Supplementary Information

Contents

1. Author Contributions

M.K. - analysed data - developed and implemented filtering steps; interpreted results; wrote manuscript; L.S. - analysed data - phylogenetic trees and dates; interpreted results; wrote manuscript; M.V. - analysed data - BSP; interpreted results; wrote manuscript; M.A.W.S. - analysed data - simulations; interpreted results; wrote manuscript; M.J. - analysed data - NRY tree annotation, comparison of sequence and STR data based haplogroup coalescent times; interpreted results; wrote manuscript; M.E. - analysed data - data manipulations; M.Mi, R.M. - analysed data - filtering, data manipulations; A.-M.I., E.-L.L. - analysed data - NRY tree annotation; H.S. - analysed data - NRY tree annotation; contributed Armenian, Assyrian and Iranian samples; T.P. - analysed data - deriving ancestral states; U.G.T. - analysed data - mappability modelling; S.R. - analysed data - SNP Sanger validation, NRY tree annotation, STR genotyping/analysis; L.P. - analysed data - wrote manuscript; C.M., D.L., M.W. - conducted anthropological research and collection of Aborigine samples; G.Z. - conducted anthropological research and collection of Albanian samples; D.M.B. - conducted anthropological research and collection of Arab, Druze and Jewish samples, Sequenced A00 genomes; L.Y. - conducted anthropological research and collection of Armenian, Assyrian and Iranian samples; A.K. - conducted anthropological research and collection of Belarus samples; B.M., M.D. - conducted anthropological research and collection of Buryat, Chukchi, Eskimo, Evenki, Even, Altaian, Koryak, Shor and Yakut samples; S.F. - conducted anthropological research and collection of Buryat, Even Evenki and Yakut samples; C.E., M.Mo - conducted anthropological research and collection of Colla, Cachi, Wichi samples; D.M., L.A. - conducted anthropological research and collection of Croatian and Albanian samples; D.Pr, V.S. - conducted anthropological research and collection of Croatian samples; C.G., J.V.,L.V. - conducted anthropological research and collection of data of the Dutch father-son pairs; G.N.N.S. - conducted anthropological research and collection of Dhaka samples; S.A. - conducted anthropological research and collection of Dusun and Murut samples; A.Me, E.Mi - conducted anthropological research and collection of Estonian, Finnish, Germans, Hungarian, Lithuanian, Moldavian, Karelian, Georgian and Swedish samples; N.A.B. - conducted anthropological research and collection of Evenk samples; G.A. - conducted anthropological research and collection of Georgian samples; J.W.T.S. - conducted anthropological research and collection of Igorot, Burmese and Vietnamese samples; R.W. - wrote manuscript; Z.S. - conducted anthropological research and collection of Kazakh samples;

E.B., O.B. - conducted anthropological research and collection of Kazakh, Kyrgyz, Yaghnobi, Tajiks, Ishkasim, Shugnan, Rushan-Vanch, Cossack, Russian, Ukrainian, Mongolian, Georgian, Circassian, Kabardin, Turkmen and Azerbaijani sampels; E.P. - conducted anthropological research and collection of Kuban Cossack samples; J.I., K.M. - conducted anthropological research and collection of Kyrgyz samples; E.K.K., F.A., I.K., N.T., R.K., S.L., V.A. - conducted anthropological research and collection of Kyrgyz, Uygur, Bashkir, Chuvashe, Karelian, Komi, Mari, Mordvin, Tatar, Mishar-Tatar, Babtized-Tatar, Udmurd, Vepsa, Abkhazian, Avar, Balkar, Circassian, Kabardin, Kumyk, Lezgin, North-Ossetian, Tabasaran, Altaian, Tuvinian, Azerbaijani, Uzbeki and Kazakh samples; D.Pi, F.-X.R., H.R., M.P.C., P.K. - conducted anthropological research and collection of Lebbo and Bajo samples; K.V., N.B. - conducted anthropological research and collection of Mbo samples; P.N. - conducted anthropological research and collection of Mongolian samples; D.V.L., L.P.O. - conducted anthropological research and collection of Nenets, Ket and Selkup samples ; I.E. - conducted anthropological research and collection of Russian samples; K.J. - conducted anthropological research and collection of Saami samples; G.C. - conducted anthropological research and collection of South Asian samples; D.D., S.T. - conducted anthropological research and collection of Turkmen samples; L.At, O.U. - conducted anthropological research and collection of Ukranian samples; A.B.M. - conducted anthropological research and sample collection Aeta, Agta, Batak, Koinambe and Kosipe samples; T.S.K. - conducted bioinformatics at Cph; G.H. - contributed to interpretations of results; B.Y. - contributed to interpretations of results; wrote paper; M.H. - contributed Mbo samples and in interpretation of results; T.M.K. - contributed Mbo samples, interpreted results; M.DG - contributed to analyses and interpretation of results; F.L.M. - contributed to data analyses; A.E., A.Ma - contributed to interpretations of results; wrote manuscript; A.C., F.C., Z.F. - contributed to phylogenetic analyses; Y.X. - contributed to STR genotyping and interpretation of results; C.T.-S., Q.A. - contributed to STR genotyping, wrote manuscript; K.T. - coordinated and contributed Estonian, Saami, Finnish, Germans, Hungarian, Lithuanian, Moldavian, Karelian, Georgian and Swedish samples; contributed to interpretation of results; M.G.T. - coordinated contribution of Aeta, Agta, Batak, Koinambe, Kosipe and Mbo samples; contributed to interpretation of results; wrote manuscript; E.Me - did general sample management and quality control; J.L., S.Ti - provided access to data; wrote manuscript; M.R. - supervised mappability modelling; E.W. - contributed Aborigine samples; wrote manuscript;

R.N. – interpreted the results; wrote manuscript; P.A.U. - contributed to phylogenetic analyses and nomenclature, wrote manuscript; M.Me; R.V - devised the study; analysed data; interpreted the results; wrote manuscript; T.K.- devised and supervised the study; analysed data; interpreted the results; wrote manuscript with input from co-authors. M.K., L.S., M.V., M.A.W.S and T.K., R.V., with M.Me. contributed equally to this work.

