Surata SP, Arthawiguna WA, Mayer L, Bamshad M, Jorde LB, *et al.* (2005) *Hum Biol* 77:93–114.
19. Mantel N (1967) *Cancer Res* 27:209–220.
20. Bellwood P (2005) *The First Farmers*: *The Origins of Agricultural Societies* (Blackwell, Oxford).
21. Hammer MF, Redd AJ, Wood ET, Bonner MR, Jarjanazi H, Karafet T, Santachiara-Benerecetti S, Oppenheim A, Jobling MA, Jenkins T, *et al.* (2000) *Proc Natl Acad Sci USA* 97:6769–6774.
22. Slatkin M (1993) *Evolution (Lawrence, Kans.)* 47:264–279.
23. Hoskins J (1993) *The Play of Time*: *Kodi Perspectives on Calendars*, *History*, *and Exchange* (Univ of California Press, Berkeley, CA).
24. Badan PS, Kabupaten ST (2004). *Sumba Timur in Figures 2003* (Percetakan Usaha Mulia, Waikabubak, Sumba).
25. Yengoyan A (1972) *Oceania* 43:85–95.
26. Renfrew C (2000) *Cambridge Archaeol J* 10:7–34.
27. Bahasa P (2002) *Kosakata Dasar Swadesh di Kabupaten Belu*, *Ngada*, *Sumba Barat*, *Sumba Timur*, *dan Timor Tengah Utara* (Departemen Pendidikan Nasional, Rawamangun, Jakarta).
28. Redd AJ, Agellon AB, Kearney VA, Contreras VA, Karafet T, Park H, de Knijff P, Butler JM, Hammer MF (2002) *Forensic Sci Int* 130:97–111.
29. Excoffier L, Laval G, Schneider S (2005) *Evol Bioinf Online* 1:47–50.
30. Downey SS, Hallmark B, Cox MP, Norquest P, Lansing JS (2007) *Working Paper of the Santa Fe Institute* 07-08-021 (Santa Fe Inst, Santa Fe, NM).

13

CONTRIBUTIONS

As is standard in my field of research, most publications in this thesis list multiple coauthors. This is especially common in my particular research area, molecular anthropology, which is by its very nature highly interdisciplinary. I routinely collaborate with anthropologists, linguists, computer scientists, statisticians and mathematicians, as well as other geneticists. With a rare background in both the sciences and humanities, I frequently was (and continue to be) the interface between these two broad areas, translating concepts between geneticists and researchers from the other disciplines. Within this collaborative multidisciplinary environment, my single sole author paper in this collection is perhaps better seen as the anomaly.

Importantly, though, all ten publications highlighted here describe research in which I provided a substantive intellectual input. My role has evolved over the decade in which these papers appeared. Initially, as a senior postdoctoral fellow, I performed most of the research myself, or sometimes in very close collaboration with graduate students or interns who I actively managed and advised as a senior researcher in what was a large international team. Today, as a full professor, I run my own group and manage research projects on a day-to-day basis with a close team of postdoctoral fellows, PhD students, research assistants and interns. What has remained constant throughout this period is my role as a lead in the genetics space and the key figure developing many of the primary ideas in each paper. These publications would have looked very different had I not provided key intellectual direction. Most – especially the later publications – would never have been written at all.

That said, two senior advisors who played important roles early in my career deserve a special mention here. The first is Mike Hammer, a molecular geneticist at the University of Arizona who funded me as a postdoctoral fellow from 2005 to 2009, and generously gave me an extraordinary level of intellectual and operational independence. The earliest papers in this collection were only possible because of the academic freedom he provided. The second is Steve Lansing, a cultural anthropologist who created a fellowship for me at the Santa Fe Institute for Complexity Science in 2007, and gave me free rein (and indeed lively encouragement) to pursue what must have often seemed at first like crazy ideas. Some of them even turned out to be right. While the papers in this thesis are representative of my intellectual contributions to the field, like all scientists before me, I stand on the shoulders of giants.

