Complete mtDNA genomes of Filipino ethnolinguistic groups:

A melting pot of recent and ancient lineages in the Asia-Pacific region

Frederick Delfin[1,2*], Albert Min-Shan Ko[1], Mingkun Li[1], Ellen D. Gunnarsdóttir[1,3], Kristina A. Tabbada[2,4], Jazelyn M. Salvador[2], Gayvelline C. Calacal[2], Minerva S. Sagum[2], Francisco A. Datar[5], Sabino G. Padilla[6,7], Maria Corazon A. De Ungria[2], Mark Stoneking[1*]

[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz, D04103, Leipzig, Germany

[2]DNA Analysis Laboratory, Natural Sciences Research Institute, University of the Philippines, Diliman, 1101, Quezon City, Philippines

[3]deCODE genetics, Sturlugata 8, 101 Reykjavik, Iceland

[4] The Babraham Institute, Babraham Research Campus Cambridge, CB22 3AT United Kingdom

[5]Department of Anthropology, College of Social Sciences and Philosophy, Faculty Center, 3055, University of the Philippines, Diliman, 1101, Quezon City, Philippines

[6]Department of Behavioral Sciences, College of Arts and Sciences, University of the Philippines, Manila, Ermita, 1000, Manila, Philippines

[7]AnthroWatch.org, 46 Unit 326, Eagle Court Condominium, 26 Matalino St. Barangay Central, 1100, Quezon City, Philippines

*Corresponding authors:

Mark Stoneking

Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology

Deutscher Platz 6, D04103 Leipzig, Germany

Phone: +49-341-3550-502; Secretary: +49-341-3550-504; FAX: +49-341-3550-555;

Email: stoneking@eva.mpg.de

Frederick Delfin

Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology

Deutscher Platz 6, D04103 Leipzig, Germany

Phone: (63-2) 925-2965

Email: forengen@googlemail.com, frederick_delfin@eva.mpg.de

SUPPLEMENTARY TEXT

FE group populations

This study included 230 individuals representing 10 FE groups who were previously included and described in a FE NRY study[1]. A more recent population sample collection, facilitated by Francisco A. Datar in 2007, provided an additional 30 individuals from the Abaknon FE group. A data set of 97 complete mtDNA genomes from three FE groups from Mindanao[2] was included to bring the FE group data set to a total of 357 complete mtDNA genomes representing 14 FE groups. Supplementary Figure 1 and Supplementary Table 1 give demographic information for these groups.

FE group DNA samples

The 230 DNA samples that represent 10 FE groups were archived DNA samples from buccal cells-saliva and/or blood (Supplementary Table 2) collected from 1997-2004. These DNA samples were available as DNA in solution and/or DNA immobilized on FTA™ paper punches (FTA™ Gene Guard system, Whatman Inc., Springfield Mill, Maidstone, Kent, UK) . These samples were previously used in autosomal SNPs[3], Y-chromosome[1] and autosomal microsatellite analyses (F Delfin, MCA DeUngria and M Stoneking, unpublished data). DNA samples were quantified using a human DNA-specific quantitative PCR method[4]. When necessary, samples were reconstituted using 100 ul TE buffer (1mM Tris, 0.01mM EDTA), with 56^{o}C overnight incubation. DNA was eluted from FTA™ punches by at least 48 hours of incubation (with agitation) in TE buffer at 25^{o}C. DNA from Abaknon FE group saliva samples (a more recent collection of 30 samples) was prepared using a salting-out method[4]. Approximately 100 ul of

DNA solution (available sample, reconstituted and/or eluted from FTA™ punches) were carried over to DNA library preparation.

DNA sequencing and sequence assembly

A total of four multiplexed DNA libraries were prepared for 260 samples using a multiple-sample indexing method developed for the Illumina Genome Analyzer platform[5]. This procedure was coupled with a target enrichment method designed to specifically capture human mtDNA from complex DNA mixtures[6]. Each DNA library was sequenced on one Illumina flow cell lane using 76 cycles of single-read multiplex sequencing on an Illumina Genome Analyzer IIx machine (Illumina Inc., San Diego CA). Post-sequencing processing was performed using Illumina's Real-Time Analysis and Bustard base calling software followed by refined base-calling analyses using the Improved Base Identification System (Ibis)[7]. Sequencing reads for each sample were mapped to the revised Cambridge Reference Sequence (rCRS) of the human mitochondrial genome (GenBank: NC012920.1)[8] using the iterative mapping assembly program MIA [9]. MIA was used as part of a MPI-EVA sequence assembly-analyses pipeline recently developed for detecting mtDNA heteroplasmy[10] and for detecting low-level mutations in next-generation sequence data[11].

