ORIGINAL PAPER

Inferring population structure and demographic history using Y-STR data from worldwide populations

Hongyang Xu · Chuan-Chao Wang · Rukesh Shrestha · Ling-Xiang Wang · Manfei Zhang · Yungang He · Judith R. Kidd · Kenneth K. Kidd · Li Jin · Hui Li

Abstract The Y chromosome is one of the best genetic materials to explore the evolutionary history of human populations. Global analyses of Y chromosomal short tandem repeats (STRs) data can reveal very interesting world population structures and histories. However, previous Y-STR works tended to focus on small geographical ranges or only included limited sample sizes. In this study, we have investigated population structure and demographic history using 17 Y chromosomal STRs data of 979 males from 44 worldwide populations. The largest genetic distances have been observed between pairs of African and non-African populations. American populations with the lowest genetic diversities also showed large genetic distances and coancestry coefficients with other populations, whereas Eurasian populations displayed close genetic affinities. African populations tend to have the oldest time to the most recent common ancestors (TMRCAs), the largest effective population sizes and the earliest expansion times, whereas the American, Siberian, Melanesian, and isolated Atayal populations have the most recent TMRCAs and expansion times, and the smallest effective population sizes. This clear geographic pattern is well consistent with serial founder model for the origin of populations outside Africa. The Y-STR dataset presented here provides the most detailed view of worldwide population structure and human male demographic history, and additionally will be of great benefit to future forensic applications and population genetic studies.

Keywords Y chromosome · STR · Population structure · Demographic history

Communicated by S. Hohmann.

H. Xu and C.-C. Wang contributed equally to this work.

H. Xu · Y. He · L. Jin
CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 200031 Shanghai, China

C.-C. Wang · R. Shrestha · L.-X. Wang · M. Zhang · L. Jin · H. Li
State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, 200433 Shanghai, China

J. R. Kidd · K. K. Kidd · H. Li (✉)
Department of Genetics, School of Medicine, Yale University, New Haven, CT 06520, USA
e-mail: lhca@fudan.edu.cn

Introduction

The paternally inherited Y chromosome has been widely used in anthropology and population genetics to explore population structure and demographic history of human populations (Jobling and Tyler-Smith 2003; Wang and Li 2013). Among the genetic markers on the Y chromosome, short tandem repeat (STR) polymorphic loci are extremely informative. The high mutation rates and lack of recombination of Y-STRs make them a kind of ideal tool in forensic identification and inference of recent population history (Jobling and Tyler-Smith 2003). Y-STR typing has been carried out in numerous populations, which has prompted the creation of many Y-STR databases (Roewer et al. 2001; Kayser et al. 2002; Lessig et al. 2003). However, most of those studies have focused on only small geographical ranges and put a greater emphasis on calculating allele or

haplotype frequencies, matching probabilities, and other forensic parameters. Thus, the evolution, migration, and genetic history of human populations at worldwide level still need information from Y-STR perspective. Previous genome-wide studies carried out at worldwide or continent-wide resolution have suggested the single origin of modern humans in sub-Saharan Africa and colonization of the other continents following serial founder events (Hellenthal et al. 2008; Li et al. 2008), while the population demographic changes accompanying these colonizing events are very important and interesting, and merit further investigation. Shi et al. (2010) investigated population demographic history using Y-STRs from samples of the Human Genome Diversity Project (HGDP). They found the oldest population time to the most recent common ancestors (TMRCAs) and expansion times together with the largest effective population sizes in Africa, and the youngest times and smallest effective population sizes in the Americas. However, some population sample sizes in their HGDP were as small as four, and such small sample sizes might be not sufficient to provide enough useful information about population history. Here, we presented a much more comprehensive analysis of 979 male samples collected from 44 worldwide populations. We aim at building a standard sample set for human paternal evolutionary genetic studies. In this study, we typed 17 Y-STRs for all the 979 samples using the Yfiler kit. This Y-STR haplotype database provides the most detailed view of worldwide population structure and human male demographic history, and additionally presents a suitable reference database for future use in numerous fields utilizing the Y chromosome to investigate male lineages, such as in forensic, genealogical, anthropological, and population genetic studies.

Materials and methods

Population samples and haplotyping

We typed 979 male individuals from a global sample of 44 populations. According to population ancestry and geographic locations, these 44 populations are categorized into six groups. The populations and sample sizes are as follows: Sub-Saharan Africa: Biaka Pygmy 33, Sandawe 25, African American 12, Hausa 16, Ethiopian Jews 7, Mbuti Pygmy 13, Chagga 24, Masai 12, Yoruba 35, and Ibo 7; Middle East: Yemenite Jews 21, Druze 40, Samaritan 25; Europe: Adygei 19, European American 34, Russian from Archangelsk 17, Finnish 16, Irish 49, Hungarian 39, Russian from Vologda 20, Dane 28, Komi 15, Khanty 10, Ashkenazi Jews 18, Chuvash 14; East and South Asia: Cambodian 8, Hakka (Chinese) 24, Korean 32, Keralite 5, (Minnan) Chinese from Taiwan 41, (Canton) Chinese from

San Francisco 25, Laotian 53, Yakut 25, Japanese 21, Ami 24, Atayal 36; Oceania: Micronesian 19, Nasioi Melanesian 8; America: Mexican Pima 22, Quechua 11, Maya 11, Karitiana 18, Ticuna 23, Surui 24. Sample descriptions can be found in the Allele Frequency Database (ALFRED) by searching on the population names or clicking on the sample ID link from the individual single-nucleotide polymorphism (SNP) frequencies. All samples were collected with informed consent under protocols approved by the relevant institutional review boards.

