

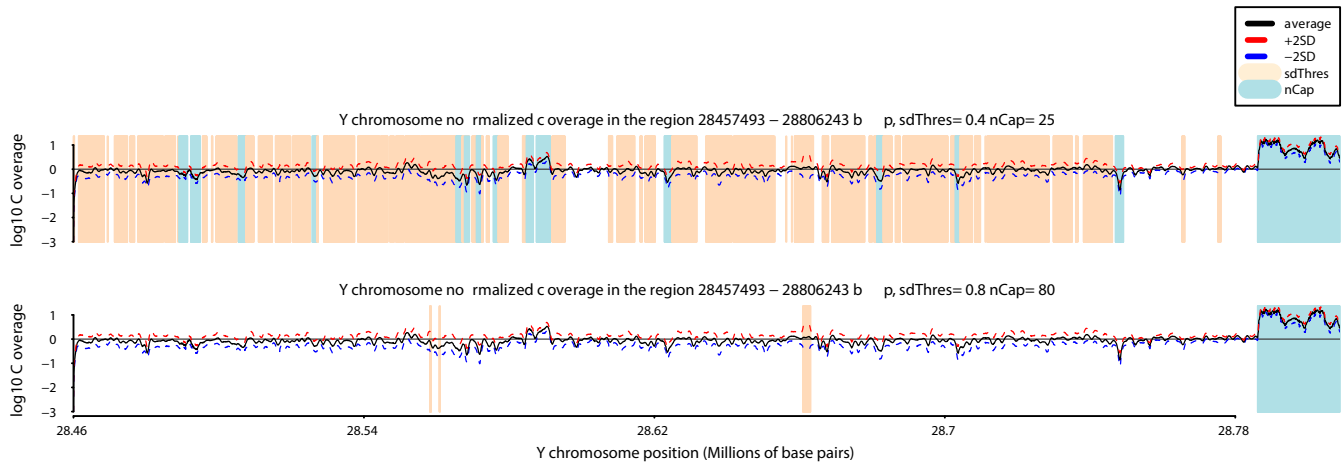


Figure S1. An example of a Y chromosome region filtering by local unique sequence coverage (UC). Combinations of filtering parameters (sdThres/nCap) applied to the same chrY region – upper panel 0.4/25; lower panel 0.8/80. In a sample of 307 Y chromosomes we normalized each individual's average UC in 1000 base pair (bp) windows (sliding by 50 bps) to that of the individual's average UC on chrX, then calculated the log10 average and standard deviation (SD) across all individuals. We then established two parameters for excluding sequence stretches of poor unique coverage: 1) sdThres – wide SD margins (exceeding a set value) indicate that coverage varies between individuals in the region; 2) nCap – very low or very high coverage ($+2 \text{ SD} < 0$ or $-2 \text{ SD} > 0$) in a set number of consecutive windows suggests a deletion/duplication in the region.

