Recombination Gives a New Insight in the Effective Population Size and the History of the Old World Human Populations

Marta Melé,[1] Asif Javed,[2] Marc Pybus,[1] Pierre Zalloua,[3] Marc Haber,[3] David Comas,[1] Mihai G. Netea,[4] Oleg Balanovsky,[5,6] Elena Balanovska,[5] Li Jin,[7] Yajun Yang,[7] R. M. Pitchappan,[8,9] G. Arunkumar,[9] Laxmi Parida,[2] Francesc Calafell,[1] Jaume Bertranpetit,[1,*] and the Genographic Consortium†

[1]IBE, Institute of Evolutionary Biology (UPF-CSIC), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

[2]Computational Biology Center, IBM T J Watson Research, Yorktown Heights, NY, USA

[3]School of Medicine, Lebanese American University, Beirut, Lebanon

[4]Department of Medicine and Nijmegen Institute for Infection, Inflammation, and Immunity, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands

[5]Research Centre for Medical Genetics, Moscow, Russia

[6]Vavilov Institute for General Genetics, Moscow, Russia

[7]MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai, China

[8]Chettinad Academy of Research and Education, Chettinad Health City, Rajiv Gandhi Salai, Kelambakkam, Chennai, India

[9]School of Biological Sciences, Madurai Kamaraj University, Madurai, India

†Membership of the Genographic Consortium is provided in the Acknowledgments.

***Corresponding author:** E-mail: jaume.bertranpetit@upf.edu.

Associate editor: Rasmus Nielsen

Abstract

The information left by recombination in our genomes can be used to make inferences on our recent evolutionary history. Specifically, the number of past recombination events in a population sample is a function of its effective population size (N_e). We have applied a method, Identifying Recombination in Sequences (IRiS), to detect specific past recombination events in 30 Old World populations to infer their N_e. We have found that sub-Saharan African populations have an N_e that is approximately four times greater than those of non-African populations and that outside of Africa, South Asian populations had the largest N_e. We also observe that the patterns of recombinational diversity of these populations correlate with distance out of Africa if that distance is measured along a path crossing South Arabia. No such correlation is found through a Sinai route, suggesting that anatomically modern humans first left Africa through the Bab-el-Mandeb strait rather than through present Egypt.

Key words: recombination, effective population size, Out of Africa.

The estimation of effective population size (N_e) in human evolution has been a subject of intense research in the recent past. The seminal papers by Takahata (reviewed in Kim et al. (2010)) established the highly cited figure of 10,000 individuals for the past human evolutionary history, which has been lately revised using either measures of heterozygosity (Kim et al. 2010; Laval et al. 2010) or of linkage disequilibrium (LD) (McEvoy et al. 2011; Hayes et al. 2003; Tenesa et al. 2007). Gene diversity estimates tend to give effective population sizes higher than 10,000, whereas LD-based estimates give lower numbers. Gene diversity estimates would reflect average N_e for long periods of time, whereas LD depends to a greater extent on the N_e in more recent times.

We have developed a method called IRiS (Identifying Recombination in Sequences) to detect specific past recombination events from extant sequences (Melé et al. 2010) based on a combinatorial algorithm (Parida et al. 2008, 2009). The algorithm yields which sequences are descendants of ancient recombination events, which sequences carry the ancestral patterns that were involved in the recombination event, and where the breakpoint is located in the genome. Here, we use IRiS to extract the number of recombinations present in diverse populations of the Old World and use this to estimate their N_e. Several aspects of this approach are novel: the detection of recombinations is not based on LD and the time distribution of the reconstructed recombinations is known (Melé et al. 2010). In our previous study (Melé et al. 2010), we showed that recent recombinations are detected by IRiS with greater sensitivity: 90% of the events detected by IRiS occurred after the out of Africa migration. Therefore, recombinations can be