Seventeen Y chromosomal STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a, DYS385b, DYS438, DYS439, DYS437, DYS448, DYS456, DYS458, DYS635, and YGATAH4) were amplified using the AmpFlSTR® Yfiler™ PCR Amplification kit (Applied Biosystems, Carlsbad, CA, USA). Amplified products were separated and detected using the ABI 3730xl Genetic Analyzer (Applied Biosystems, Carlsbad, CA, USA) according to the manufacturer's recommended protocol. The data were analyzed using GeneMapper ID v3.2 (Applied Biosystems, Carlsbad, CA, USA). For use in the analyses, DYS389II was calculated by subtracting the DYS389I allele size.

Y-STR haplotyping was carried out at the MOE Key Laboratory of Contemporary Anthropology; the Lab's proficiency has recently been certified through participation in the Y-STR haplotyping quality test organized by the YHRD (http://www.yhrd.org). Data presented herein have also been submitted to the YHRD for further quality checks in advance of publication and received the accession numbers from YA003930 to YA003973 (Table S1).

Population structure and demographic history inferring

Allele frequencies, haplotype frequencies, and average gene diversity were estimated. The extent of population genetic structure in our data was assessed by means of analysis of molecular variance (AMOVA). More specifically, genetic distances between groups of males were quantified by Fst, R_{ST}, and average number of pairwise differences. The DYS385a/b marker was not included in the population structure and the subsequent demographic history analysis. Multidimensional scaling (MDS) analysis served to visualize differences in Y-STR genetic variation between populations and was based upon pairwise linearized F_{ST} values, that is $F_{ST}/(1 - F_{ST})$. Reynolds' coancestry coefficients $(-\ln(1 - F_{ST}))$ (Reynolds et al. 1983) between every two populations were also calculated. All analyses were performed using R statistical software v3.0.2 (R Core Team 2013) or Arlequin v3.5.1.3 (Excoffier et al. 2007), as appropriate. YPredictor by Vadim Urasin v1.5.0 (http://predictor.ydna.ru/) was used for haplogroup prediction. YPredictor is based on the phylogenetic trees of

each haplogroup and uses the difference in marker values, marker mutation rates and age of parent node to calculate prediction probability.

Demographic history estimations for the studied populations and predicted haplogroups were made using 15 STRs in BATWING (Wilson et al. 2003) under a model of exponential growth from an initially constant-sized population. The parameters used in estimation were following Xue et al. (2006). Four sets of Y-STR mutation rates were applied in time estimations following Wei et al. (2013): (1) a widely used evolutionary mutation rate (EMR) (Zhivotovsky et al. 2004), (2) two observed genealogical mutation rates (OMRB and OMRS) (Burgarella and Navascués 2011; Shi et al. 2010), and (3) a genealogical mutation rate adjusted for population variation using a logistic model (lmMR) (Burgarella and Navascués 2011). A total of 104 samples of the program's output representing 10^6 MCMC cycles were taken after discarding the first 3×10^3 samples as burn-in. The TMRCA was calculated using the product of the estimated effective population size N_e and the height of the tree T (in coalescent units) (Wilson et al. 2003). In addition, TMRCAs of 25 Y chromosome predicted haplogroups were also estimated by average squared distance (ASD) method, which assumes the median haplotype is the founder haplotype within a lineage (Zhivotovsky 2001; Ramakrishnan and Mountain 2004; Sengupta et al. 2006). A generation time of 25 years was used to produce a time estimate in years.

Results

We observed 136 different alleles in 17 Y-STR loci of the 979 haplotypes analyzed (Table S2 and Supporting doc1). The largest average gene diversity over loci was observed in Europe and the smallest was found in the Americans at the continental level. The Americans also showed the lowest expected heterozygosity over loci. Average gene diversity over loci also varied a lot in different populations, from the highest value of 0.6881 in Cambodians and 0.6755 in Hausa to the lowest of 0.1383 in Atayal and 0.1457 in Surui. The low locus diversities in Atayal and American populations were also confirmed by small allele size ranges and number of alleles at different loci of those populations (Supporting doc1).

Various parameters of genetic distances were estimated to infer population structure. At the continental level, the largest genetic distances were noted for the Sub-Saharan populations and the other five groups (all pairwise $F_{ST} > 0.20$, $p < 10^{-4}$; all Slatkin linearized $F_{ST} > 0.25$) (Table S3 and Supporting doc2). Coancestry coefficients between Sub-Saharan populations and the other five groups were all above 0.22 (Table S4). Genetic distances among

non-African populations were much smaller although still significant. The smallest genetic distance was observed between the Far East (East & South Asia) and Oceania (pairwise $F_{ST} = 0.03138$, Slatkin linearized $F_{ST} = 0.0324$) (Table S3). Coancestry coefficients between the Far East and Oceania and between the Middle East and Europe were relatively small, only 0.03188 and 0.06742 (Table S4 and Supporting doc2), respectively, revealing those populations were genetically similar. At the population level, pairs of African and non-African populations showed much larger genetic distances than pairs of African populations or pairs of non-African populations. The largest genetic distances are found between African populations paired with American populations, two Oceanian populations, two Taiwan aboriginal populations (Ami and Atayal), and Yakut (Fig. 1; Table S3). Coancestry coefficients of the above pairs were relatively large, even reached two for Ibo and Atayal and for Ibo and Surui (Fig. 2; Table S4). In contrast, the smallest genetic distances are observed between pairs of Eurasian populations (Fig. 1; Table S3). Most of the coancestry coefficients between Eurasian pairs were below 0.1 (Fig. 2; Table S4). MDS analysis was performed based upon linearized R_{ST}, which also shows a clear geographic clustering pattern. European populations are tightly clustered in the middle of the MDS plot, surrounded by East Asian populations on the top and Sub-Saharan populations on the right. Druze and Samaritan of Middle East tended to be placed between Africa and Europe. American populations were plotted away from the Eurasian center, but showed affinities with the Siberian population (Yakut) and East Asian populations (Melanesian, Han Chinese, and Ami) (Fig. 3). By AMOVA analysis, 76.98 % of the overall variation was within populations, 13.76 % was among populations within groups, defined according to continental residency, and 9.26 % was among groups (Table 1). The larger within population variation was also confirmed using locus by locus AMOVA (Table S5).