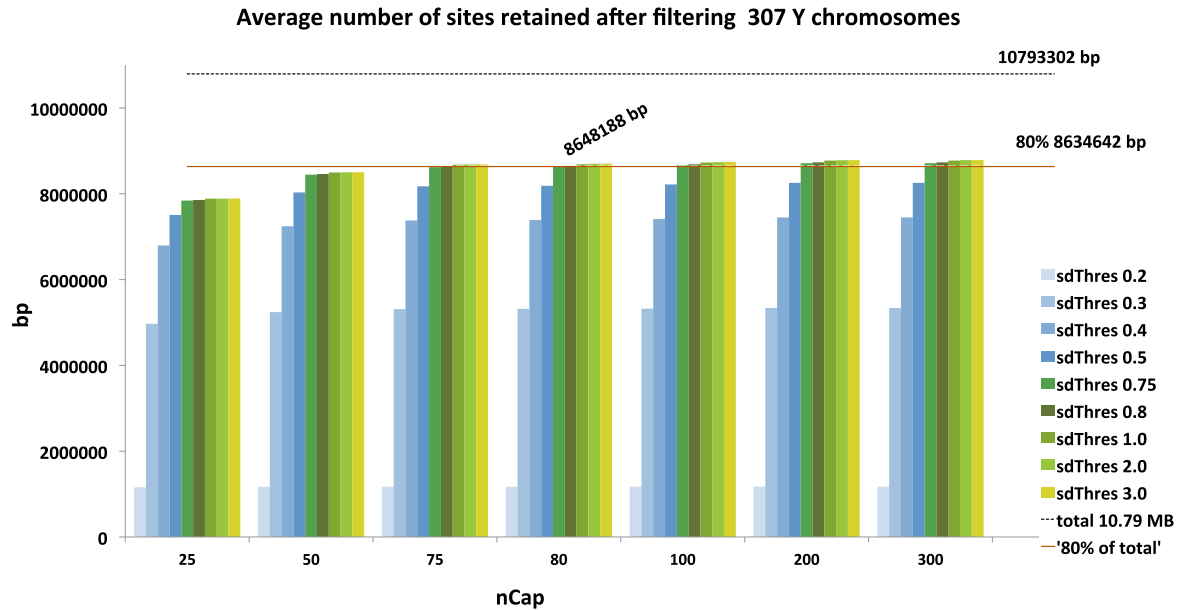


Figure S2. The effect of different combinations of nCap/sdThres on the average number of sites retained. The effect of different combinations of nCap/sdThres (filter b) on the sequence length (average number of sites) retained. Within 10.79 MB of chrY regions we applied a re-mapping filter, removed individual bases with less than 5x unique coverage and then tested the effect of nCap/sdThres combinations on the average number of nucleotides in the sample of 307 Ychromosomes for which coverage data was available. The combination of these two parameters reaches a plateau approximately at nCap=75/sdThres=0.75 where the gains in retaining the sequence length from further relaxing the parameters is negligible – the difference between nCap100/sdThres1.0 and nCap300/sdThres 3.0 is only 0.01%. To retain at least 80% of the original data we picked a point between nCap75/sdThres0.75 and nCap100/sdThres1.0 – and made final analyses with the parameters nCap80/sdThres0.8. Dashed line marks the 10.79MB, orange line marks the point at which 80% of the original sequence length was retained.

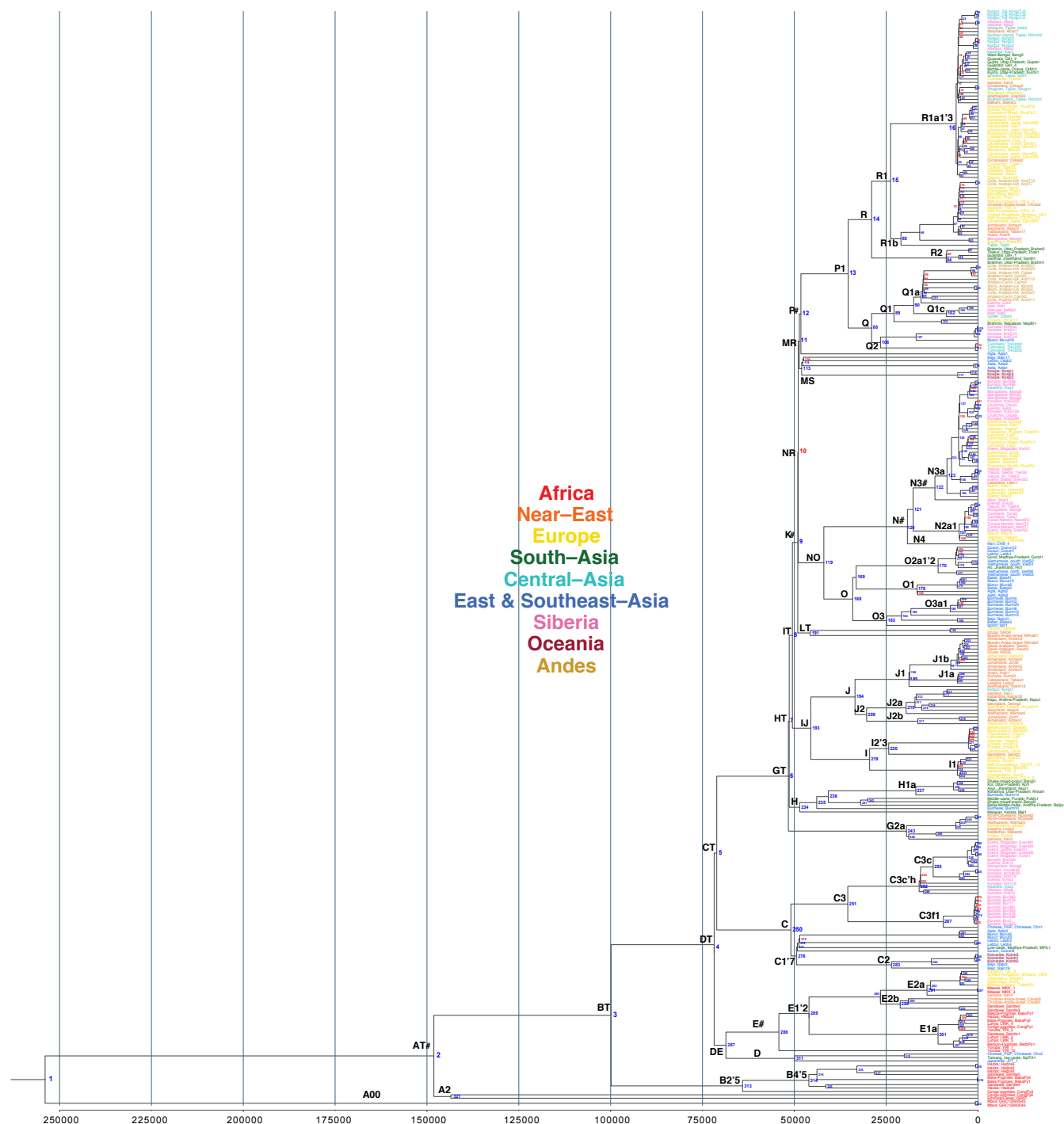


Figure S3. Y chromosome phylogeny based on BEAST analyses. Sample names correspond to those reported in Table S6. Age and posterior support estimates for each numbered clade are reported in Table S7. Six clades that showed significant ($p_{\Delta 1} < 0.05$) phylogenetic asymmetry in SymmeTree analyses are highlighted by # after the haplogroup label.