Letter

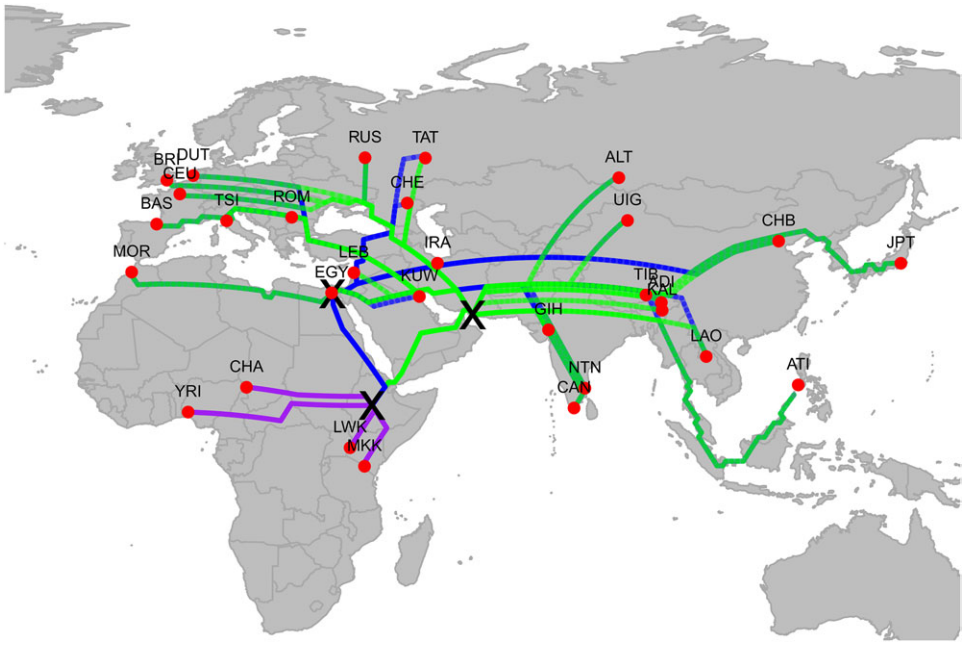

FIG. 1. Map of the Old World with the samples of the present study. The two tested routes out of Africa are marked: a route through the North of Egypt in blue and a South Arabian route in green. Wherever the two lines overlap, colors are added up. In purple, the putative path followed by sub-Saharan African populations. The X symbol marks a presumable origin (or a later stage) of AMH in Addis Ababa and the two locations in which AMH are forced to go through in their way out of Africa for each of the two possible routes (present Egypt or Iran).

used as recent genetic markers and they can potentially help to make inferences on the most recent events of human evolutionary history, such as the estimation of population-specific N_e. In fact, most of the reconstructed recombinations are population-specific (93.1%).

The dataset was taken from a large survey of 1250 SNPs or single-nucleotide polymorphisms (Javed A, Melé M, Marc Pybus, Zalloua P, Haber M, Comas D, Netea MG, Balanovsky O, Balanovska E, Jin L, Yang Y, Arunkumar G, Pitchappan R, Bertranpetit J, Calafell F, Parida L,The Genographic Consortium, unpublished data) (supplementary table S1, Supplementary Material online) belonging to five gene-free regions of the X chromosome spanning 2 Mb genotyped in 1240 males from 30 Old World populations (fig. 1, supplementary table S1, Supplementary Material online). High uniform SNP coverage was necessary to detect as many recombination events as possible and therefore a customized genotyping array was used. By choosing only male samples, we could overcome the uncertainty associated with phasing haplotypes. Finally, regions known to contain genes were avoided in order to eschew the possible confounding effects of natural selection. Further details about region selection and genotyping can be found in the supplementary text, Supplementary Material online.

We used the expression $\rho = 3N_e r$ (Hudson 1987) to infer N_e which is analogous to the $\theta = 4N_e\mu$ formula (Nei 1987) for recombination and the X chromosome, where rho (ρ) is the population recombination parameter and r is the recombination rate. In a similar way in which the number of segregating sites can be used to infer theta (θ) (Nei 1987), the number of recombinations can be used to calculate rho by means of the equation $\rho = 3N_e r$ (3 is for the X chromosome), where $\rho = \frac{R}{\sum_{i=1}^{n-1}\frac{1}{i}}$, where R is the number of recombinations inferred for each population and n is the number of sequences analyzed.