Posterior estimates of TMRCA, expansion time, effective population size N_e, and growth rate were then calculated for the 44 worldwide populations using one evolutionary mutation rate (EMR) and three genealogical mutation rates (OMRB, OMRS, and lmMR) (Fig. 4; Table S6). For the first three parameters, median values obtained using EMR were two or three times larger than those calculated by the three genealogical rates, whereas for growth rate, the opposite pattern was seen. There is also a tendency for the population with an older TMRCA to have an older expansion time and a larger effective population size but a lower growth rate and vice versa. Those demographic parameters also show strong geographical patterns. Sub-Saharan African populations tend to have the oldest TMRCAs, the largest population sizes and the earliest expansion times, whereas the American, Siberian, Melanesian,
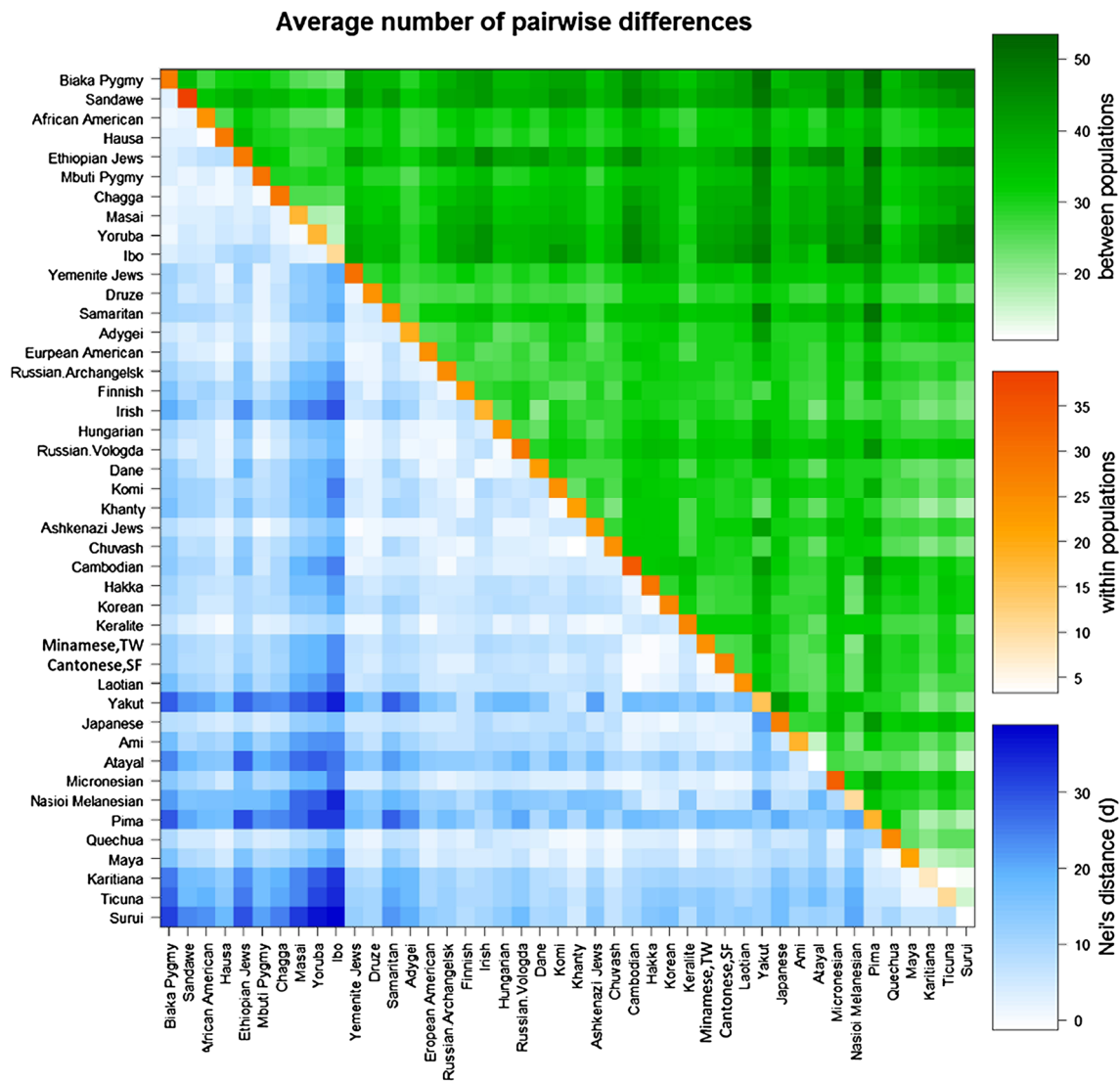
Average number of pairwise differences