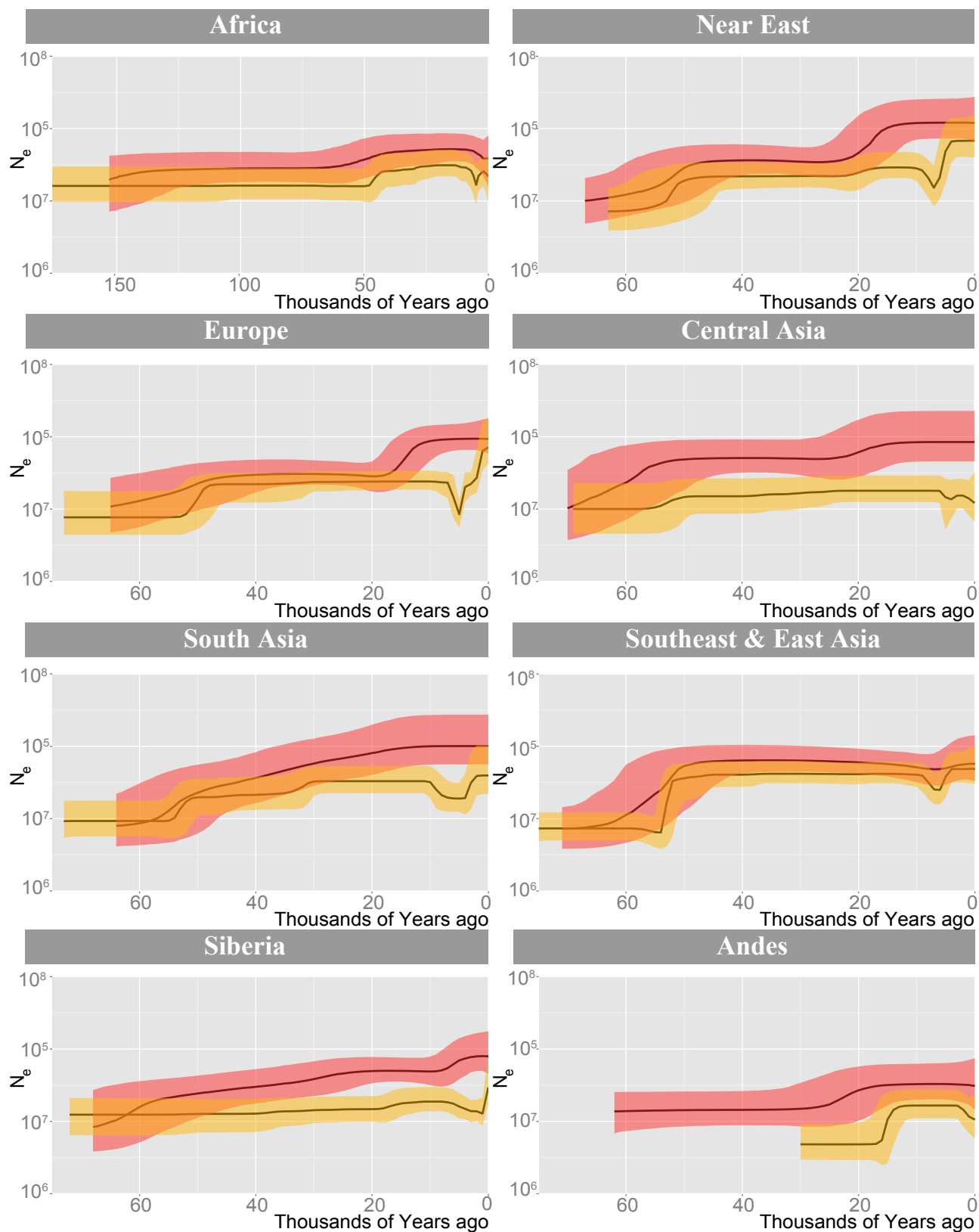


Figure S4A. NRY and mtDNA Bayesian Skyline Plots for each geographical region.

The 95% CI for mtDNA BSPs are represented in red and for NRY in yellow. The BSPs were created using a piecewise-linear coalescence model.