Extensive simulations were performed in order to assess whether the number of recombinations detected by IRiS could be used as a proxy to infer the effective population size both at populations under equilibrium and under different demographic models. In total, 14 demographic models were tested, three under equilibrium differing in their N_e, and 11 based on the calibrated demography published in Schaffner et al. (2005) in which simulations are performed such that they resemble empirical human data of three populations. Demographic models differed either in the N_e of the different populations or else in the time of the split between populations (see supplementary fig. S1 and supplementary materials and methods, including supplementary table S3, Supplementary Material online). Results show first, that under equilibrium, estimates of N_e obtained based on the number of recombinations inferred by IRiS correspond to the value of N_e set in the simulations (supplementary table S2, Supplementary Material online). Second, we show that under a human-like demography in which three different populations are modeled, differences in N_e between populations will be detected by IRiS as a proportional increase in the number of recombinations detected (supplementary table S3, Supplementary Material online). Third, it is shown that the time of the split between European and Asians or the time set for the out of Africa in the simulations does not affect our results. Thus, the method is robust for the proposed goals.

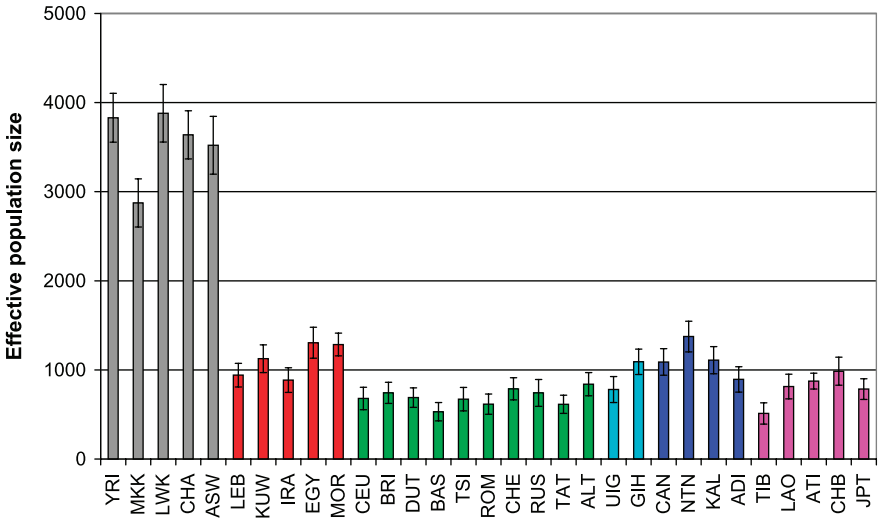

Fig. 2. Inferred effective population sizes from the number of recombinations detected and the corresponding sampling standard deviations calculated based on the 100 permuted datasets. Population abbreviations as in supplementary table S4, Supplementary Material online.

We then extracted the number of recombinations per population with the IRiS method run on 100 permuted datasets with 18 chromosomes per population. The average number of recombinations detected per population was multiplied by the corresponding sensitivity of the method (7.7%) assessed by simulations (see supplementary material, Supplementary Material online) at a recombination rate of 1.8×10^{-8}. Sensitivity was not affected by SNP diversity (Spearman's $r = -0.161$; $P = 0.111$) and therefore, the same value could be applied for all populations.