Fig. 1 Plots of average number of pairwise differences at population level

and isolated Atayal populations located at the terminals of human out-of-Africa dispersal routes have the most recent TMRCAs and expansion times, and the smallest population sizes. However, some populations stand out from this general pattern. Yoruba and Ibo showed low TMRCAs and population sizes but higher growth rates compared with other African populations. Micronesian and Pima have relatively large TMRCAs and population sizes, but Pima also showed a very recent expansion time (4–8 thousand years ago, kya) coupled with a low growth rate. Adygei, European American, Hungarian, Dane, Laotian, Chinese, Japanese, Ashkenazi Jews, and Cambodian all showed relatively high population growth rates, whereas for populations in Sub-Sahara and Middle East (such as Biaka Pygmy, Sandawe, Mbuti Pygmy, and Druze), the growth rates were extremely low.

Discussion

In this study, the large Y-STR dataset from 44 worldwide populations presented and analyzed here demonstrates the great value of the Y-STRs for male population structure and demographic history inference on a global and regional scale. The largest genetic distances have been observed between pairs of African and non-African populations. American populations with the lowest genetic diversities also showed large genetic distances and coancestry coefficients with other populations, whereas Eurasian populations displayed close genetic affinities. African populations tend to have the oldest TMRCAs, the largest population sizes and the earliest expansion times, whereas the American, Siberian, Melanesian, and isolated Atayal populations have the most recent TMRCAs and expansion times and
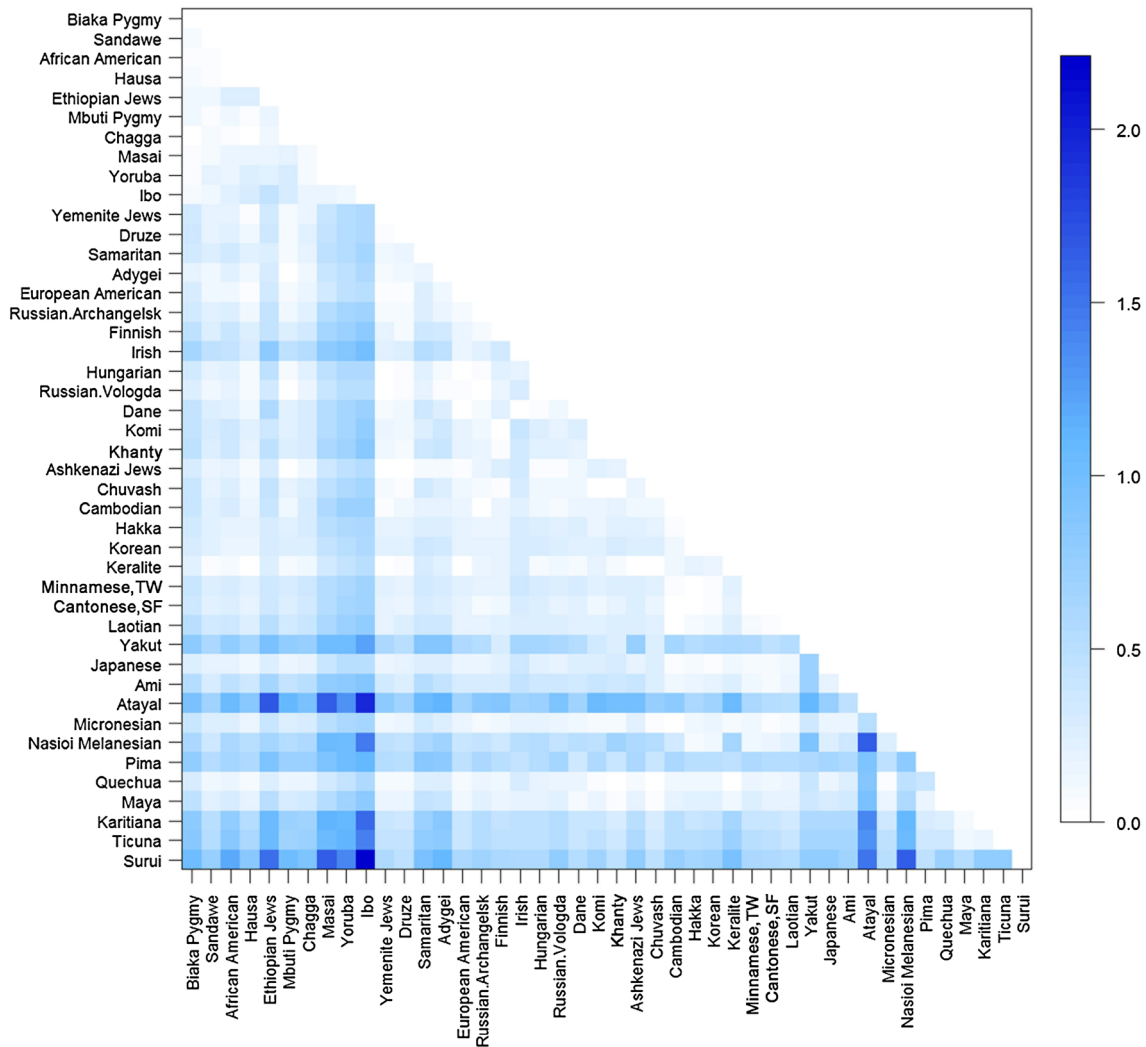
Fig. 2 Plots of matrix of coancestry coefficients at population level

the smallest population sizes. This clear geographic pattern is well consistent with serial founder model for the origin of populations outside Africa.