Data quality

Due to their archived nature and extensive use in previous studies, the available FE group samples yielded varying degrees of sequence coverage (Supplementary Figure 2; Supplementary Table 2). Highest average coverage was for the Ivatans (>2,000x coverage). Samples from the Aetas of Bataan, the Agtas and the Aetas of Zambales showed the lowest average coverage of

100x, 30x and 10x, respectively. This likely reflects the limited amount of remaining DNA that could be eluted from the FTA paper punches. To ensure quality data analyses, all sequences used in the study had less than 1% missing data [ambiguous bases (N) plus gaps < 165). Haplogroup assignment (via mtDNAble and/or Haplogrep) using complete sequence data served as another gauge for data quality (Supplementary Table 4).

Multiple sequence alignment (MSA) and haplogroup assignment of FE group mtDNA

The global iterative refinement method of MAFFT v6.833b[12] was used in multiple sequence alignment of the FE group data set of 357 complete mtDNA genomes (260 from this study and 97 from Gunnarsdóttir *et al*. 2011a), including the RSRS[13] (http://www.mtdnacommunity.org). Aligned complete mtDNA sequences were assigned haplogroups using a custom Perl script (available from the DREEP website: http://dmcrop.sourceforge.net) that combines the MUSCLE v3.8[14,15] sequence alignment software with mtDNAble v1.1.0.0[13] (http://www.mtdnacommunity.org), a program for haplogroup assignment and phylogeny based quality checking. Haplogrep[16] (http://www.haplogrep.uibk.ac.at) with PhyloTree mtDNA tree build 13[17] (http://www.phylotree.org), was also used to check haplogroup assignment.

mtDNA diversity and genetic relationships of FE groups

The MSA data of the 357-mtDNA genome FE group data set was 'cleaned' [removal of nucleotide positions with missing data: ambiguous bases (N) and gaps] using BioEdit [18] and used in haplotype analyses. Haplogroup assignment output of mtDNAble was used in haplogroup analyses. Arlequin 3.5.1.3[19] was used for the following analyses: estimation of genetic diversity

indices [number of polymorphic sites, number of haplotypes and haplogroups, shared haplotypes, haplotype diversity (HtD), haplogroup diversity (HgD) and mean number of pairwise differences (MPD)]; Φ_{ST} genetic distance estimation; Mantel tests and Analysis Of Molecular Variance (AMOVA). Diversity indices were also checked using DnaSP v5[20]. Statistica v8[21] was used for the following analyses: Multidimensional scaling (MDS) using Φ_{ST} genetic distance estimates; Correspondence analysis (CA) using haplogroup counts and Mann-Whitney U tests. Matrix correlation analyses via Mantel tests for genetics-geography and genetics-language association made use of the following components: Φ_{ST} genetic distance matrix generated by Arlequin 3.5.1.3[19]; geographic distance matrix generated by Geographic Distance Matrix Generator v1.2.3[22] (http://biodiversityinformatics.amnh.org/open_source/gdmg) and a language distance matrix kindly provided by Russell Gray[23].

FE group mtDNA and Filipino RC group mtDNA data

FE group mtDNA data was compared to a Filipino data set of unknown FE group affiliation from[24] composed of the following: mtDNA hypervariable segment 1 (HVS-1) data of 100 Filipinos sampled from the main Philippine island groups of Luzon (from the National Capital Region), Visayas (from Cebu City) and Mindanao (Zamboanga City); mtDNA HVS-1 data of 320 Filipinos sampled in Taiwan, and 30 complete mtDNA genomes of Filipinos also sampled in Taiwan. The local iterative refinement method of MAFFT[12] was used in the MSA of this data set and the 360-mtDNA genome FE group data, including the RSRS. Two data sets were extracted from this MSA using BioEdit[18]. One data set was composed of a total of 780 (260: this study; 97: Gunnarsdóttir *et al*. 2011a; 420: Tabbada *et al*. 2010) HVS-1 [nucleotide (nt) 16037 to nt16369] sequences and the other data set was composed of 390 (260: this study;

97: Gunnarsdóttir *et al*. 2011a; 30: Tabbada *et al*. 2010) complete mtDNA genomes. BioEdit was used to clean the data. Phi_{ST} (Φ_{ST}) molecular genetic distance was computed for each data set using Arlequin 3.5.1.3[19]. Phi_{ST} distances were used in MDS analysis using Statistica v8[21]. Haplogroup counts for this data set were used in CA using Statistica v8[21].