Estimates of N_e for each of the populations based on the 100 permutations and the corresponding standard deviations are given in supplementary table S4, Supplementary Material online and plotted in figure 2. As expected, results consistently show that Sub-Saharan Africans have much higher N_e than all other populations; values are roughly 4-fold larger or, in absolute terms, of ~4,000 for African populations and of ~1,000 for the rest. This result is in line with the low values obtained with LD-based estimates (Tenesa et al. 2007; Laval et al. 2010) and the ~2.5 times higher African effective sizes found from genetic diversity estimates (Laval et al. 2010) and from LD (Hayes et al. 2003; Tenesa et al. 2007). On the other hand, the estimation of N_e is subject to uncertainties that may carry into the inference of N_e as observed in the simulation analysis, but if significant differences in N_e exist between populations, they will be recovered by our method.

For the first time, we provide specific effective sizes for a wide range of Old World populations in relative and absolute values (supplementary table S4, Supplementary Material online) and a number of interesting patterns are revealed. The populations with the largest sizes other than Sub-Saharan Africans are North Africans (Moroccans and Egyptians) due to their known Sub-Saharan admixture (Krings et al. 1999; Bosch et al. 2001; Brakez et al. 2001). Outside of Africa, the largest N_e is found in South Asia; only recently, the high internal diversity of Indian populations is being appreciated (Xing et al. 2010). Europeans and East Asians have similar N_e. Tibetans and Basques showed the lowest values, a direct measure of small population size and isolation.

We further investigated the geographic variation of both SNPs and recombinations to understand the general pattern of genetic variation and population history. In order to compare patterns of diversity across populations, we used Nei's nucleotide diversity statistic (Nei 1987) to calculate diversity using either SNP allele frequencies (SNP diversity) or population frequencies of each recombination event (recombination diversity). We provide a geographic framework to these values by plotting them against the geographic distance of each population to Eastern Africa, the presumed place of origin of modern humans (Quintana-Murci et al. 1999; Tishkoff et al. 2009) if leaving Africa through the north of Egypt. Note that the relative differences in the distance between non sub-Saharan populations to a putative origin would not change even if an origin in more Central or Southern in Africa was considered (Liu et al. 2006; Betti et al. 2009; Henn et al. 2011) (see fig. 1). Geographic distances were calculated as in Prugnolle et al. (2005) in which the shortest landmass path between a specific point of origin and our 30 populations can be calculated based on graph theory.

As expected, SNP diversity was found to be highly correlated with geographical distance with East Africa (Spearman's $r = -0.596$; $P = 0.00064$) (supplementary fig. S2, Supplementary Material online) (Prugnolle et al. 2005; Ramachandran et al. 2005; Li et al. 2008), even if sub-Saharan African samples were removed ($r = -0.441$, $P = 0.024$). With recombination diversity, however, the correlation that is found is lower ($r = -0.391$; $P = 0.033$) and vanishes if African samples are removed ($r = -0.077$; $P = 0.71$) (fig. 3A). African populations show significantly higher recombination diversity than any other population (Mann–Whitney U test; $P = 0.0015$), in a proportion that goes to a 4- or 5-fold higher diversity than the mean for non-Africans; European populations