Some individual populations stand out as exceptions to the above general patterns. For instance, Yoruba showed low TMRCAs, expansion time, and population size, which was consistent with previous study (Shi et al. 2010), probably due to the Bantu Neolithic expansion that erased much of the ancient genetic diversity from this region. Two Neolithic expanded lineages associated with Bantu-speaking people, E-U175 and E-M191 (Pour et al. 2013), comprise more than 80 % of Yoruba (Table S7), which further support the above inference. It is also surprising to find the oldest TMRCA outside Africa is in Pima living in Arizona, which probably reflect their history of recent admixture. Majority of paternal lineages in Pima are indigenous haplogroup Q, while European specific lineage G has also been detected (Table S7). The large TMRCAs of some recent

admixed populations make us to review the method of calculating population TMRCAs using Y-STRs. Although this method has been widely used (Xue et al. 2006; Shi et al. 2010), we have to point out that Y-STR haplotypes in a same population probably belong to different Y-SNP lineages, and grouping those haplotypes together to infer population TMRCA might cause bias. First, Y-STRs under different Y-SNP lineages probably have different mutational mechanisms. Furthermore, taken the mutation rates of the Y-STRs and the time depth of the Y-SNP lineage ramifications into consideration, it is possible to find the same or similar Y-STR haplotypes from different lineages. Thus, inferring TMRCAs using Y-STRs for specific Y-SNP lineages might be more reliable than for populations.

The lowest growth rates are found not only among the Pygmy and Sandawe populations, which might be due to their hunting and foraging lifestyles that could not support rapid population growth, but also among Middle Eastern
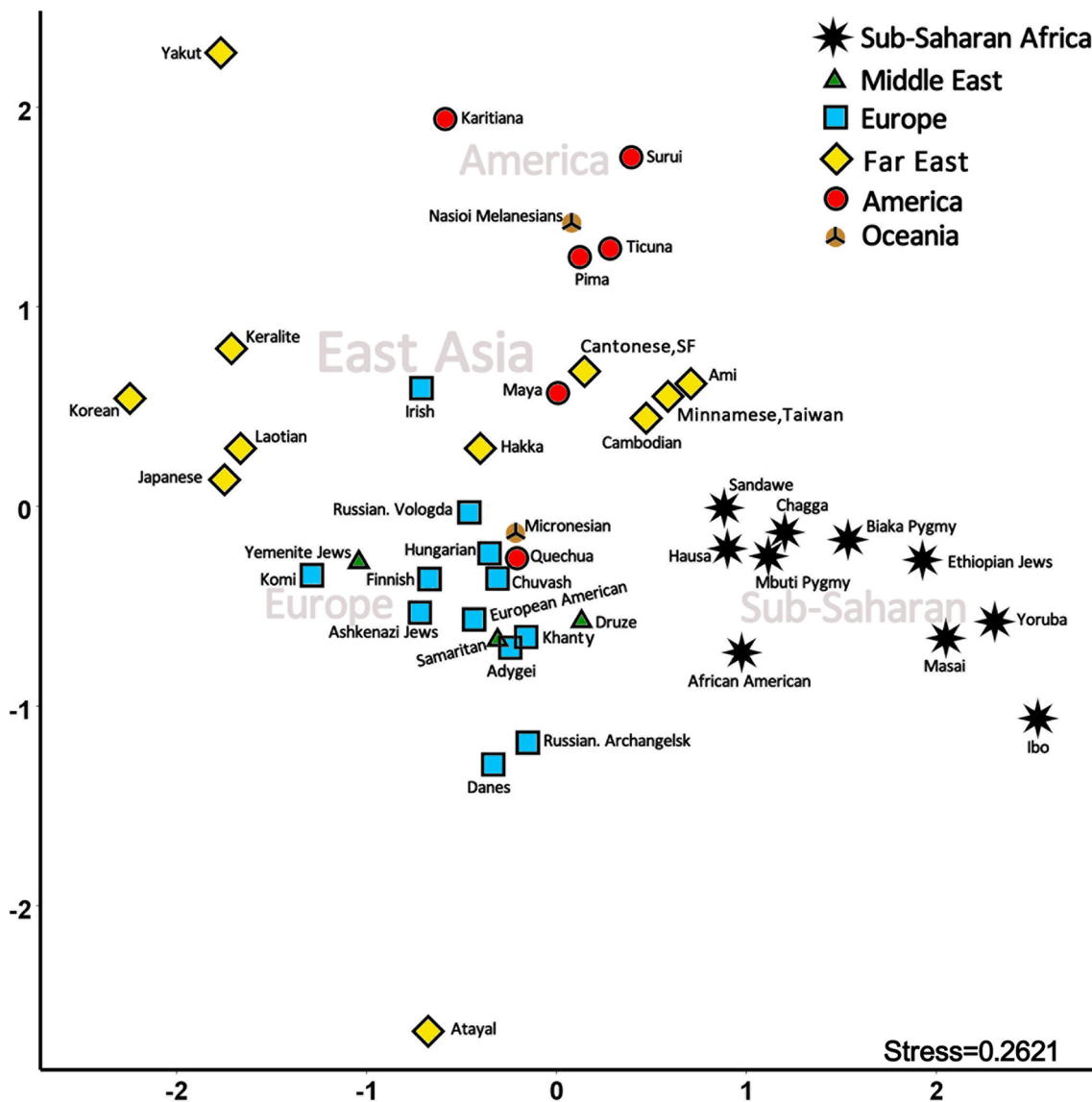
Fig. 3 Population structure revealed by MDS plot