2. Filtering the sequence data

We filtered the variant sites by the quality scores provided by Complete Genomics and kept only biallelic SNPs with VQHIGH (samples analysed with Analysis Pipeline versions earlier than 2.4) or PASS score (Analysis Pipeline version 2.4) and call-rate higher than 95%. We developed several additional filters to improve the quality of the resulting data set. The human Chr Y is known to have large segments of sequence that are difficult to map (Skaletsky et al. 2003; Poznik et al. 2013; Wei et al. 2013a). When selecting the Chr Y regions for further analyses, we relied on previous knowledge of Chr Y structure – considering the X-degenerate sequence class (Skaletsky et al. 2003) and the regions considered reliable for next-generation sequencing data (Poznik et al. 2013; Wei et al. 2013a). We initially applied a combination of regional filters previously defined on the basis of analyses of Illumina HiSeq data (Poznik et al. 2013; Wei et al. 2013a), resulting in ten regions of Chr Y sequence, altogether capturing 10.8 Mb (filter c, Table S2). However, the application of the regional filters led only to a modest reduction of false positive calls judged by the number of father-son/brother-brother (FS) differences and the count of recurrent mutations (Table S2). The number of FS differences was approximately 10 fold higher than the expected number of *de novo* mutations considering the range of published Chr Y mutation rates (Xue et al. 2009; Francalacci et al. 2013; Mendez et al. 2013; Poznik et al. 2013). This finding prompted us to explore additional filters.

Altogether, we tested four filters (Table S2): a) *>5x unique sequence coverage filter*, where regions with less than 5x unique coverage on Chr Y were removed; b) *X chromosome normalized coverage filter*, where we tracked the fluctuations of relative unique coverage (UC) normalized to that of the X chromosome (chrX) to highlight the deviation of local sequence coverage from the expected mean; c) *regional exclusion mask*, where we exclude all of Chr Y outside 10.8 Mb sequence mostly overlapping with X-degenerate regions shown to yield reliable NGS data; d) *re-mapping filter*, where we modelled poorly mapping regions on Chr Y and identified those that also map to sequence data derived from female individuals.

We tested several thresholds for the relative unique coverage filter and assessed the combined effect on the overall number of variant sites, recurrent mutations and *de novo* mutations (Table S2). After applying several combinations of these filters, we conclude that out of the 56.3 Mb total sequence length of Chr Y, only 8.3 – 10.8 Mb produce data where we can have high confidence that the variants map uniquely to haploid segments of Chr Y. For further analyses we selected the combination of filters a, b and d (Table S2a_b_d) that reduced the proportion of recurrent mutations to 1.6% and the average number of father-son differences to less than 1 (Table S2) while retaining 35,672 variable positions. Phylogenetic analyses, including Bayesian skyline plots and coalescent time calculations were based on these filter settings while haplogroup annotations in Supplementary Table Annotations are reported for a wider set of 42,340 binary SNPs that had call rate of 95% and were captured by filter c (Table S2, Table S2c).

While combining A00, A2, A3 sequences with the rest of the phylogeny we noticed among these deep lineages clustering of mutations at close ranges likely reflecting structural variation. We applied additional filters to mask mutations that were within 70 bp range in these samples.

Detailed description of the filters

Filter a: The >5x unique sequence coverage in >95% individuals

This filter excludes positions where more than 5% of the individuals have unique coverage below 5X. The total length of the blacklist is 10,016,681 bp, leaving 46,367,847 bp of usable sequence. The filter is best applied together with other filters.

Filter b: ChrX normalized coverage

We normalized each individual’s coverage on Chr Y (in 1,000 bp windows sliding by 50 bp increments) to that of his average chrX coverage. To consider the range of variation of coverage in Chr Y segments we calculated the log10 average and SD across the given sliding window for

307 individuals for which we had coverage information available. In the same sample we used two parameters to create a blacklist of stretches of Chr Y based on poor unique coverage – the amount of variability (sdThres) and the 'glue'-parameter (nCap). These two parameters define the regions to be excluded:

1) sdThres – the tolerance of variation in unique coverage in the given window. If the SD margins are wider than a set value the window is excluded to provide uniform coverage for bases used in downstream analyses.

2) nCap – the 'glue'-parameter defining the number of consecutive windows with extreme coverage values. If coverage is significantly lower or higher by the sdThres parameter in a set number of consecutive windows, e.g. because of a deletion/duplication in this region, we exclude it from further analyses.

Example of this approach is shown in Figure S1. To choose the best combination of nCap/sdThres, we tested the effect of several parameter settings together with other filters and monitored the number of father-son differences, proportion of recurrent sites and the total number of sites retained (Figure S2). Other filters tested included the regional masks from published sources (Poznik et al. 2013; Wei et al. 2013a), sites that also re-mapped in females and sites with less than 5x unique coverage. We tested nCap values 25–150 with 25 bp step size of the sliding window, and also 200, 300 consecutive windows. sdThres values were tested in the range of 0.2 to 3.0 (small intervals at lower values and increasing at higher values) (Figure S2). To minimize both the data loss and the proportion of false positives we decided to use in downstream analyses the nCap=80 and sdThres=0.8 combination. The filter, when applied alone, keeps 12,315,045 bp of Chr Y sequence, removing 47,058,521 bp.

Filter c: Regional exclusion mask

This filter includes the X-degenerate sequence class (Skaletsky et al. 2003) and the regions that in the previous studies have been considered reliable for next-generation sequence data (Poznik et al. 2013; Wei et al. 2013a). We used a combination of the coordinates defined in Wei et al. 2012 and Poznik et al. 2013 (Poznik et al. 2013; Wei et al. 2013a), resulting in ten regions of high quality Y-chromosome sequence, altogether capturing 10,793,302 bp and consisting of 10 distinct regions.

Filter d: Re-mapping filter

For additional data filtering we also applied extrapolated information on next generation sequencing read mismapping areas, obtained from modelled Illumina datasets. These masks eliminate additional areas of low mappability on a short read level, inherent to genomic areas of frequent repeats as well as high sequence homology content. Since our mappability model takes into account the contribution of known SNVs into short sequence read placement ambiguity, it may become an essential correction factor when comparing datasets with highly variable phylogenetic distance from the reference genome.