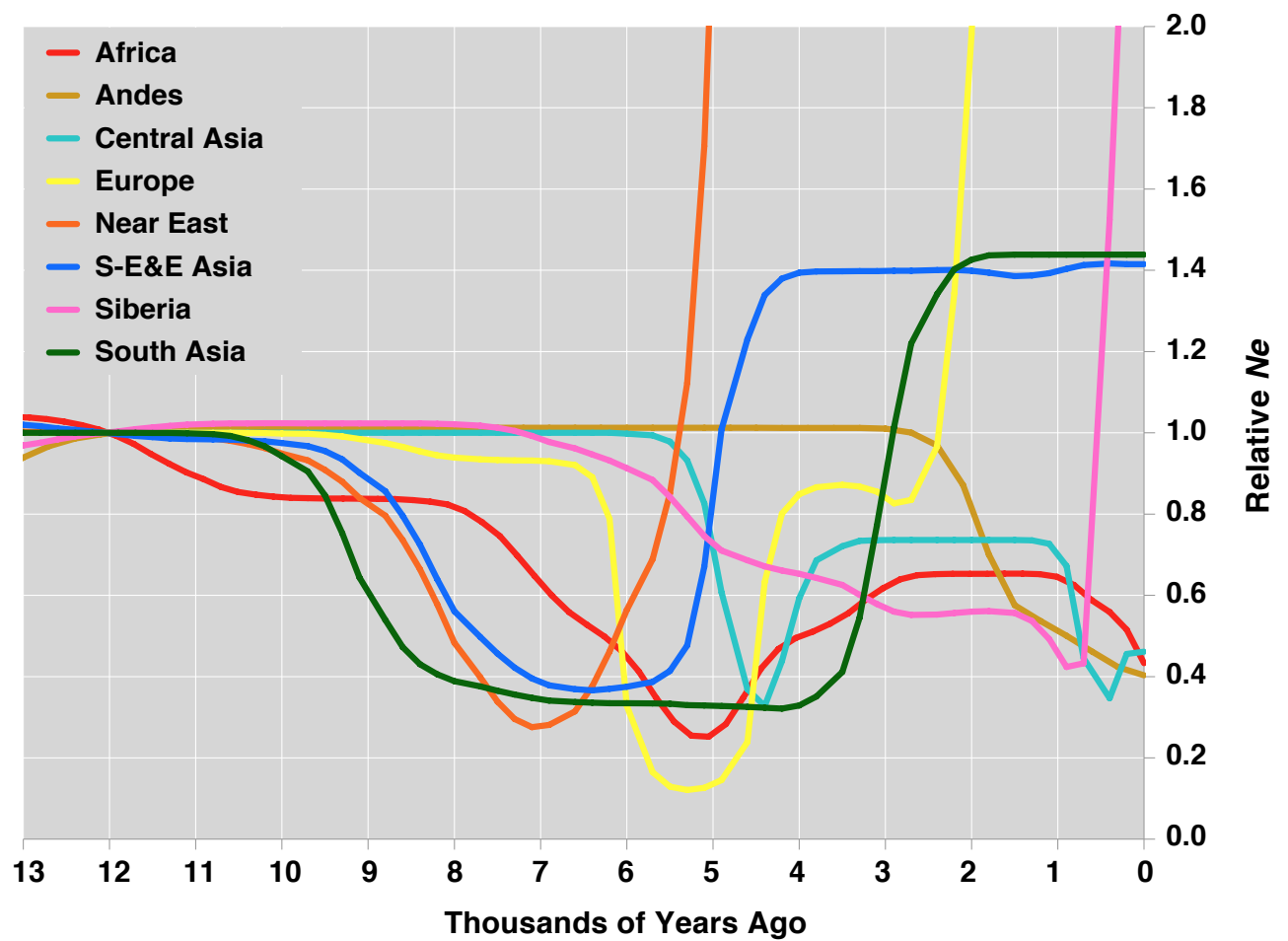


Figure S4B. Variation of Y chromosome effective population size over the past 13 kya. The regional NRY BSPs presented in Figure S4A are shown relative to their values at 12 kya.