Table 1 AMOVA results

Source of variation	Df	No. of different alleles (F_{ST})					
		Sum of squares	Variance components	Percentage of variation	F_{CT}	F_{SC}	F_{ST}
Among groups	5	473.2	0.48 ($p < 10^{-4}$)	9.26	0.09259	0.15162	0.23017
Among populations within groups	38	744.54	0.72 ($p < 10^{-4}$)	13.76	($p < 10^{-4}$)	($p < 10^{-4}$)	($p < 10^{-4}$)
Within populations	935	3,746.29	4.01 ($p < 10^{-4}$)	76.98			
Total	978	4,964.03	5.2				

populations (Druze, Samaritan, and Yemenite Jews), Siberian populations (Russian from Archangelsk and Vologda, Yakut, Komi, Khanty, Chuvash), and Oceanian populations (Micronesian and Nasioi Melanesian); probably because those climates and environments are not appropriate for agricultural prosperity.
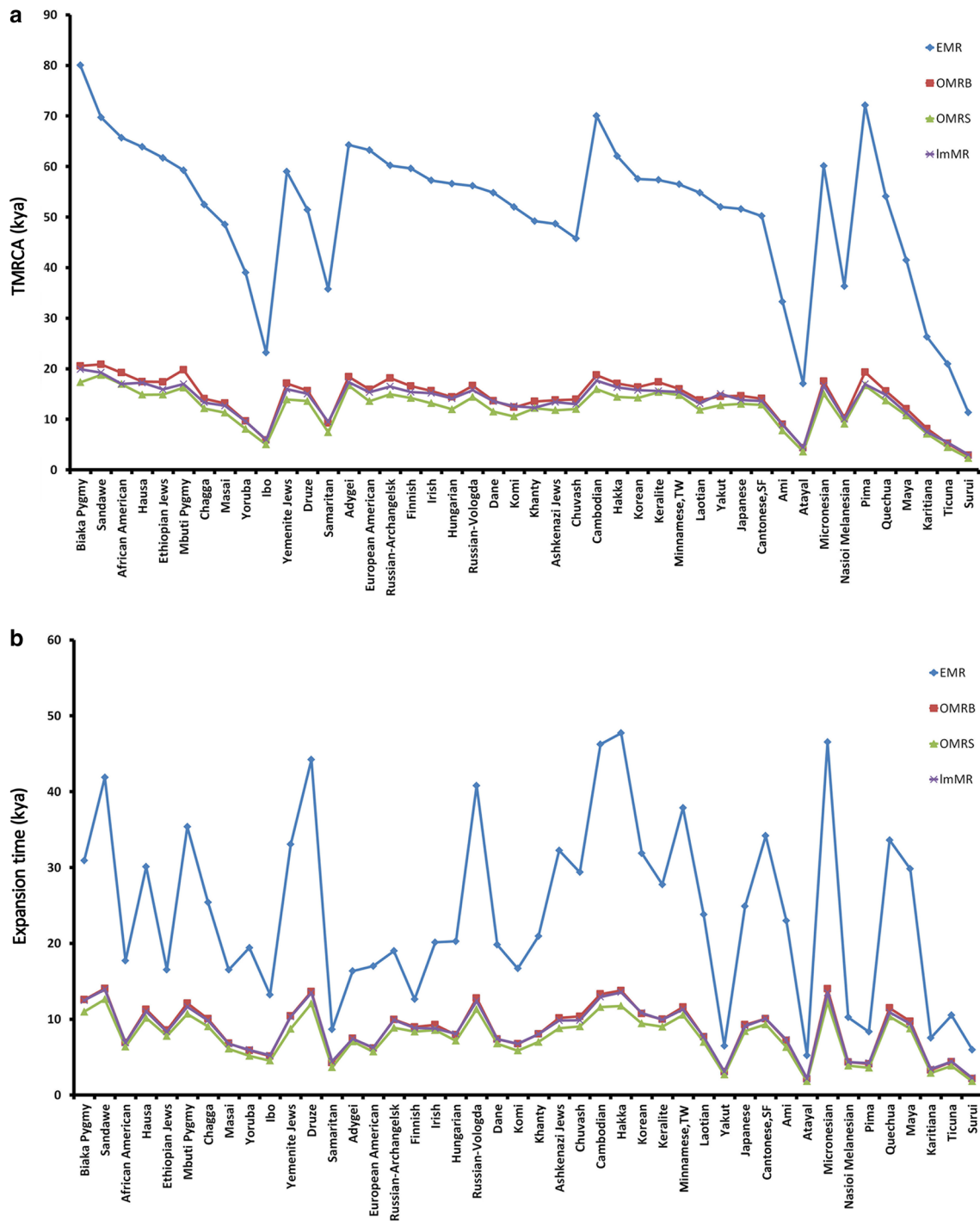
Fig. 4 TMRCA, expansion time, effective population size, and population growth rate for 44 worldwide populations

For population expansion times, estimates based on evolutionary rate (Zhivotovsky et al. 2004) all lie within the Paleolithic period, except three American populations (Pima, Karitiana, and Surui). However, Paleolithic expansions are not in close agreement with estimates from Y chromosome sequencing data and neutral autosomal data, which suggest large population growth in the Neolithic (Wang et al. 2013;