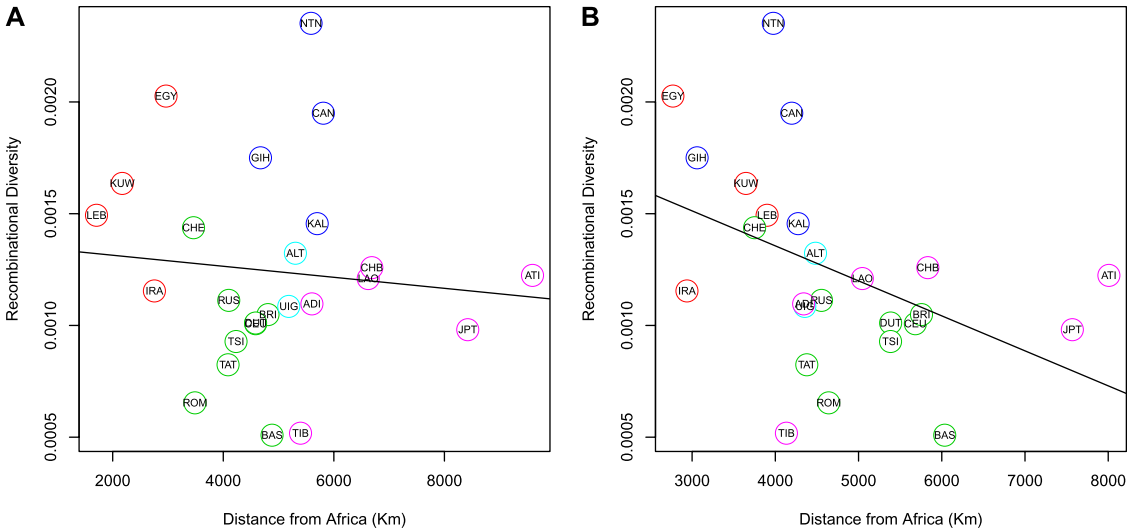

Fig. 3. Correlation between recombinational diversity and geographic distance from East Africa (in kilometers). (*A*) Following the North of Egypt route. (*B*) Following the South Arabian route. Populations are color-coded by continent as in figure 2.

show similar diversity values as East Asian populations, whereas Indian populations showed significantly more diversity than Europeans (Mann–Whitney *U* test; $P = 0.0055$) and East Asians (Mann–Whitney *U* test; $P = 0.011$). Finally, Moroccans appeared as clear outliers in both analyses, especially in the SNP diversity estimate (supplementary fig. S2, Supplementary Material online) due to their high levels of sub-Saharan African admixture. This compromised the geography-based analysis and, therefore, they were removed from the subsequent analyses.

The present results stress the wide differences between Sub-Saharan Africans and the rest of the Old World populations and point to a special role for South Asia (India) in the Out of Africa expansion of modern humans. Unfortunately, the density of sampled population within sub-Saharan Africa does not allow discussing the place of origin of anatomically modern humans. Recently, evidence has been published for the expansion of anatomically modern humans throughout Asia through a single and fast route, most likely via a Southern coastal path through India and onward into Southeast Asia and Australasia (Macaulay et al. 2005; Thangaraj et al. 2005; Mellars 2006; Armitage et al. 2011). Previous studies on microsatellites (Liu et al. 2006) and morphological variation (Betti et al. 2009) among others show the strong patterns left by the out of Africa settlement of Eurasia, a pattern that can be refined using the present recombination analysis.

Given that recombination diversity was not correlated with the distance from East Africa following a route through Northern Egypt, we assessed whether a route going through the Bab-el-Mandeb Strait (South Arabian route) as proposed by Mellars (2006) could better explain the relationship between recombination diversity and distance from East Africa. Interestingly, the correlation when considering only non sub-Saharan African samples became significant ($r = -0.569$, $P = 0.0029$) (fig. 3*B*) and, as expected, it became stronger when taking into account sub-Saharan Africans ($r = -0.723$, $P = 9 \times 10^{-6}$). In order to assess how much more robust the South Arabian route was compared with the Northern Egypt route, we performed a bootstrap analysis in which the populations used in the correlation were randomly sampled with replacement. Our results showed that 95.24% of the times, the South Arabian route had higher r^2 values than the North Egypt route.

Whereas a route through South Arabia explained better the patterns of recombination diversity, it did not for SNP diversity values (which have a slightly lower $r = -0.570$, $P = 0.0012$). These differences may be explained by two different factors. SNP diversity calculations may be affected by the ascertainment bias toward high frequency alleles in Europeans typical in HapMap2 SNPs (Clark et al. 2005), whereas recombination diversity estimates are more robust to this kind of bias (Melé et al. 2010). And the two approaches may reflect processes taking place in different time frames with the recombination-based analysis being more sensitive to more recent events.