FE group mtDNA and a Reference data set

For data comparison, a Reference data set composed of 1,759 complete mtDNA genomes representing 43 Asian and Pacific groups was retrieved from the data archive of PhyloTree.org (http://www.phylotree.org/mtDNA_seqs.htm). The Asian and Pacific groups considered were the following: 28 Indian groups[25-27], including Andaman Islanders[28]; Vietnam[24,29,30]; Besemah and Semende groups of Sumatra, Indonesia[31]; a designated 'general' Indonesian group of 41 complete mtDNA genomes from different studies[30,32]; Bidayuh, Jehai, Seletar and Temuan groups of Malaysia[33]; Han Chinese[34]; Japan [35]; Taiwan[24,30,32,36]; Australia[37,38]; Near Oceania[32,39-42] and Remote Oceania[32,42] (Supplementary Table 3). MSA of this data set with the FE group data, including the RSRS was done using the global iterative refinement method of MAFFT v6.833b. Aligned complete genomes were run through the same Perl script-MUSCLE v3.8-mtDNAble pipeline for haplogroup assignment. MSA data was cleaned using BioEdit for haplotype analysis. Genetic distance (Φ_{ST}) was estimated using Arlequin 3.5.1.3. MDS using Φ_{ST} and CA using haplogroup counts were performed using Statistica v8. Phylogeographic analysis of specific haplogroups was performed by comparison of diversity indices (described in the preceding section) and coalescent times (described in the following section).

In MDS analyses, five dimensions were required to obtain an acceptable degree of correspondence (Stress=<0.15) between the distances among MDS mapped points and the Φ_{ST}

genetic distance matrix generated for the FE group and Reference data sets. For haplogroup data, CA initially included 18 Reference data set groups that share haplogroups with FE groups (Supplementary Table 7). Ten runs of CA were performed, with the removal of outliers in each subsequent run (Supplementary Figure 7) until the last CA, where only five Reference data set groups were observed to be the closest to FE groups (Supplementary Table 8; Figure 4) because of shared haplogroups B4a1a (and other B4 and B5 sublineages), E1a1a (and other E sublineages), M7c3c, M52'58, M52a and Y2a1.

Bayesian evolutionary analyses

The program package BEAST v1.7.2[43] (http://beast.bio.ed.ac.uk/) was used for all Bayesian Markov Chain Monte Carlo (MCMC) analyses. The mtDNA coding region (nt577-nt16023) was extracted from the MSA of all included complete mtDNA genomes (FE groups, Reference data set and RSRS) and cleaned using BioEdit. DnaSP v5 was used to convert coding region FASTA data files into NEXUS data files for subsequent use in the BEAST utility program BEAUti v1.7.2 to generate the BEAST input XML data file. The following BEAST parameters were used: TN93 substitution model[44]; Gamma + Invariant sites site heterogeneity model; strict molecular clock model with a rate of 1.691×10^{-8} substitutions/site/year[45]; coalescent tree prior with a piecewise linear model[46]. Each BEAST run involved a 40 million MCMC chain length with 10% burn-in and parameters logged every 4,000 steps. For each type of analysis [Bayesian skyline plot (BSP) generation and coalescent time estimation] at least two separate BEAST runs were performed, each with different starting random number seeds. Tracer v1.5 (http://tree.bio.ed.ac.uk/software/tracer) was used to evaluate each BEAST run for effective MCMC chain mixing and convergence. LogCombiner v1.7.2 (included in the BEAST package)

was used to combine MCMC log files of independent BEAST runs into one results file and BEAST trees into one tree results file. TreeAnnotator v1.7.2 (included in BEAST package) was used to summarize the combined trees into a single target tree [maximum clade credibility tree (MCC tree)]. FigTree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree) was used to display the target trees. To check for consistency in the clustering of sequences on the MCC trees, phylogenetic analyses were performed using DNA Alignment version 1.3.1.1 (http://fluxus-engineering.com) to generate input files for a Median-joining (MJ) network analyses[47] using Network version 4.6.1.0 and Network Publisher version 1.3.0.0 (http://fluxus-engineering.com). The described BEAST parameters and implementation were used to estimate coalescent times of haplogroups found in FE groups and shared with groups in the Reference data set. For coalescent time estimation, the coalescent tree was calibrated using the RSRS as tree root with a normal prior distribution for root age with a mean of about 177,000 years and standard deviation of 11,000 years[13]. The same Bayesian MCMC framework in BEAST was used to generate Bayesian Skyline Plots (BSPs) for each of the FE groups to show the fluctuations in the effective population sizes (N_e) through time. BSP construction and N_e estimation was performed in two ways: 1) using FE group only and 2) FE group including the RSRS as root. N_e values were extrapolated using a generation time of 28 years based on various studies of mother-daughter generation intervals[48-51].