CHAPTER 3 Guillot, E.G., M.L. Hazelton, T.M. Karafet, J.S. Lansing, H. Sudoyo, and **M.P. Cox**. 2015. Relaxed observance of traditional marriage rules allows social connectivity without loss of genetic diversity. *Molecular Biology and Evolution* 32:2254-2262.

MY CONTRIBUTIONS Proposed the central ideas; designed the project; co-developed individual analyses; managed the overall project; co-wrote the manuscript; funded the project; senior/last author.

COAUTHOR CONTRIBUTIONS

Elsa G. Guillot *Massey University, NZ*	Graduate student (supervised); implemented code and performed population genetic analyses.
Martin L. Hazelton *Massey University, NZ*	Statistics collaborator; provided advice on statistical inference procedures.
Tatiana M. Karafet *University of Arizona, USA*	Laboratory manager; prepared DNA samples for commercial marker screening.
J. Stephen Lansing *Nanyang Technological University, Singapore*	Anthropology collaborator; provided insight regarding the cultural anthropology of the study region.
Herawati Sudoyo *Eijkman Institute for Molecular Biology, Indonesia*	Medical collaborator; performed fieldwork in Indonesia and provided samples.

CHAPTER 4 Sanderson, J., H. Sudoyo, T.M. Karafet, M.F. Hammer and **M.P. Cox**. 2015. Reconstructing past admixture processes from local genomic ancestry using wavelet transformation. *Genetics* 200:469-481.

MY CONTRIBUTIONS Proposed the central ideas; designed the project; developed individual analyses, particularly the simulation model; ran individual analyses, including most of the simulations; managed the project; co-wrote the manuscript; funded the project; senior/last author.

COAUTHOR CONTRIBUTIONS

Jean Sanderson *Massey University, NZ*	Postdoctoral fellow (supervised); implemented code and performed population genetic analyses.

Herawati Sudoyo *Eijkman Institute for Molecular Biology, Indonesia*	Medical collaborator; performed fieldwork in Indonesia and provided samples.
Tatiana M. Karafet *University of Arizona, USA*	Laboratory manager; prepared DNA samples for commercial marker screening.
Michael F. Hammer *University of Arizona, USA*	Genetics collaborator; provided samples and funded commercial marker screening.

CHAPTER 5 Tumonggor, M.K., T.M. Karafet, B. Hallmark, J.S. Lansing, H. Sudoyo, M.F. Hammer, and **M.P. Cox**. 2013. The Indonesian archipelago: An ancient genetic highway linking Asia and the Pacific. *Journal of Human Genetics* 58:165-173.

MY CONTRIBUTIONS Co-designed the project; co-developed individual analyses; ran individual analyses, particularly for sex-biased processes; managed the overall project; co-wrote the manuscript; co-funded the project; senior/last author.

COAUTHOR CONTRIBUTIONS

Meryanne K. Tumonggor *University of Arizona, USA*	Graduate student (supervised); generated mitochondrial DNA sequences and performed population genetic analyses.
Tatiana M. Karafet *University of Arizona, USA*	Laboratory manager; extracted DNA from samples and sequenced mitochondrial DNA.
Brian Hallmark *University of Arizona, USA*	Graduate student; performed some of the statistical tests.
J. Stephen Lansing *Santa Fe Institute, USA*	Anthropology collaborator; provided insight regarding the cultural anthropology of the study region and co-funded the project.
Herawati Sudoyo *Eijkman Institute for Molecular Biology, Indonesia*	Medical collaborator; performed fieldwork in Indonesia and provided samples.
Michael F. Hammer *University of Arizona, USA*	Genetics collaborator; provided samples and co-funded the project.

CHAPTER 6 **Cox, M.P.**, M.G. Nelson, M.K. Tumonggor, F.-X. Ricaut, and H. Sudoyo. 2012. A small cohort of Island Southeast Asian women founded Madagascar. *Proceedings of the Royal Society B* 279:2761-2768.