We modelled reference Chr Y sequence mappability back to itself with and without reported SNPs, including regions of Chr Y that also map to the autosomes and chrX. The filter retains 9,831,618 bp of sequence including 38,943 binary SNPs.

Validation of SNVs between fathers and sons

We validated with Sanger sequencing SNPs between three fathers and their six sons and between two brothers in one case where father's genome failed QC. We compared altogether seven pairs from four Estonian families. At first we applied VQHIGH, 95% call-rate, custom filter combination (a+b+d) and excluded all individual N-s from comparisons. This filtering scheme revealed within these seven pairs in total 6 differences. All these occurred within two families and in 4 unique positions. These were 1) two positions where both fathers had each one position where they were in derived state but their two sons in reference state, thus implicating a possible sequencing error (or back mutation); 2) two unique positions where one father had two sons carrying one derived allele each, implicating possible mutation between generations. To see the effect of different filtering we also used only filters b+d which revealed that both sons from the family already carrying four differences have each one additional derived allele. For these six SNV-s we were able to design Chr Y specific primers and to get homozygous Chr Y sequence calls for five of these positions. Sanger sequencing revealed that all of the studied positions were false positives.

Ancestral allele inference

We used two different outgroups to infer the ancestral status of the binary SNPs reported in Table S8. Firstly, we used the ancestral inference based on chimpanzee outgroup data http://pipeline.lbl.gov/data/hg18_panTro2/ that is based on AVID alignment of chimpanzee and human reference sequences http://genome.cshlp.org/content/13/1/97.short. Secondly, we used the Illumina high coverage sequence data for the most basal haplogroup A00 to define the ancestral states of the SNPs in our Complete Genomics sequences that were all from the A2'T clade (Figure S14). In case of both approaches we assumed the ancestral status to the alleles that were shared with the outgroup.

mtDNA sequence filtering and haplogroup calling

In order to contextualize the Chr Y phylogeny and BSP plots we analyzed the mtDNA genomes of the same set of 323 male individuals. In the phylogenetic analyses we used mtDNA variants reported in CG mastervar file as "snp", "sub" and "complex" variants. We ignored insertion and deletion variants as the mechanisms by which these evolve are different from the single nucleotide substitutions. The variants had been mapped against the rCRS sequence positions (Andrews et al. 1999) in the testvar file we generated by cgatools software of CG. We created the final mtDNA sequence files by converting CG diploid calls (for sequences analyzed by CG pipeline 2.4) to haploid (00 to 0, 11 to 1, all other states to N) and added sites that in the CG pipeline are listed as 'substitutions' and 'complex calls' in the next stage using locally developed scripts. Among the 2,146 mtDNA variants 85 were annotated as substitutions in the original testvar file and these were converted to SNPs before further downstream analyses. Similar to Chr Y analyses, variants with 5% or more missing calls were excluded from further analyses. In total, we removed 7 such positions. With these final settings we obtain an overall call rate of 0.9999, with the call rate per sample ranging from 0.999 and 1. Finally, haplogroups were inferred with HaploFind (Vianello et al. 2013), that uses haplogroup definitions reported by Phylotree (van Oven and Kayser 2009).

3. Y chromosome mutation rate and haplogroup age estimation

Mutation rate

Different Chr Y mutation rate estimates, broadly in the range of 0.6-1x10^{-9} /bp /year, have been offered over the past few years on the premises of calibrations based on deep patrilineal pedigrees (1x10^{-9} /bp/year rate, (Xue et al. 2009)), transfer of autosomal pedigree rate on Chr Y data (0.62x10^{-9} /bp/year rate, (Mendez et al. 2013)), archaeological evidence for peopling of the Americas (0.82x10^{-9} /bp/year rate, (Poznik et al. 2013)), and Sardinia (0.65x10^{-9} /bp/year rate, (Francalacci et al. 2013)). The application of the minimum and maximum ranges of these mutation rates on Chr Y phylogeny results in two fold differences in age estimates of haplogroups. For example, the age of haplogroup F that captures the majority of non-African lineages and is considered to be a proxy of the out of Africa dispersal varies in the range of 50 ky (Poznik et al. 2013; Wei et al. 2013a) to 110 ky (Francalacci et al. 2013; Mendez et al. 2013).

In order to minimize the effects of NGS differences and autosomal versus sex chromosome specifics on mutation rate calibration, and to avoid the need to make assumptions about the extent of genetic variation in relation to archaeological evidence, we calibrated the Chr Y mutation rate in our CG data by using inferences of the coalescent times of two Chr Y haplogroups, Q1 and Q2b, from ancient DNA data. We used Chr Y data of the 12.6 ky old Anzick (Q1b) and 4 ky old Saqqaq (Q2b) specimens (Rasmussen et al. 2010; Rasmussen et al. 2014). In both cases we used only transversion polymorphisms and the approach described in Rasmussen et al. 2014.

In case of haplogroup Q1a’c-L53 (Figure S35) variation we fixed the age of this clade in which Native American and Siberian Q lineages coalesce at 16.9 ky following the SI 13 of Rasmussen et al. 2014 (Rasmussen et al. 2014). From the phylogenetic distances to this ancestral node in 17 descendant samples that were sequenced using the CG platform, we estimated the mutation rate of binary SNPs as 0.73×10^{-9} /bp/year (95% CI 0.63–0.95 x10-9 /bp/year).