For another aspect of population demography, the same Bayesian MCMC framework was used to estimate the fluctuations in N_e of FE groups through time, summarized in BSPs. BSPs and N_e estimation was initially performed using FE group data only. However, because of small sample size the AetaZ (which have effectively only one haplotype) and the Agta data did not generate typical BSPs (Supplementary Figure 4). Analyses were repeated for all FE groups, this

time including the RSRS to root the tree. FE group BSPs (excluding AetaZ and Agta) and N_e estimates from these analyses (Supplementary Figure 5; Supplementary Table 5, 'B' columns) showed a shift towards greater N_e values, but did not drastically differ from estimates that did not include the RSRS (Supplementary Figure 4; Supplementary Table 5, 'A' columns) except that by including the RSRS a current N_e estimate was available for AetaZ. However, typical BSPs still could not be generated for the AetaZ and Agta likely because of their small sample size.. For the other FE groups, BSPs (Supplementary Figure 5) show that from more than 20kya to about 10kya, FE groups experienced different trends in N_e change. The Ivatan, Abaknon, Surigaonon, Mamanwa, Manobo and Maranao show a marked increasing trend in N_e change. The Ifugao shows a slight increasing trend, while the Kankanaey and Ibaloi have an almost unnoticeable increasing trend in N_e change. The Kalangoya showed no N_e fluctuation. The Bugkalot and AetaB show a decreasing trend in N_e change. Generally, a decline in N_e is observed between 10 kya to 5 kya and a drop in N_e around 2 kya to the present time. Exceptions to these are the Bugkalot and AetaB whose N_e was in decline even prior to 10kya; the Surigaonon and Maranao who show an increasing N_e trend up to the present time; and the Kalangoya which show an increase in N_e around 2kya. For estimates of current N_e (Supplementary Table 5), the Bugkalot, AetaZ and the Agta have the smallest N_e ($\leq$ 1,000), while the Maranao has the largest N_e (>20,000). In between, current N_es ranged from as low as 1,000 to >4,000. While not completely up to date, FE group census sizes[52] are shown in Supplementary Table 5 ('C' column) to show that the N_e estimates for females are but a fraction of the actual group sizes. The differences in the trends of N_e change and the current N_e estimates may suggest different demographic histories for the FE groups included in this study

Analyses of specific mtDNA lineages

Haplogroups are specific lineages whose distribution are often discussed and related to population history and demographic events. The 54 mtDNA haplogroups detected among the FE groups belong to the following lineages: B4, B5, D, E, F, M, N, P, Q, R and Y2 (Table 2). Of the 54 haplogroups, seven haplogroups [B4a1a, B4b1a2, B5b1c, E1a1a, E1a1a1 (which is effectively E1a1a), M7c3c, Y2a1] have appreciable frequencies (≥5%) and have different geographical distributions (Figure 4). The most common haplogroups, B4a1a and M7c3c are most frequent in northern FE groups (excluding the Ivatan and two Aeta groups) and are less frequent in southern groups and present in only one central group, the Abaknon. The northern Bugkalot group does not have B4a1a and has M7c3c in low frequency. B4b1a2 is present in three northern groups, one central group and three southern groups with the highest frequency in the AetaB group. B5b1c is mainly found in northern FE groups except the two Aeta groups. E1a1a and its sublineage E1a1a1 are both present in the Surigaonon, Mamanwa and Manobo southern FE groups, while the Maranao only has the E1a1a1 sublineage and the Abaknon only has E1a1a. For northern groups, E1a1a is present only in the Ivatan and Ibaloi while E1a1a1 is present only in the Kankanaey and Bugkalot. Y2a1 is present in only four northern and three southern groups. Haplogroup B4c1b2a2 is present in the Ivatan, the Abaknon and in southern groups except the Mamanwa. Haplogroup F1a4a is present in three northern groups, the Abaknon and the Manobo. Other haplogroups of interest are M52’58 and its sublineage M52a which are exclusive to AetaZ and Agta FEN groups. Previously reported as N^{*2}, N11b is still only found in the Mamanwa, while haplogroup P is found only in the AetaB FEN group. The patchy distribution of these haplogroups further demonstrates the diversity of the different FE groups in this study.