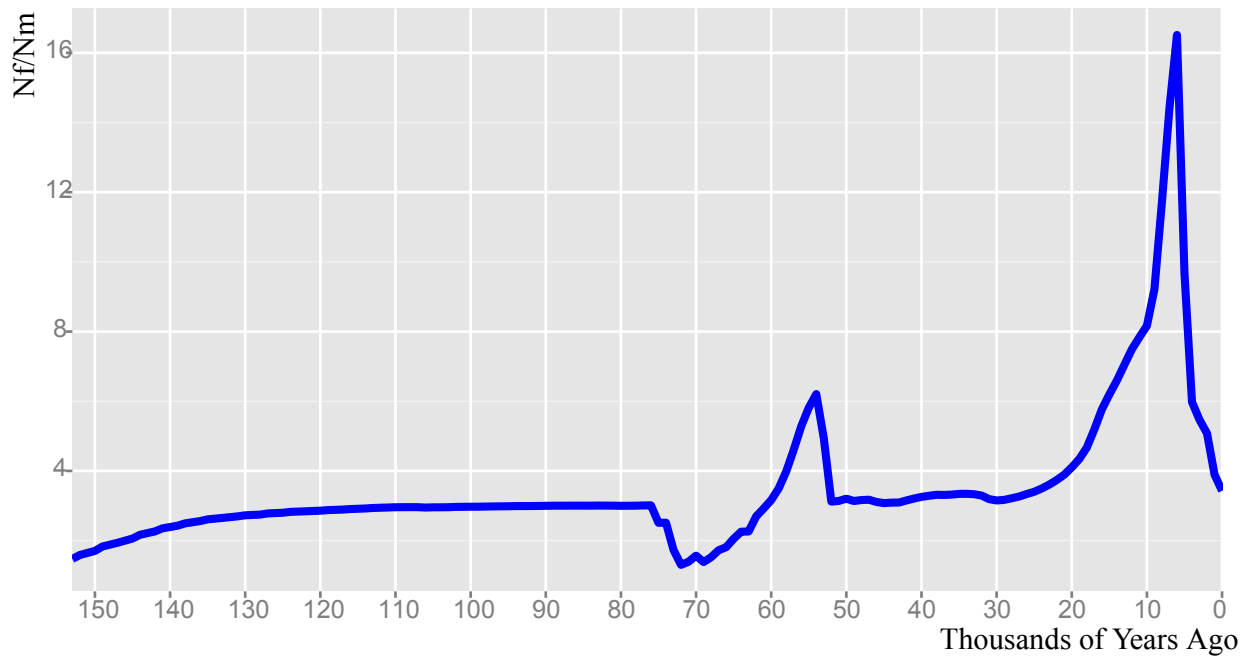


Figure S5 – The temporal dynamics of the ratio of female (N_f) and male (N_m) effective population size in the last 140KY. The ratios of the global accumulative N_e estimates of mtDNA (N_f) and Y chromosome (N_m), as presented in Figure 2, are plotted against the time (in thousands of years) back from the present (0). The BSPs estimates of N_e were obtained in BEAST using a piecewise-linear coalescence model.

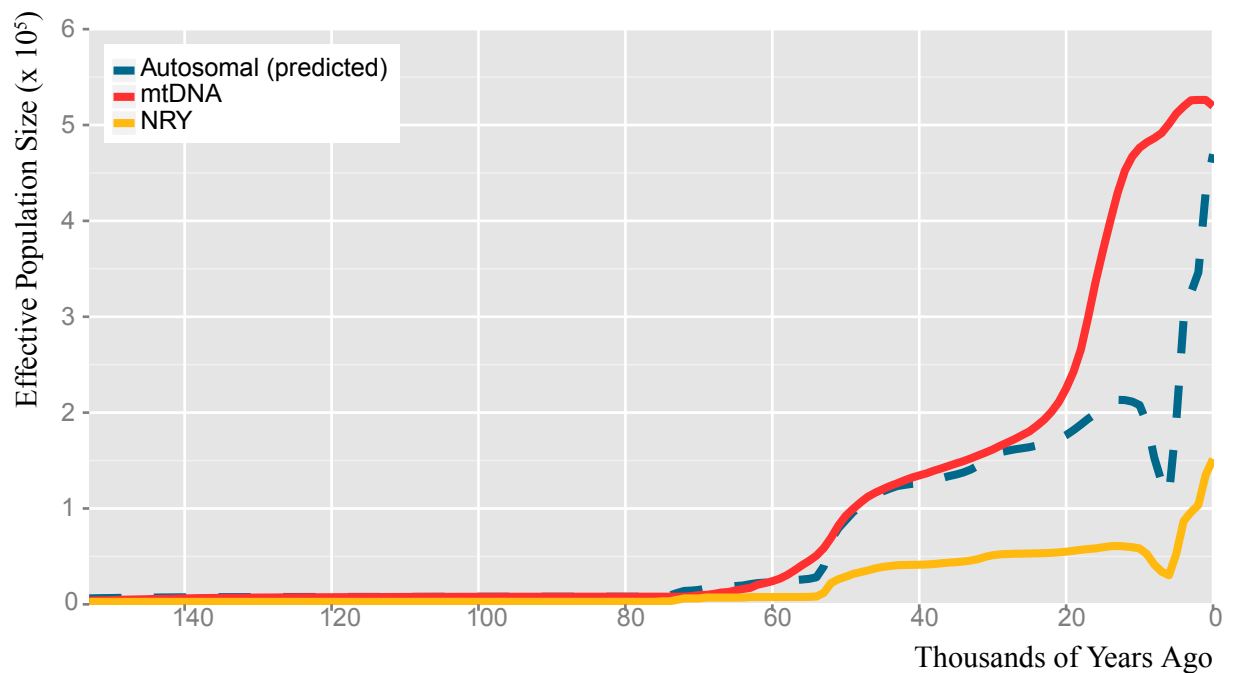


Figure S6 – The temporal dynamics of the autosomal effective population size in the last 140KY as predicted from female (N_f) and male (N_m) effective population sizes. The autosomal estimates were derived applying the $4N_fN_m/(N_f+N_m)$ formula on the mtDNA and Y chromosome based estimates.