The Saqqaq sequence shared the highest number of derived alleles in our data with four Koryak sequences in haplogroup Q2b-B143 (Figure S35). We applied the a+b+d filter (Table S2) on the Saqqaq data and found that out of the 53 transversions that were shared by the four Koryak Q2b sequences and absent in their closest outgroup sequence of the Murut individual in haplogroup Q2c, 18 were also called in the Saqqaq data at 4x or higher coverage after the removal of polymorphisms that had another sequence variant in the ancient DNA data within the neighbouring 10 bps. Of these 18 positions, Saqqaq carried the derived allele in 12 and the ancestral allele in 6 cases. From the derived/ancestral ratio observed in these 18 sites we estimated 2/3 as the derived and 1/3 as the ancestral allele probabilities for the 35 Q2b-B143 defining sites for which the Saqqaq sequence had no calls. We estimated the average call rate of transversions in the filtered Saqqaq data as 0.2532 at positions where haplogroup Q2 differs from the reference sequence. We used this empirical call rate to correct the number of private mutations of the Saqqaq sequence. We observed two private transversions in the filtered Saqqaq data which when divided by the call rate gives us the estimate of the length of the private branch leading to the Saqqaq sample as 7.9 transversions. Considering that the four Koryak sequences are on average at a distance of 19.5 transversions from the most recent common ancestor they share with Saqqaq, we estimate that the Saqqaq misses 11.6 transversions per 4,000 years, which considering 1.7 as the average transition/transversion ratio in our a+b+d filtered data for 8,819,704 positions, yields a rate estimate of binary SNPs of 0.89x10^{-9} /bp/year.

For the calculations of Chr Y haplogroup coalescent times and BSP analyses we combined the two ancient DNA based mutation rate estimates using weights proportional to the product of age and coverage of both ancient DNA samples yielding the final estimate of 0.74 x10^{-9} /bp/year with 95% CI of 0.63–0.95 x10-9 /bp/years.

Haplogroup coalescent times based on sequence and STR data

The coalescent ages of Chr Y haplogroups were estimated using two methodologies: Bayesian inference applied on sequence data (SI4) and using short tandem repeat (STR) data. The STR base age estimates were drawn using the method developed by Zhivotovsky et al. (2004) (Zhivotovsky et al. 2004) and modified by Sengupta et al. (2006) (Sengupta et al. 2006) which calculates the average squared difference in the repeat count of all sampled chromosomes from their inferred ancestral haplotype. The ancestral haplotype is determined as the median repeat count at each STR locus. We used 21 STR loci genotyped using the PowerPlex® Y23 System (Table S9, data for the multi-copy DYS385a/b locus was excluded from the age estimation analyses). The 'evolutionary' mutation rate of 6.9x10^{-4} per 25 years (Zhivotovsky et al. 2004) was used, as it has been shown repeatedly (Dulik et al. 2012; Underhill et al. 2014) that the various 'pedigree' rates (Heyer et al. 1997; Ge et al. 2009; Goedbloed et al. 2009; Wei et al. 2013b) are too fast for haplogroup age estimation. Although the rate of 6.9x10^{-4} per 25 years applies to tri- and tetranucleotide markers and STRs with longer repeat units are known to evolve slower (Jarve et al. 2009), the removal of the two penta- and one hexanucleotide STR in the PowerPlex® Y23 System did not have a notable effect on the age estimates (data not shown). STR-based coalescent time estimates were calculated for clades that had at least 5 samples (an exception was made for hg I1, for which the STR-based age was computed based on 4 samples).

The STR and sequence data based age estimates of haplogroups (Figure S8) were generally consistent with those presented in Figure 4 of (Wei et al. 2013b). The trend, confirmed by our larger sample size and diverse set of haplogroups, clearly shows that STRs overestimate the ages of younger haplogroups, but agree better with sequencing-based estimates for older (20–50 ky) haplogroups (Figure S8). To illustrate the time-dependence of STR based coalescence time estimates, we calculated age class dependent STR mutation rates assuming linear relationship with haplogroup age estimates derived from sequence data (Figure S8 right y-axis). The age class dependent STR mutation rates were estimated for 10 bins, each spanning 5,000 years. Haplogroups were assigned to the 10 bins on the basis of their coalescent times estimated from sequence data. The average STR mutation rate for each bin was calculated from the observed STR variation within constituent haplogroups assuming linear relationship with sequence data based age estimates. The resulting graph (green line, Figure S8) shows that the STR-based ages deviate most strongly for the youngest haplogroups (ages up to 5,000 years), where the STR mutation rate would need the largest adjustment (towards a faster rate). On the other hand, for haplogroups older than 30,000 years, the corrected STR mutation rate converges to the 'evolutionary' rate of 6.9×10^{-4} per 25 years (Zhivotovsky et al. 2004).

4. Phylogenetic analyses

Summary statistics, such as nucleotide diversity, mean pairwise differences and AMOVA were computed in Arlequin v3.5.1.3 (Excoffier and Lischer 2010). Both mitochondrial DNA (mtDNA) and Chr Y showed approximately two-fold higher nucleotide diversity in African than in non-African populations (Table S3). The proportion of genetic variation on the global scale that was observed within populations was 76% in case of mtDNA and 66% in case of Chr Y while 12% of mtDNA and 18% of Chr Y variation was apportioned among regional groups and 12% of mtDNA and 16% of Chr Y variation was among populations within the regional groups (Figure S10). The Papuan, Central Asian, and Siberian populations show a higher male-specific diversity among populations. More mtDNA than Chr Y variation was observed among populations in our African dataset (Figure S10).

We used software package BEAST v.1.8.0 (Drummond et al. 2012) to reconstruct phylogenetic trees, estimate coalescent ages of haplogroups and sex specific effective population sizes. Prior to the BEAST analyses we executed a jModelTest run (Darriba et al. 2012) to identify the best fitted substitution model for the Chr Y and mtDNA. Based on Akaike and Bayesian information criteria the GTR substitution model was selected as the best fitted for the Chr Y data and the HKY+I+G for the mitochondrial genomes.

The BEAST analyses for the Chr Y regional Bayesian Skyline Plots (BSPs) were run using a relaxed lognormal clock with a mutation rate 0.74×10^{-9} /bp/year (SI3) and a strict clock model with a rate of 1.665×10^{-8} /bp/year for mtDNA (Soares et al. 2009). At the end of each BEAST run, the results were inspected and visualized in Tracer v1.6 and it was confirmed that all ESS values were above 200 (Drummond et al. 2012). For the 8 geographically explicit regions (Table S1), we generated BSPs for both Chr Y and mtDNA data (Figure S4). We did not perform BSP analyses on the Papua New Guinea region because of the small sample size (n=6). In order to reduce the computational load, the Chr Y BEAST analysis only contained the variable positions. However, the BEAST input xml file was modified by adding another parameter under the *patterns* section that specifies the nucleotide composition at invariable sites.