Finally, given the higher N_e values for the Indian populations, it is tempting to speculate whether our results point toward India as having had a major role in a maturation phase prior to the expansion of modern humans to the whole of Eurasia as suggested by Atkinson et al. (2008) (see supplementary fig. S3, Supplementary Material online). The correlation between recombination diversity and geographic distance of Eurasians from South Asia ($r = -0.532$; $P = 0.0016$) (supplementary fig. S4, Supplementary Material online) is similar to the correlation between recombination diversity and geographic distance from Iran (South Arabian route) ($r = -0.569$, $P = 0.0029$) and therefore we cannot draw any conclusion at this end. Nonetheless, the higher N_e values present in South India compared with the East Asians do give support to the recent Southern coastal path (Macaulay et al. 2005; Thangaraj et al. 2005; Mellars 2006;

Armitage et al. 2011) proposed for the colonization of East Eurasia through India and onward into Southeast Asia and Australasia.

Supplementary Material

Supplementary tables S1–S7, figures S1–S4, materials and methods, and material are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to Mònica Vallés, UPF, for excellent technical support and to Sònia Sagristà for her help. Funding for this project was provided by National Geographic and IBM within the Genographic Project initiative and by the Spanish Ministry of Science and Innovation project BFU2010-19443 (subprogram BMC). M.M. was supported by grant AP2006-03268. Genotyping and bioinformatic services were provided respectively by CEGEN (Centro Nacional de Genotipado) and INB (National Bioinformatics Insitute), Spain. HapMap phase III population samples were obtained from the Coriell Cell Repository.

The Genographic Consortium includes: Syama Adhikarla (Madurai Kamaraj University, Madurai, Tamil Nadu, India), Christina J. Adler (University of Adelaide, South Australia, Australia), Danielle A. Badro (Lebanese American University, Chouran, Beirut, Lebanon), Andrew C. Clarke (University of Otago, Dunedin, New Zealand), Alan Cooper (University of Adelaide, South Australia, Australia), Clio S. I. Der Sarkissian (University of Adelaide, South Australia, Australia), Matthew C. Dulik (University of Pennsylvania, Philadelphia, PA, USA), Christoff J. Erasmus (National Health Laboratory Service, Johannesburg, South Africa), Jill B. Gaieski (University of Pennsylvania, Philadelphia, PA, USA), Wolfgang Haak (University of Adelaide, South Australia, Australia), Angela Hobbs (National Health Laboratory Service, Johannesburg, South Africa), Matthew E. Kaplan (University of Arizona, Tucson, AZ, USA), Shilin Li (Fudan University, Shanghai, China), Begoña Martínez-Cruz (Universitat Pompeu Fabra, Barcelona, Spain), Elizabeth A. Matisoo-Smith (University of Otago, Dunedin, New Zealand), Nirav C. Merchant (University of Arizona, Tucson, AZ, USA), R. John Mitchell (La Trobe University, Melbourne, Victoria, Australia), Amanda C. Owings (University of Pennsylvania, Philadelphia, PA, USA), Daniel E. Platt (IBM, Yorktown Heights, NY, USA), Lluis Quintana-Murci (Institut Pasteur, Paris, France), Colin Renfrew (University of Cambridge, Cambridge, UK), Daniela R. Lacerda (Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil), Ajay K. Royyuru (IBM, Yorktown Heights, NY, USA), Fabrício R. Santos (Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil), Theodore G. Schurr (University of Pennsylvania, Philadelphia, PA, USA), Himla Soodyall (National Health Laboratory Service, Johannesburg, South Africa), David F. Soria Hernanz (National Geographic Society, Washington, DC, USA), Pandikumar Swamikrishnan (IBM, Somers, NY, USA), Chris Tyler-Smith (The Wellcome Trust Sanger Institute, Hinxton, UK), Kavitha Valampuri John (Madurai Kamaraj University, Madurai, Tamil Nadu, India), Arun Varatharajan Santhakumari (Madurai Kamaraj University, Madurai, Tamil Nadu, India), Pedro Paulo Vieira (Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil), R. Spencer Wells (National Geographic Society, Washington, DC, USA), and Janet S. Ziegle (Applied Biosystems, Foster City, CA, USA).