Specific haplogroups that were observed in CA (Figure 4) to drive associations between FE groups and Reference data set groups as well as those found interesting in FE groups, were subjected to Bayesian-MCMC evolutionary analyses using BEAST v1.7.2[43]. The mtDNA coding region (nt577-nt16023) was used for each sample and the RSRS [13] was used to root coalescent trees. Results for these analyses are MCC trees where, common to all figures (Supplementary Figures 8-15), the tree topology and coalescent time estimates are the results of BEAST analyses, the tree tips are labelled with sample names, and the haplogroup labels reflect haplogroup assignment with mtDNAble [13]. MCC trees showed consistency with MJ networks (data not shown).

REFERENCES

1. Delfin F, Salvador JM, Calacal GC *et al*: The Y-chromosome landscape of the Philippines: extensive heterogeneity and varying genetic affinities of Negrito and non-Negrito groups. *Eur J Hum Genet* 2011; **19:** 224-230.

2. Gunnarsdóttir ED, Li M, Bauchet M, Finstermeier K, Stoneking M: High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Res* 2011a; **21:** 1-11.

3. Abdulla MA, Ahmed I, Assawamakin A *et al*: Mapping Human Genetic Diversity in Asia. *Science* 2009; **326:** 1541-1545.

4. Quinque D, Kittler R, Kayser M, Stoneking M, Nasidze I: Evaluation of saliva as a source of human DNA for population and association studies. *Anal Biochem* 2006; **353:** 272-277.

5. Meyer M, Kircher M: Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harbor Protocols* 2010; **2010:** pdb.prot5448.

6. Maricic T, Whitten M, Pääbo S: Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PLoS ONE* 2010; **5:** e14004.

7. Kircher M, Stenzel U, Kelso J: Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* 2009; **10:** R83.

8. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N: Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 1999; **23:** 147-147.

9. Briggs AW, Good JM, Green RE *et al*: Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes. *Science* 2009; **325:** 318-321.

10. Li M, Schönberg A, Schaefer M, Schroeder R, Nasidze I, Stoneking M: Detecting Heteroplasmy from High-Throughput Sequencing of Complete Human Mitochondrial DNA Genomes. *Am J Hum Genet* 2010; **87:** 237-249.

11. Li M, Stoneking M: A new approach for detecting low-level mutations in next-generation sequence data. *Genome Biol* 2012; **13:** R34.

12. Katoh K, Kuma K-i, Toh H, Miyata T: MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 2005; **33:** 511-518.

13. Behar Doron M, van Oven M, Rosset S *et al*: A Copernican Reassessment of the Human Mitochondrial DNA Tree from its Root. *Am J Hum Genet* 2012; **90:** 675-684.

14. Edgar R: MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004b; **5:** 113.

15. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004a; **32:** 1792-1797.

16. Kloss-Brandstätter A, Pacher D, Schönherr S *et al*: HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Human Mutat* 2010; **32:** 25-32.

17. van Oven M, Kayser M: Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutat* 2009; **30:** E386-E394.

18. Hall TA: BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* 1999; **41:** 95-98.

19. Excoffier L, Lischer HEL: Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 2010; **10:** 564-567.

20. Librado P, Rozas J: DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 2009; **25:** 1451-1452.

21. StatSoft I. STATISTICA (data analysis software system), 2007.

22. Ersts PJ. Geographic Distance Matrix Generator v1.2.3. New York USA, Center for Biodiversity and Conservation, American Museum of Natural History, 2006.

23. Gray RD, Drummond AJ, Greenhill SJ: Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. *Science* 2009; **323:** 479-483.

24. Tabbada KA, Trejaut J, Loo J-H *et al*: Philippine Mitochondrial DNA Diversity: A Populated Viaduct between Taiwan and Indonesia? *Mol Biol Evol* 2010; **27:** 21-31.

25. Chandrasekar A, Kumar S, Sreenath J *et al*: Updating Phylogeny of Mitochondrial DNA Macrohaplogroup M in India: Dispersal of Modern Human in South Asian Corridor. *PLoS ONE* 2009; **4:** e7447.