Two independent runs with piecewise-linear model with 10 groups (6 for the Andeans) and MCMC between 20 and 50 million iterations (depending on the group size), with a sampling in every 1,000 steps were made for Chr Y and mtDNA geographical BSP analyses. After inspection in Tracer, the 2 independent runs were merged under the LogCombiner v1.8.0 with a burn-in of 20% discarded. For reconstructing the Chr Y phylogenetic trees of the global geographic as well as the 456 sample dataset, which only excludes close relatives and duplicates, 6 or 8 (depending on the number of samples) independent BEAST analyses were run for 200 million iterations, sampling every 5,000 steps, with strict clock, 15 skyline groups, and 10% burn-in when combining results. Maximum likelihood (ML) trees inferred with RAxML v. 7.8.6 (Stamatakis 2006) were provided as starting trees. Other relevant settings and steps were as described above. The ML starting trees had fairly uniform distances from the root to the tips. The applicability of strict clock was confirmed by comparing AICM scores (Baele et al. 2012) of strict and relaxed lognormal clock models of the smaller dataset. The BSPs for Chr Y and mtDNA were plotted together in R (R Core Team 2012) using the package ggplot2 (Wickham 2009).

The Chr Y specific dip that was observed in Figure 2 in most regional populations examined (see also Figure S4) has the region-specific age estimates and minima reported in Table S4.

In order to assess if the Chr Y BSP dips were caused by potential biases caused by our specific data filtering scheme we replicated the analyses using only the regional exclusion mask (Filter C). We were unable to detect a significant difference in the BSPs and the magnitude of the dip (data not shown).

To test for significant deviations in diversification rates along the branches of Y chromosome tree we used SymmeTree 1.1 (Chan and Moore 2005). Six branches that showed significant ($p<0.05$) P_Lambda1 values are shown in Figure S3 with the number symbol (#).

5. Simulations

Settings

FastSimCoal2 simulations of 500,000 sites of Chr Y and 16,569 sites of mtDNA were performed using mutation rates specified in SI4 and starting population size 10,000. Coalescent times of all nodes in the resulting trees were estimated under a constant size and exponential growth models. Each scenario was run 1000 times, and we sampled either 100 or 500 individuals after each run. The growth model assumed constant size until 400 generations followed by exponential growth (Keinan and Clark 2012). The starting population sizes for the exponential growth model were chosen so that the true height of the tree (on average) was the same between the exponential and constant population models.

For each model we plotted the histogram of coalescent times of all the nodes of the simulated trees by six different deme formation scenarios: 1: no deme structure; 2: formation of 10 demes 400 generations ago; 3-6: formation of 25, 50, 75 and 100 demes 400 generations ago, respectively. Under each scenario we modelled six different Nm/Nf ratios which were held constant within each scenario over time: 1) $N_m/N_f = 1$; 2) $N_m/N_f = 0.75$; 3) $N_m/N_f = 0.5$; 4) $N_m/N_f = 0.25$; 5) $N_m/N_f = 0.1$; 6) $N_m/N_f = 0.01$. We show the results of the formation of 1 and 100 demes and N_m/N_f ratios of 1, 0.5 and 0.1 in Figure S7.

Results

Figure S7 shows the results of the simulations under constant size and growth models. As a sanity check, in all scenarios (for any number of deme formations, and for either constant pop size or exponential growth) we confirmed first that the distributions of node ages for Chr Y versus mtDNA assuming N_m/N_f =1 were not significantly different. Under an assumption of constant population size, we indeed observe what looks like reduction in node ages around 400 generations, at the same time as deme formation. Under the simulation that samples 100 individuals, this reduction, and recovery, is apparent for the formation of 25, 50, and 75 demes (not at 100 demes because there is no resolution with only 100 samples). Under the scenario where we sample 500 individuals, as expected, we can pick up more rare variants, so we observe a larger number of recent coalescent events. Nevertheless, we still observe an effect of deme formation (a reduction in diversity) around 400 generations ago.

Considering the models with constant population size, we observe that as N_m/N_f decreases, there is an *increasing* effect of deme formation on the Chr Y, and a *decreasing* effect of deme formation on the mtDNA. This suggests that, if N_m/N_f ratios have been skewed for much of human history, then we might expect deme formation to more drastically effect Chr Y than mtDNA. However, since this does not take into account the recent exponential growth of the human population, we also investigate the same set of models, with deme formation, and variations in the ratio of the effective numbers of males and females contributing to the next generation, under a model that assumes exponential growth, starting at the same time as deme formation, 400 generations ago.

For the exponential growth models, we do observe a shift in node ages around 400 generations ago, around the start of the exponential growth, and very little effect of the deme formation. Thus, under a model of exponential growth, deme formation likely cannot explain a reduction in diversity on the Chr Y. However, in this set of models, deme formation and exponential growth both occur at 400 generations. So, we conducted two additional sets of models, both keeping deme formation at 400 generations ago, and: 1) delaying the start of exponential growth until 200 generations ago; or 2) allowing exponential growth to start 600 generations ago (200 generations before deme formation).

These simulations confirm that exponential growth dilutes the effect of deme formation. More specifically, what these simulations suggest is that if exponential growth started before deme formation, it would be very unlikely that deme formation could be responsible for a dip in effective population size on the Chr Y. However, if exponential growth started after deme formation (and there was an extreme difference in the effective number of males and females), then deme formation may be able to explain the reduction in N_e on the Chr Y, and the lack of a reduction on the mtDNA.

6. Nomenclature

The initial Y Chromosome Consortium (YCC) nomenclature was based on the genotyping of almost all known Chr Y SNPs of the time, 237 altogether, in a panel of 74 male YCC cell lines (Y Chromosome Consortium 2002). The resulting tree defined 153 Y chromosome haplogroups by an alternating alphanumeric system. This initial YCC nomenclature combined the horizontal labelling of 19 primary branches in the tree by capital letters and a vertical labelling scheme derived from a set theory that assigned unique alphanumerical labels to each clade. The nested (vertical) scheme of labelling clades allows them to be mapped precisely on a tree. This approach has been successfully applied over the past decade in many Chr Y studies which have introduced tens of thousands of new Chr Y markers that have been catalogued and maintained in tree format by ISOGG http://www.isogg.org. Advances in sequencing technology have significantly increased the capacity to generate new data and it can be foreseen that within the next couple of years hundreds of thousands of whole Chr Y sequences will be generated thanks to the concerted efforts of evolutionary population genetic and forensic research, as well as personal genomics.