MY CONTRIBUTIONS Proposed the central ideas; designed the project; developed individual analyses, particularly the simulation model framework; ran all final analyses; managed the overall project; co-wrote the manuscript; funded the project; first author.

COAUTHOR CONTRIBUTIONS

Michael Nelson *Massey University, NZ*	Intern (supervised); implemented the first version of the simulation model under direct supervision.
Meryanne K. Tumonggor *University of Arizona, USA*	Graduate student (supervised); generated mitochondrial DNA sequences.
François-Xavier Ricaut *University of Toulouse (Paul Sabatier), France*	Genetics collaborator; provided insight into the Malagasy component of the demographic model.
Herawati Sudoyo *Eijkman Institute for Molecular Biology, Indonesia*	Medical collaborator; performed fieldwork in Indonesia and provided samples.

CHAPTER 7 Lansing, J.S., **M.P. Cox**, T.A. de Vet, S.S. Downey, B. Hallmark and H. Sudoyo. 2011. An ongoing Austronesian expansion in Island Southeast Asia. *Journal of Anthropological Archaeology* 30:262-272.

MY CONTRIBUTIONS Co-developed the central ideas; co-designed the project; ran individual analyses, particularly for sex-biased processes; co-managed the overall project; co-wrote the manuscript.

COAUTHOR CONTRIBUTIONS

J. Stephen Lansing *University of Arizona, USA*	Anthropology collaborator; provided insight regarding the cultural anthropology of the study region and co-developed the central ideas.
Thérèse A. de Vet *University of Arizona, USA*	Anthropology collaborator; provided insight on the social anthropology of the study region.
Sean Downey *University College London, UK*	Postdoctoral fellow; provided advice on linguistic data.

Brian Hallmark *University of Arizona, USA*	Graduate student; performed some of the statistical tests.
Herawati Sudoyo *Eijkman Institute for Molecular Biology, Indonesia*	Medical collaborator; performed fieldwork in Indonesia and provided samples.

CHAPTER 8 **Cox, M.P.**, T.M. Karafet, J.S. Lansing, H. Sudoyo, and M.F. Hammer. 2010. Autosomal and X-linked single nucleotide polymorphisms reveal a steep Asian-Melanesian ancestry cline in eastern Indonesia and a sex bias in admixture rates. *Proceedings of the Royal Society B* 277:1589-1596.

MY CONTRIBUTIONS Proposed the central idea; designed the project; developed, implemented and ran all analyses; managed the overall project; wrote the manuscript; first author.

COAUTHOR CONTRIBUTIONS

Tatiana M. Karafet *University of Arizona, USA*	Laboratory manager; prepared DNA samples for commercial marker screening.
J. Stephen Lansing *Santa Fe Institute, USA*	Group head; provided insight regarding the cultural anthropology of the study region and funded the anthropological field program.
Herawati Sudoyo *Eijkman Institute for Molecular Biology, Indonesia*	Medical collaborator; performed fieldwork in Indonesia and provided samples.
Michael F. Hammer *University of Arizona, USA*	Group head; provided samples, funded commercial marker screening and funded staff, including my postdoctoral fellowship.

CHAPTER 9 **Cox, M.P.** 2008. Accuracy of molecular dating with the rho statistic: Deviations from coalescent expectations under a range of demographic models. *Human Biology* 80:335-357.

MY CONTRIBUTIONS Performed the entire work.

CHAPTER 10 **Cox, M.P.**, F.L. Mendez, T.M. Karafet, M. Metni Pilkington, S.B. Kingan, G. Destro-Bisol, B.I. Strassmann, and M.F.

Hammer. 2008. Testing for archaic hominin admixture on the X-chromosome: Model likelihoods for the modern human *RRM2P4* region from summaries of genealogical topology under the structured coalescent. *Genetics* 178:427-437.

MY CONTRIBUTIONS Developed the central ideas; designed the project; performed most of the lab work; developed the summary statistics; developed, implemented and ran all analyses; managed the overall project; wrote the manuscript; first author.