26. Eaaswarkhanth M, Haque I, Ravesh Z *et al*: Traces of sub-Saharan and Middle Eastern lineages in Indian Muslim populations. *Eur J Hum Genet* 2009; **18:** 354-363.

27. Palanichamy Mg, Sun C, Agrawal S *et al*: Phylogeny of Mitochondrial DNA Macrohaplogroup N in India, Based on Complete Sequencing: Implications for the Peopling of South Asia. *Am J Hum Genet* 2004; **75:** 966-978.

28. Thangaraj K, Chaubey G, Kivisild T *et al*: Reconstructing the Origin of Andaman Islanders. *Science* 2005; **308:** 996.

29. Hill C, Soares P, Mormina M *et al*: Phylogeography and Ethnogenesis of Aboriginal Southeast Asians. *Mol Biol Evol* 2006; **23:** 2480-2491.

30. Soares P, Trejaut JA, Loo J-H *et al*: Climate Change and Postglacial Human Dispersals in Southeast Asia. *Mol Biol Evol* 2008; **25:** 1209-1218.

31. Gunnarsdóttir ED, Nandineni MR, Li M *et al*: Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. *Nat Commun* 2011b; **2:** 228.

32. Soares P, Rito T, Trejaut J *et al*: Ancient Voyaging and Polynesian Origins. *Am J Hum Genet* 2011; **88:** 239-247.

33. Jinam TA, Hong L-C, Phipps ME *et al*: Evolutionary history of continental South East Asians: â€œearly trainâ€ hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. *Mol Biol Evol* 2012.

34. 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* 2003; **73:** 671-676.

35. Tanaka M, Cabrera VM, GonzÃ¡lez AM *et al*: Mitochondrial Genome Variation in Eastern Asia and the Peopling of Japan. *Genome Res* 2004; **14:** 1832-1850.

36. Trejaut JA, Kivisild T, Loo JH *et al*: Traces of Archaic Mitochondrial Lineages Persist in Austronesian-Speaking Formosan Populations. *PLoS Biol* 2005; **3:** e247.

37. van Holst Pellekaan SM, Ingman M, Roberts-Thomson J, Harding RM: Mitochondrial genomics identifies major haplogroups in Aboriginal Australians. *Am J Phys Anthropol* 2006; **131:** 282-294.

38. Hudjashov G, Kivisild T, Underhill PA *et al*: Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *PNAS* 2007; **104:** 8726-8730.

39. Friedlaender JS, Friedlaender FoR, Hodgson JA *et al*: Melanesian mtDNA Complexity. *PLoS ONE* 2007; **2:** e248.

40. Friedlaender J, Schurr T, Gentz F *et al*: Expanding Southwest Pacific Mitochondrial Haplogroups P and Q. *Mol Biol Evol* 2005; **22:** 1506-1517.

41. Merriwether DA, Hodgson JA, Friedlaender FoR *et al*: Ancient mitochondrial M haplogroups identified in the Southwest Pacific. *PNAS* 2005; **102:** 13034-13039.

42. Pierson MJ, Martinez-Arias R, Holland BR, Gemmell NJ, Hurles ME, Penny D: Deciphering Past Human Population Movements in Oceania: Provably Optimal Trees of 127 mtDNA Genomes. *Mol Biol Evol* 2006; **23:** 1966-1975.

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

44. Tamura K, Nei M: Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 1993; **10:** 512-526.

45. Atkinson QD, Gray RD, Drummond AJ: mtDNA Variation Predicts Population Size in Humans and Reveals a Major Southern Asian Chapter in Human Prehistory. *Mol Biol Evol* 2008; **25:** 468-474.

46. Drummond AJ, Rambaut A, Shapiro B, Pybus OG: Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Mol Biol Evol* 2005; **22:** 1185-1192.

47. Bandelt H-J, Forster P, Röhl A: Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 1999; **16:** 37-48.

48. Storz JF, Ramakrishnan U, Alberts SC: Determinants of Effective Population Size for Loci With Different Modes of Inheritance. *J Hered* 2001; **92:** 497-502.

49. Forster L, Forster P, Lutz-Bonengel S, Willkomm H, Brinkmann B: Natural radioactivity and human mitochondrial DNA mutations. *PNAS* 2002; **99:** 13950-13954.

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

51. Matsumura S, Forster P: Generation time and effective population size in Polar Eskimos. *Proc R Soc Lond B Biol Sci* 2008; **275:** 1501-1508.

52. Lewis MP (ed): *Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/*. 2009.