The increasing amount of sequence data poses a substantial burden to the existing Chr Y haplogroup nomenclature because the number of branches (or haplogroups) of a tree is approximately twice the number of its tips: (n-1)x2 in case of fully bifurcating trees. Given that unlimited number of parallel branches of the same clade can be labelled alphanumerically (e.g. O1, O2, O3..) and the YCC nomenclature rules allow the ancestral branches, as these are revealed by increasing sequence resolution, to be called as the “join” or union of two subclades (e.g. the ancestral clade joining O and N is called NO) we propose to simplify the Chr Y haplogroup nomenclature by defining a limited number of levels of alphanumeric depth to be used in the haplogroup names. In line with van Oven et al. (2014) we use the apostrophe symbol (‘) to denote the ‘joined’ names of related haplogroups at depths greater than Level I (Table S5). For example, O1’2 refers to a clade (Figure S32) that joins O1 and O2 (with O3 being an outgroup) and the use of the ‘join’ rule here, thus, circumvents the need to create an additional layer of haplogroup names. While applying the ‘join’ rule on bifurcating clades it is sufficient and can be preferable in practice for the sake of brevity to combine the names of only one representative both from the left and right hand side of the phylogeny – for example, O1’2 instead of O1’2’4’5, considering that O2 forms a clade with O4 and O5 (Figures S32-33). While joining Level II+ clade names it is sufficient to spell out the full haplogroup name for only one of the subclades that are merged together and report only the final symbol from the name of the other – e.g. Q1b’c instead of Q1b’Q1c (Figure S35).

We propose that the alphanumeric depth is approximately proportional to the time depth of respective haplogroups estimated from the sequence data. Table S5 includes explicit examples of the proposal as applied on our data using a relaxed set of filters. Above each temporally restricted depth level the “join” rule (in which unlabeled clades can be named as the union of two most distant subclades) can be used alongside the Haplogroup nomenclature by Marker style of coding described in Y Chromosome Consortium 2002 to define uniquely any branch that occurs on the tree. To maximize consistency with historically defined haplogroup labels we keep these within our revised nomenclature depth levels as much as possible.

We assigned sequenced individuals into Chr Y haplogroups (Table S6) according to the phylogenetic tree estimated in BEAST (Figure S3). Posterior estimates of phylogenetic support for each clade along with age estimates and their confidence intervals are provided in Table S7. Haplogroup defining mutations are reported in Table S8 and the STR profiles of a subset of individuals reported in Table S9.

In haplogroup A we keep the A00, A0, and A1 definitions according to (Cruciani et al. 2011; Mendez et al. 2013). Following Table S5 scheme we define four subclades (A2-A5) in the A2’5 clade defined by L419 (Figure S14). In haplogroup B we keep B1 and B3 labels whereas in B2 the sequence depth allows us to recognize at least three level II clades (B2, B4 and B5; Figure S15). In haplogroup E (Figures S16-19), we simplify the structure of former E1b (van Oven et al. 2014) clade which is now divided between E1-V38, E2-M215 and a minor branch E3-P75, and former E1a, which is now called E4-M33. Subsequently, the predominantly non-African branch E1b1b1a-M78 (Karafet et al. 2008) has been renamed as E2a (Figure S19).

Coherent with previous studies, all non-African samples that we sequenced allocate to haplogroups C, DE, and F. Consistent with Poznik et al. (2013) (Poznik et al. 2013) we observe an early split in our haplogroup F data that separates haplogroup G from HT-M578. However, intersecting our data with Malaysian Chr Y sequence data (Wong et al. 2013) reveals a split in haplogroup F that predates the G/HT split by one mutation, F1329 (Figure S13). This finding is in accordance with the two Lahu F2-M427 individuals reported in Poznik et al. (2013) as having an ancestral allele of M578. In combination with the presence of deep branches of K in Southeast Asia, this further strengthens the model proposing that the initial radiation of the non-African Chr Y lineages may have taken place somewhere in Southeast Asia (Karafet et al. 2014). Following PhyloTreeY (van Oven et al. 2014) we re-define the internal structure of haplogroup H-M3035 that now incorporates South Asian lineages H1-M69 (predominantly found in Indian peninsula), H2-B108 (detected in one of our Burmese samples) and H3-Z5857 (India) that previously (Karafet et al. 2008) were recognized as F* (Figure S23). Although all F* lineages from South Asia in our data belong to H the phylogenetic depth (40-44 kya) of its division into three primary subclades suggests that their distribution patterns may also be considered informative about the process of initial radiation of non-African Y chromosomes (Figure S9). Although absent in 728 South Asian samples (Sengupta et al. 2006) the rare H4-M282/P96 lineage (van Oven et al. 2014) has been observed in two Iranians (Regueiro et al. 2006), one French (Poznik et al. 2013) and one Dutch individuals (Karafet et al. 2008) as well as several low coverage Sardinian sequences (Francalacci et al. 2013). Intersecting the Sardinian variants with our data allows us to approximately position the H4 lineage within haplogroup H (Figure S23).

Only 24 mutational events distinguish the progression of two major non-African founder haplogroups F to K (Figure S13). Similarly small number of differences separate haplogroups LT, NO, S and P from their MRCA in haplogroup K (Figure S28), consistent with the suggestion of Karafet et al. 2014 (Karafet et al. 2014) that the initial diversification of Eurasian and Oceanian founder haplogroups was a rapid process limited to a few thousand years overall. We estimate that a peak of the coalescent events of the oldest non-African haplogroups falls into a time window of 47-52 kya (Figure S9). Altogether in our data we detect 32 branches that are older than 30 ky. In addition, we consider it plausible that at least 11 other "old" non-African branches have living descendants in extant populations. The geographic distribution of the extant members of these ~40 oldest out-of-Africa (OOA) branches of the Chr Y tree (Figure S9) is, however, more geographically balanced than the distribution of the four basic lineages of K

(Karafet et al. 2014): haplogroups G and IJ have their spread restricted predominantly to West Eurasia, H is restricted to South Asia, and only the newly defined F2 branch is potentially restricted to Southeast Asia. Considering the geographic distribution of the F, C, D derived clades in combination with the four primary subclades of K, overall, the Chr Y data appear to support the model by which the initial radiation of non-African genetic diversity was a rapid process involving almost simultaneous splits of the ancestors of West and East Eurasian, as well as Oceanian populations rather than supporting a model of a slow serial founder process by which the regional pools of founding haplogroups of populations more distant from Africa would be seen as nested subsets of populations living more closer geographically to Africa.