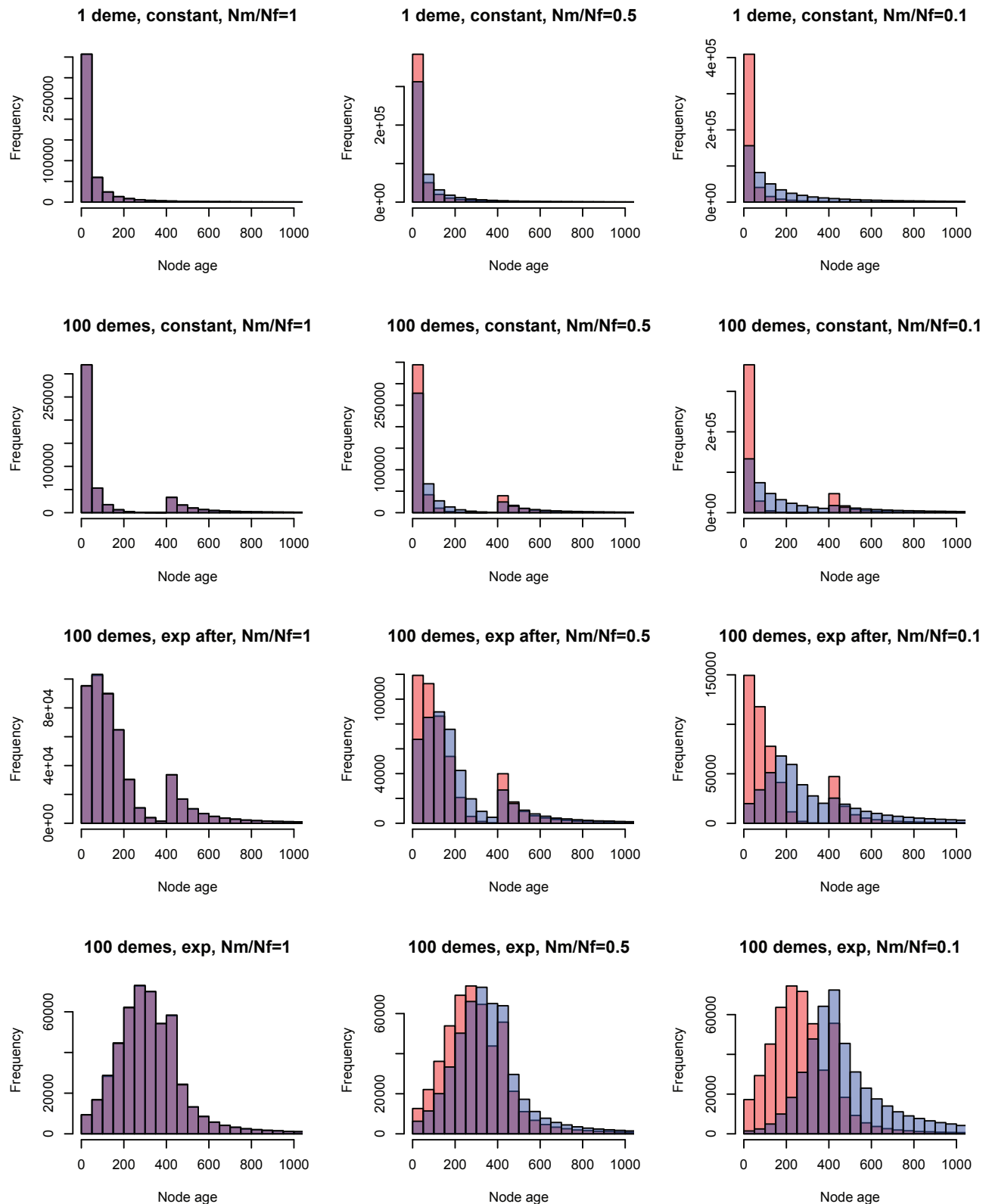


Figure S7 - Distribution of node ages for 1000 coalescent simulation replicates of Y chromosome (red) and mtDNA (blue) lineages under varying scenarios of population growth, deme formation, and effective number of males (N_m) and females (N_f).

Plots show histograms of node ages, measured in generations, from fastsimcoal2 coalescent simulations run for four scenarios: Row 1 - one deme with a constant population size; Row 2 - 1 deme divided into 100 equally-sized demes formed 400 generations ago with a constant population size; Row 3 - formation of 100 demes 400 generations ago, with exponential growth starting 200 generations ago; Row 4 - formation of 100 demes 400 generations ago, with exponential growth starting 400 generations ago. The first column shows simulations run, assuming the effective number of males relative to the effective number of females (N_m/N_f) is equal. The second column shows simulations where $N_m/N_f = 0.5$, and the third column shows simulations with an extreme $N_m/N_f=0.1$. N_m and N_f are computed assuming a constant autosomal effective size of 10,000. All simulations sample 500 individuals, distributed evenly between demes.