References

Armitage SJ, Jasim SA, Marks AE, Parker AG, Usik VI, Uerpmann H-P. 2011. The Southern Route Out of Africa: Evidence for an Early Expansion of Modern Humans into Arabia. *Science* 331:453–456.

Atkinson QD, Meade A, Venditti C, Greenhill SJ, Pagel M. 2008. Languages evolve in punctuational bursts. *Science* 319: 588.

Betti L, Balloux F , Amos W, Hanihara T, Manica A. 2009. Distance from Africa, not climate, explains within-population phenotypic diversity in humans. *ProcBiol Sci*. 276:809–814.

Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA, Bertranpetit J. 2001. High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between Northwestern Africa and the Iberian Peninsula. *Am J Hum Genet*. 68:1019–1029.

Brakez Z, Bosch E, Izaabel H, Akhayat O, Comas D, Bertranpetit J, Calafell F. 2001. Human mitochondrial DNA sequence variation in the Moroccan population of the Souss area. *Ann Hum Biol*. 28:295–307.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res*. 15:1496–1502.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res*. 13:635–643.

Henn BM, Gignoux CR, Jobin M, et al. (13 co-authors). 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc Natl Acad Sci U S A*. 108:5154–5162.

Hudson RR. 1987. Estimating the recombination parameter of a finite population-model without selection. *Genet Res*. 50:245–250.

Kim HL, Igawa T, Kawashima A, Satta Y, Takahata N. 2010. Divergence, demography and gene loss along the human lineage. *Philos Trans R Soc Lond B Biol Sci*. 365:2451–2457.

Krings M, Salem A-eH, Bauer K, et al. (13 co-authors). 1999. mtDNA analysis of Nile river valley populations: a genetic corridor or a barrier to migration? *Am J Hum Genet*. 64:1166–1176.

Laval G, Patin E, Barreiro LB, Quintana-Murci L. 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS One* 5:e10284.

Li JZ, Absher DM, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

Liu H, Prugnolle F, Manica A, Balloux F. 2006. A geographically explicit genetic model of worldwide human-settlement history. *Am J Hum Genet*. 79:230–237.

Macaulay V, Hill C, Achilli A, et al. (21 co-authors). 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* 308:1034–1036.

McEvoy BP, Powell JE, Goddard ME, Visscher PM. 2011. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res*. 21:821–829.

Melé M, Javed A, Pybus M, Calafell F, Parida L, Bertranpetit J. Genographic Consortium Memebers. 2010. A new method to reconstruct recombination events at a genomic scale. *PLoS Comput Biol*. 6:e1001010.

Mellars P. 2006. Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science* 313:796–800.

Nei M. 1987. Molecular evolutionary genetics New York: Columbia University Press.

Parida L, Javed A, Mele M, Calafell F, Bertranpetit J. 2009. Minimizing recombinations in consensus networks for phylogeographic studies. *BMC Bioinformatics*. 10(Suppl 1):S72.

Parida L, Mele M, Calafell F, Bertranpetit J. 2008. Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns. *J Comput Biol*. 15:1133–1154.

Prugnolle F, Manica A, Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. *Curr Biol*. 15:R159–R160.

Quintana-Murci L, Semino O, Bandelt H-J, Passarino G, McElreavey K, Santachiara-Benerecetti AS. 1999. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat Genet*. 23:437–441.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A*. 102:15942–15947.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res*. 15:1576–1583.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res*. 17:520–526.

Thangaraj K, Chaubey G, Kivisild T, Reddy AG, Singh VK, Rasalkar AA, Singh L. 2005. Reconstructing the origin of Andaman Islanders. *Science* 308:996.

Tishkoff SA, Reed FA , Friedlaender FR. (25 co-authors). 2009. The genetic structure and history of Africans and African Americans. *Science* 324:1035–1044.

Xing J, Watkins WS, Hu Y, Huff C, Sabo A, Muzny D, Bamshad M, Gibbs R, Jorde L, Yu F. 2010. Genetic diversity in India and the inference of Eurasian population expansion. *Genome Biol*. 11:R113.