We report 44 additional SNPs that support the IJ branch, equivalent to M429, a group containing both the major European clade I and major Western Asian clade J. Consistent with previous studies (Rootsi et al. 2004; Underhill and Kivisild 2007), the distribution of haplogroup I subclades I1 and I2 is restricted to Europe in our sample, with a single Georgian sample belonging to the newly defined subclade I3 (formerly I*) (Figure 1, Figures S24-25). The I1 and I2’3 subclades have a coalescent time ~25-29 kya (Figure 1, Figure S25). While the I1 clade is defined by a single long branch and has a recent, late-Neolithic (~5 kya) coalescent date similar to other Y chromosome haplogroups common in Europe, such as R1b and R1a, clade I2`3 has a more complex structure with a wider distribution area and a number of internal branches coalescing at 15-25 kya. These age estimates are consistent with the results of ancient DNA studies showing that pre-Neolithic lineages of Northwest Europe belonged to the I2’3 branch of haplogroup I (Lazaridis et al. 2014; Skoglund et al. 2014). Considering that three out of four pre-Neolithic European I lineages belonged specifically to the I2a-M423 clade (Lazaridis et al. 2014), while not belonging to either of its extant L621 or L161 subclades it is worthwhile noticing that the age estimate of I2a-M423 drawn from our data is 13 kya (95% CI 11.5-14.5 kya) consistent broadly with the age of this clade as suggested by ancient DNA.

In our dataset haplogroup J combines three lineages, J1, J2a and J2b (Figure S26-27), with similar coalescent times at around 15 kya (Figure 1, Figure S3, Table S7).

We re-define haplogroup S, common in Papua New Guinea, by marker B254, allowing it to embrace branches formerly unresolved in K*. Haplogroup M and S form a monophyletic clade MS-PR2099, which is common and diverse in Island Southeast Asian populations (Figure S28).

In haplogroup O, common in Southeast and East Asia, our revised nomenclature recognizes seven basic subclades, O1-O7, that descend from the MRCA of O at ca 35 kya. O2a, which is common among Austro-Asiatic populations, is characterized by low diversity (MRCA 10.8, 95% CI 7.4-14.3 ky) and is composed of two major branches, O2a1 and O2a2 that are both 5-7 ky old (Figure S3, Table S7). Albeit significantly younger than earlier Y-STR based age estimates of 17-28 kya (Chaubey et al. 2011), the age estimates of O2a1 and O2a2 are broadly consistent with the dates associated with the spread of rice in SE Asia (Higham 2003; Bellwood 2005). In haplogroup Q, we define Q1c, represented by West Siberian and Central Asian samples, as a sister-clade of Q1a and Q1b, which are specific to Native Americans (Figure S35). In Q1a, we find 17 mutations that are equivalent to M3 in all Native American samples; two of these are in the ancestral state in a Siberian Eskimo. A newly defined clade, Q2b, joins 4 Koryak sequences with the Paleo-Eskimo Saqqaq (Rasmussen et al. 2010). The newly defined Q2b and formerly known Q2c-M120, represented in our data by a single sequence from Borneo, share a common ancestor (at around 15 kya) that is comparable in age to the MRCA of Q1a and Q1b. Rare haplogroup Q lineages in Europe and Nepal belong to clade Q1d which, again, has a pre-Holocene coalescent time.

Among the haplogroups that are characterized by high frequency but low sequence diversity, such as N throughout Northern Eurasia, R1a across Eurasia, especially South, Central and West Asia and Eastern Europe, R1b in Western Europe, we define a number of changes in haplogroup nomenclature. Our haplogroup N sample set belongs entirely to two basal branches – N1’3 (former N2; Phylotree) and N4 (Figures S29-31). In R1a, 68 out of 70 of our sequences fall into just two previously known (Pamjav et al. 2012; Underhill et al. 2014) basic clades, R1a1-Z283 and R1a2-Z93. Although our dataset is fairly large and originates from a wide geographic area, we detect no branches of R1a more basal than R-CTS4385 (van Oven et al. 2014), which we name R1a3, even though three more basal branches, R1a4'6, are assumed to exist from previous genotyping studies (Karafet et al. 2008; Underhill et al. 2010; Underhill et al. 2015). This makes the rapid (major inner branches are short) and widespread diversification of R1a all the more striking. We report a new subclade, R1a2a, found among Altay and Kyrgyz samples, that possibly represents the cryptic South Siberian branch of R1a defined earlier as R1a-Z93* (Underhill et al. 2014). In R1b, the simplified nomenclature recognizes 19 basic subclades (Figures S38-41), while in R1a, there are 6 (Figures S36-37).

References

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**(2): 147.

Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular biology and evolution* **29**(9): 2157-2167.

Bellwood PS. 2005. *First Farmers : the origins of agricultural societies*. Blackwell Pub., Malden, MA.

Chan KM, Moore BR. 2005. SYMMETREE: whole-tree analysis of differential diversification rates. *Bioinformatics* **21**(8): 1709-1710.

Chaubey G, Metspalu M, Choi Y, Magi R, Romero IG, Soares P, van Oven M, Behar DM, Rootsi S, Hudjashov G et al. 2011. Population Genetic Structure in Indian Austroasiatic Speakers: The Role of Landscape Barriers and Sex-Specific Admixture. *Molecular biology and evolution* **28**(2): 1013-1024.

Cruciani F, Trombetta B, Massaia A, Destro-Bisol G, Sellitto D, Scozzari R. 2011. A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. *Am J Hum Genet* **88**(6): 814-818.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature methods* **9**(8): 772.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution* **29**(8): 1969-1973.