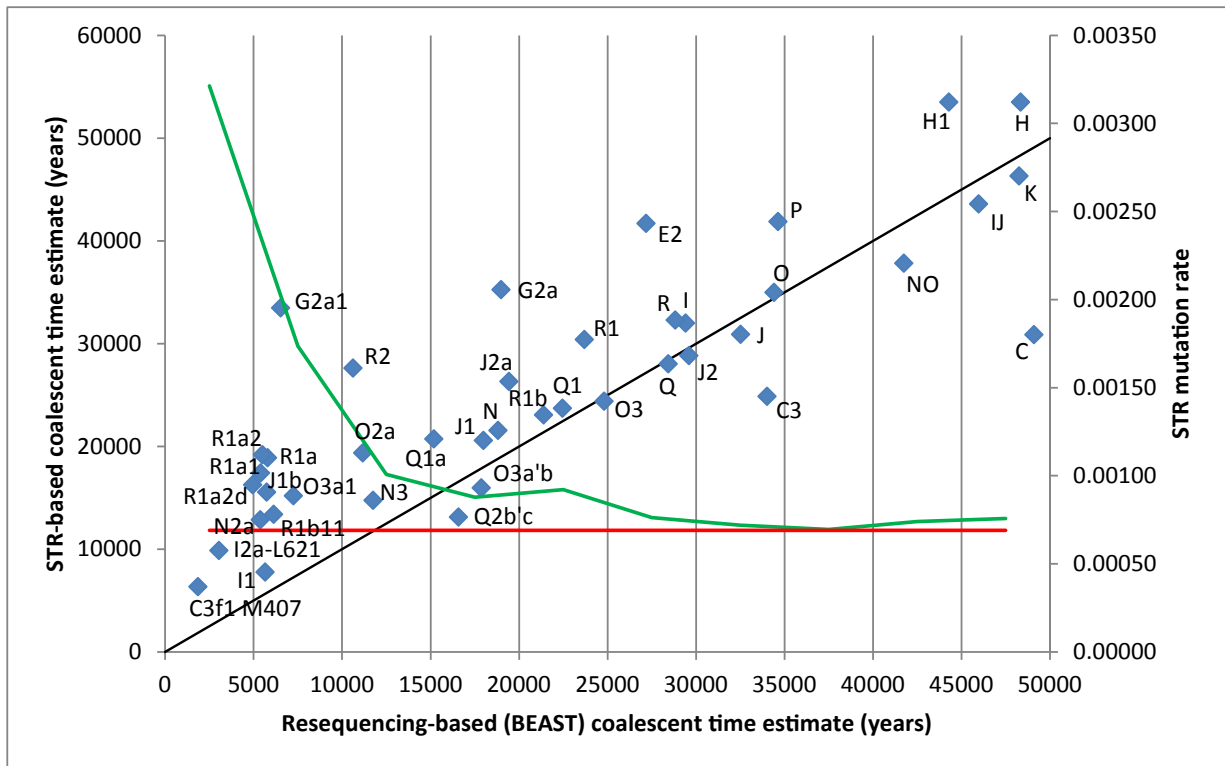


Figure S8. Comparison of STR and sequence data based coalescent time estimates.

STR-based estimates: 21 STR markers, “evolutionary” mutation rate 6.9×10^{-4} per 25 years (Zhivotovsky et al. 2004). Sequence based estimates: BEAST v.1.8.0, GTR substitution model, lognormal relaxed clock, mutation rate 7.4×10^{-10} mut/bp/year (SI3). The majority of the STR-based age estimates of “young” haplogroups deviate from the expected linear relationship (black line) with age estimates derived from sequence data. Adjusted STR mutation rates (green line, and the right y-axis) were calculated by 5000 year bins of haplogroup ages assuming linear relationship with sequence data based haplogroup age estimates. The “evolutionary” STR mutation rate that was used to calculate the raw STR-based age estimates is shown as a red line.