Yan et al. 2013; Gazave et al. 2014). In fact, we have found genealogical rates lead to more plausible time estimates for Neolithic coalescent compared with sequence-based dating in our previous work (Wang and Li 2014). In this study, the expansion times estimated using three genealogical rates for most populations have fallen into the Neolithic period, which matches better with sequencing data, archeological
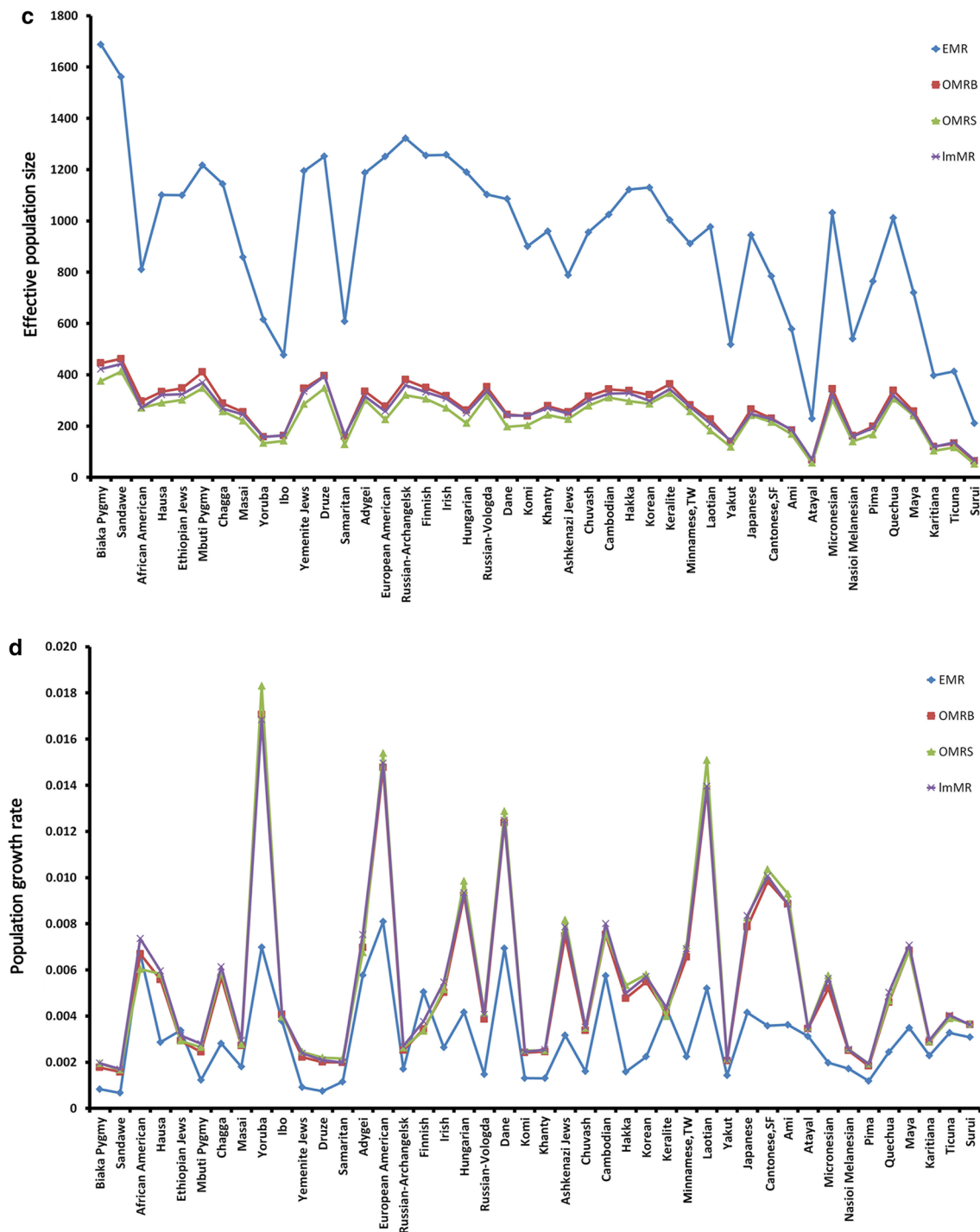
Fig. 4 continued

and historical records. However, we note that all the population-level time estimations are calculated in Batwing using stepwise mutation model (SMM) for all the STRs. It is also possible that different time estimation methods use different algorithms and assumptions might give different results. This interesting inconsistency in time estimation using STR data provides directions for future work.

Another issue we have to address here is that samples from Central or South Asia are under-represented in our presented dataset. We actually lack samples from Central and South Asia, but the most frequent lineages in those regions (haplogroup R, J, L, and H) are very well represented in our dataset, comprising near 25 % of whole samples, and it does not seem to influence the conclusions we

draw from now available dataset. It is noteworthy that the haplogroups and probabilities generated by YPredictor could be only intended as a rough guide for lineage assigning and population paternal admixture analysis, although time estimations for haplogroups are consistent very well with previous publications. The reasons are that: (i) prediction algorithm is reliant on databases; and (ii) the convergence of Y chromosome STR haplotypes might reduce the accuracy of haplogroup prediction. Thus, our paper mainly focuses on the interpretation of Y-STR rather than Y-SNP haplogroup results.

To assist future studies utilizing Y-STRs, complete dataset of Y-STR haplotypes obtained from the 979 individuals analyzed here is made available via Table S2 or YHRD (http://www.yhrd.org). This worldwide dataset will be of great benefit to genealogical studies, population genetic structure and genetic history studies, and forensic applications.