Dulik MC, Zhadanov SI, Osipova LP, Askapuli A, Gau L, Gokcumen O, Rubinstein S, Schurr TG. 2012. Mitochondrial DNA and Y chromosome variation provides evidence for a recent common ancestry between Native Americans and Indigenous Altaians. *Am J Hum Genet* **90**(2): 229-246.

Excoffier L, Lischer HE. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources* **10**(3): 564-567.

Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, Atzeni R, Pilu R, Busonero F, Maschio A, Zara I et al. 2013. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science* **341**(6145): 565-569.

Ge J, Budowle B, Aranda XG, Planz JV, Eisenberg AJ, Chakraborty R. 2009. Mutation rates at Y chromosome short tandem repeats in Texas populations. *Forensic science international Genetics* **3**(3): 179-184.

Goedbloed M, Vermeulen M, Fang RN, Lembring M, Wollstein A, Ballantyne K, Lao O, Brauer S, Kruger C, Roewer L et al. 2009. Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR Yfiler PCR amplification kit. *International journal of legal medicine* **123**(6): 471-482.

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P. 1997. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Human molecular genetics* **6**(5): 799-803.

Higham C. 2003. Languages and Farming Dispersals: Austroasiatic Languages and Rice Cultivation. In *Examining the farming/language dispersal hypothesis*, (ed. P Bellwood, C Renfrew). The McDonald Institute for Archaeological Research, Cambridge.

Jarve M, Zhivotovsky LA, Rootsi S, Help H, Rogaev EI, Khusnutdinova EK, Kivisild T, Sanchez JJ. 2009. Decreased rate of evolution in Y chromosome STR loci of increased size of the repeat unit. *Plos One* **4**(9): e7276.

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res* **18**(5): 830-838.

Karafet TM, Mendez FL, Sudoyo H, Lansing JS, Hammer MF. 2014. Improved phylogenetic resolution and rapid diversification of Y-chromosome haplogroup K-M526 in Southeast Asia. *European journal of human genetics : EJHG*.

Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**(6082): 740-743.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Sudmant PH, Schraiber JG, Castellano S, Kirsanow K, Economou C et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**(7518).

Mendez FL, Krahn T, Schrack B, Krahn AM, Veeramah KR, Woerner AE, Fomine FL, Bradman N, Thomas MG, Karafet TM et al. 2013. An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. *Am J Hum Genet* **92**(3): 454-459.

Pamjav H, Feher T, Nemeth E, Padar Z. 2012. Brief communication: new Y-chromosome binary markers improve phylogenetic resolution within haplogroup R1a1. *Am J Phys Anthropol* **149**(4): 611-615.

Poznik GD, Henn BM, Yee MC, Sliwerska E, Euskirchen GM, Lin AA, Snyder M, Quintana-Murci L, Kidd JM, Underhill PA et al. 2013. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**(6145): 562-565.

R-Core-Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW, Jr., Rasmussen S, Moltke I, Albrechtsen A, Doyle SM et al. 2014. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**(7487): 225-229.

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R et al. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**(7282): 757-762.

Regueiro M, Cadenas AM, Gayden T, Underhill PA, Herrera RJ. 2006. Iran: tricontinental nexus for Y-chromosome driven migration. *Hum Hered* **61**(3): 132-143.

Rootsi S, Magri C, Kivisild T, Benuzzi G, Help H, Bermisheva M, Kutuev I, Barac L, Pericic M, Balanovsky O et al. 2004. Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in europe. *Am J Hum Genet* **75**(1): 128-137. Epub 2004 May 2025.

Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE, Lin AA, Mitra M, Sil SK, Ramesh A et al. 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.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**(6942): 825-837.

Skoglund P, Malmstrom H, Omrak A, Raghavan M, Valdiosera C, Gunther T, Hall P, Tambets K, Parik J, Sjogren KG et al. 2014. Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* **344**(6185): 747-750.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl 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.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**(21): 2688-2690.

Underhill PA, Kivisild T. 2007. Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations. *Annu Rev Genet* **41**: 539-564.

Underhill PA, Myres NM, Rootsi S, Metspalu M, Zhivotovsky LA, King RJ, Lin AA, Chow CE, Semino O, Battaglia V et al. 2010. Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. *European journal of human genetics : EJHG* **18**(4): 479-484.

Underhill PA, Poznik GD, Rootsi S, Jarve M, Lin AA, Wang J, Passarelli B, Kanbar J, Myres NM, King RJ et al. 2015. The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. *European journal of human genetics : EJHG* **23**(1): 124-131.

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human mutation* **30**(2): E386-394.

van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MH. 2014. Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. *Human mutation* **35**(2): 187-191.

Vianello D, Sevini F, Castellani G, Lomartire L, Capri M, Franceschi C. 2013. HAPLOFIND: a new method for high-throughput mtDNA haplogroup assignment. *Human mutation* **34**(9): 1189-1194.

Wei W, Ayub Q, Chen Y, McCarthy S, Hou Y, Carbone I, Xue Y, Tyler-Smith C. 2013a. A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Res* **23**(2): 388-395.

Wei W, Ayub Q, Xue Y, Tyler-Smith C. 2013b. A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping. *Forensic science international Genetics* **7**(6): 568-572.

Wickham H. 2009. *Ggplot2 : elegant graphics for data analysis*. Springer, New York.

Wong LP, Ong RT, Poh WT, Liu X, Chen P, Li R, Lam KK, Pillai NE, Sim KS, Xu H et al. 2013. Deep whole-genome sequencing of 100 southeast Asian Malays. *Am J Hum Genet* **92**(1): 52-66.

Xue Y, Wang Q, Long Q, Ng BL, Swerdlow H, Burton J, Skuce C, Taylor R, Abdellah Z, Zhao Y et al. 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr Biol* **19**(17): 1453-1457.

Y Chromosome Consortium 2002. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res* **12**(2): 339-348.

Zhivotovsky LA, Underhill PA, Cinnioglu C, Kayser M, Morar B, Kivisild T, Scozzari R, Cruciani F, Destro-Bisol G, Spedini G et al. 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.