COAUTHOR CONTRIBUTIONS

Fernando L. Mendez *University of Arizona, USA*	Graduate student; provided advice on the mathematical framework described in the appendix.
Tatiana M. Karafet *University of Arizona, USA*	Laboratory manager; extracted DNA from samples and generated some of the *RRM2P4* sequences.
Maya Metni Pilkington *University of Arizona, USA*	Graduate student; generated some of the *RRM2P4* sequences.
Sarah B. Kingan *University of Arizona, USA*	Graduate student; generated some of the *RRM2P4* sequences.
Giovanni Destro-Bisol *University of Rome "La Sapienza", Italy*	Genetics collaborator; provided samples.
Beverly I. Strassmann *University of Michigan, USA*	Genetics collaborator; provided samples.
Michael F. Hammer *University of Arizona, USA*	Group head; first identified the *RRM2P4* locus and funded staff, including my postdoctoral fellowship.

CHAPTER 11 Lansing, J.S., J.C. Watkins, B. Hallmark, **M.P. Cox**, T.M. Karafet, H. Sudoyo, and M.F. Hammer. 2008. Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations. *Proceedings of the National Academy of Sciences USA* 105:11645-11650.

MY CONTRIBUTIONS Co-developed central ideas; co-designed the project; developed and implemented individual analyses; co-wrote the manuscript.

COAUTHOR CONTRIBUTIONS

J. Stephen Lansing *Santa Fe Institute, USA*	Group head; provided insight regarding the cultural anthropology of the study region and co-developed the central ideas.
Joseph C. Watkins *University of Arizona, USA*	Mathematics collaborator; provided advice on mathematical testing.
Brian Hallmark *University of Arizona, USA*	Graduate student; performed many of the statistical tests and implemented the final simulation models.
Tatiana M. Karafet *University of Arizona, USA*	Laboratory manager; extracted DNA from samples and screened Y chromosome markers.
Herawati Sudoyo *Eijkman Institute for Molecular Biology, Indonesia*	Medical collaborator; performed fieldwork in Indonesia and provided samples.
Michael F. Hammer *University of Arizona, USA*	Group head; funded staff, including my postdoctoral fellowship.

CHAPTER 12 Lansing, J.S., **M.P. Cox**, S.S. Downey, B.M. Gabler, B. Hallmark, T.M. Karafet, P. Norquest, J.W. Schoenfelder, H. Sudoyo, J.C. Watkins, and M.F. Hammer. 2007. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proceedings of the National Academy of Sciences USA* 104:16022-16026.

MY CONTRIBUTIONS Co-developed central ideas; co-designed the project; developed, implemented and ran individual analyses, particularly the Bayesian statistical inference components; assisted in managing the project; co-wrote the manuscript.

COAUTHOR CONTRIBUTIONS

J. Stephen Lansing *Santa Fe Institute, USA*	Group head; provided insight regarding the cultural anthropology of the study region and co-developed the central ideas.
Sean S. Downey *University of Arizona, USA*	Graduate student; provided advice on the linguistic data.
Brandon M. Gabler *University of Arizona, USA*	Graduate student; provided advice on the archeological data.
Brian Hallmark *University of Arizona, USA*	Graduate student; performed some of the statistical tests.

Tatiana M. Karafet *University of Arizona, USA*	Laboratory manager; extracted DNA from samples and screened Y chromosome markers.
Peter Norquest *University of Arizona, USA*	Graduate student; analyzed the linguistic data.
John W. Schoenfelder *University of Arizona, USA*	Graduate student; provided advice on the archeological data.
Herawati Sudoyo *Eijkman Institute for Molecular Biology, Indonesia*	Medical collaborator; performed fieldwork in Indonesia and provided samples.
Joseph C. Watkins *University of Arizona, USA*	Collaborator; provided advice on mathematical testing.
Michael F. Hammer *University of Arizona, USA*	Group head; funded staff, including my postdoctoral fellowship.