In summary, the Y-STR dataset presented and analyzed here provides the most detailed view of worldwide population structure and human male demographic history. African and non-African population pairs show the largest genetic distances. American populations have the lowest genetic diversities, but showed large genetic distances and coancestry coefficients with other populations, whereas populations in Eurasia displayed close genetic affinities with each other. African populations tend to have the oldest TMRCAs, the largest effective population sizes and the earliest expansion times, while the populations at the tail end of out-of-Africa migration have the most recent TMRCAs and expansion times, and the smallest effective population sizes. Additionally, the dataset exhibits tremendous potential for future forensic application and population genetic studies.

References

Burgarella C, Navascués M (2011) Mutation rate estimates for 110 Y-chromosome STRs combining population and father-son pair data. Eur J Hum Genet 19(1):70–75

Core Team R (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Excoffier L, Laval G, Schneider S (2007) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50

Gazave E, Ma L, Chang D, Coventry A, Gao F, Muzny D, Boerwinkle E, Gibbs RA, Sing CF, Clark AG, Keinan A (2014) Neutral genomic regions refine models of recent rapid human population growth. Proc Natl Acad Sci USA 111(2):757–762

Hellenthal G, Auton A, Falush D (2008) Inferring human colonization history using a copying model. PLoS Genet 4(5):e1000078

Jobling MA, Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4(8):598–612

Kayser M, Brauer S, Willuweit S, Schädlich H, Batzer MA, Zawacki J, Prinz M, Roewer L, Stoneking M (2002) Online Y-chromosomal short tandem repeat haplotype reference database (YHRD) for U.S. populations. J Forensic Sci 47:513–519

Lessig R, Willuweit S, Krawczak M, Wu FC, Pu CE, Kim W, Henke L, Henke J, Miranda J, Hidding M, Benecke M, Schmitt C, Magno M, Calacal G, Delfin FC, de Ungria MC, Elias S, Augustin C, Tun Z, Honda K, Kayser M, Gusmao L, Amorim A, Alves C, Hou Y, Keyser C, Ludes B, Klintschar M, Immel UD, Reichenpfader B, Zaharova B, Roewer L (2003) Asian online Y-STR haplotype reference database. Leg Med 5:S160–S163

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866):1100–1104

Pour NA, Plaster CA, Bradman N (2013) Evidence from Y-chromosome analysis for a late exclusively eastern expansion of the Bantu-speaking people. Eur J Hum Genet 21(4):423–429

Ramakrishnan U, Mountain JL (2004) Precision and accuracy of divergence time estimates from STR and SNPSTR variation. Mol Biol Evol 21:1960–1971

Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105(3):767–779

Roewer L, Krawczak M, Willuweit S, Nagy M, Alves C, Amorim A, Anslinger K, Augustin C, Betz A, Bosch E, Cagliá A, Carracedo A, Corach D, Dekairelle AF, Dobosz T, Dupuy BM, Füredi S, Gehrig C, Gusmaõ L, Henke J, Henke L, Hidding M, Hohoff C, Hoste B, Jobling MA, Kärgel HJ, de Knijff P, Lessig R, Liebeherr E, Lorente M, Martínez-Jarreta B, Nievas P, Nowak M, Parson W, Pascali VL, Penacino G, Ploski R, Rolf B, Sala A, Schmidt U, Schmitt C, Schneider PM, Szibor R, Teifel-Greding J, Kayser M (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. Forensic Sci Int 118:106–113

Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE, Lin AA, Mitra M, Sil SK, Ramesh A, Usha Rani MV, Thakur CM, Cavalli-Sforza LL, Majumder PP, Underhill PA (2006) 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(2):202–221

Shi W, Ayub Q, Vermeulen M, Shao RG, Zuniga S, van der Gaag K, de Knijff P, Kayser M, Xue Y, Tyler-Smith C (2010) A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. Mol Biol Evol 27(2):385–393

Wang CC, Li H (2013) Inferring human history in East Asia from Y chromosomes. Investig Genet 4(1):11

Wang CC, Li H (2014) Comparison of Y-chromosomal lineage dating using either evolutionary or genealogical Y-STR mutation rates. bioRxiv, doi:10.1101/004705

Wang CC, Huang Y, Wen SQ, Chen C, Jin L, Li H (2013) Agriculture driving male expansion in Neolithic Time. arXiv preprint arXiv:1311.6857

Wei W, Ayub Q, Xue Y, Tyler-Smith C (2013) A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping. Forensic Sci Int Genet 7(6):568–572

Wilson IJ, Weale ME, Balding DJ (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J R Stat Soc 116:155–188

Xue Y, Zerjal T, Bao W, Zhu S, Shu Q, Xu J, Du R, Fu S, Li P, Hurles ME, Yang H, Tyler-Smith C (2006) Male demography in

East Asia: a north-south contrast in human population expansion times. Genetics 172(4):2431–2439

Yan S, Wang CC, Zheng HX, Wang W, Qin ZD, Wei LH, Wang Y, Pan XD, Fu WQ, He YG, Xiong LJ, Jin WF, Li SL, An Y, Li H, Jin L (2013) Y chromosomes of 40% Chinese are descendants of three Neolithic super-grandfathers. arXiv preprint arXiv:1310.3897

Zhivotovsky LA (2001) Estimating divergence time with the use of microsatellite genetic distances: impacts of population growth and gene flow. Mol Biol Evol 18:700–709

Zhivotovsky LA, Underhill PA, Cinnioğlu C, Kayser M, Morar B, Kivisild T, Scozzari R, Cruciani F, Destro-Bisol G, Spedini G, Chambers GK, Herrera RJ, Yong KK, Gresham D, Tournev I, Feldman MW, Kalaydjieva L (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet 74(1):50–61