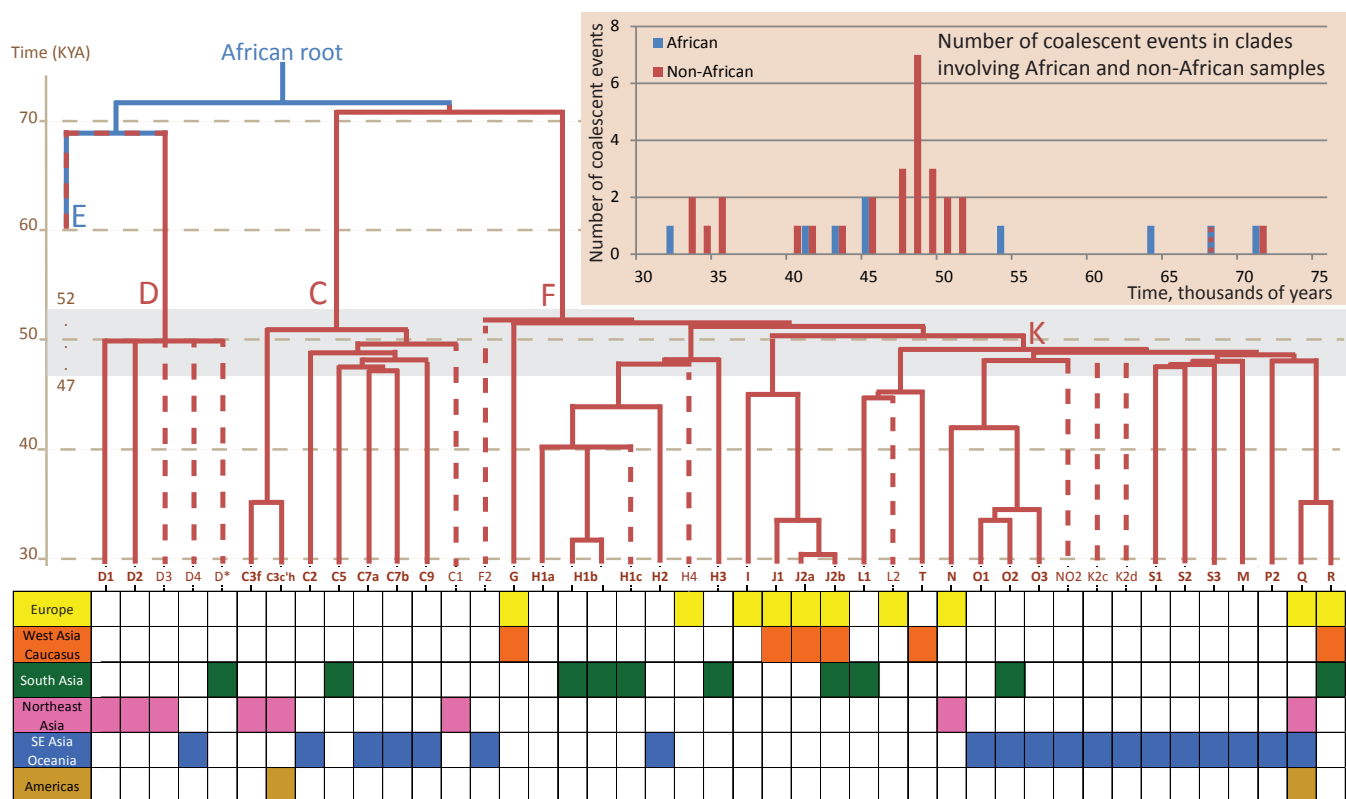


Figure S9. Coalescent times of the oldest Y chromosome haplogroups outside

Africa. The tree was generated and coalescent times (presented in thousands of years, KY) calculated in BEAST. Shown are only those 32 branches that had coalescent date older than 30 KY. These branches are shown in solid lines. The plausible branching structure of 11 additional branches that have been reported in the literature is shown in dotted lines. The distribution of relevant clades by geographic regions is presented on the basis of our data and inferences made from literature ignoring cases of the presence of the haplogroup in a region that can be explained by historic gene flow. The distribution of coalescent events in time for African and non-African populations was obtained using a sliding window of 2000 and step size 1000 years. The red and blue line connecting D and E reflects the proposed model of an early back to Africa event involving putative DE* lineages (Hammer et al. 1998).

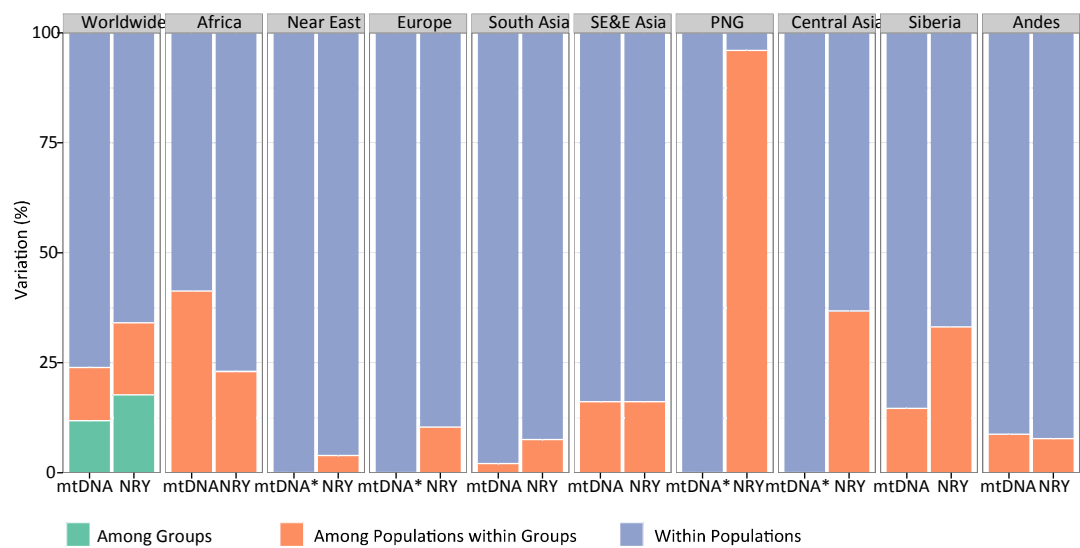


Figure S10. Distribution Y chromosome and mitochondrial diversity within and among populations.

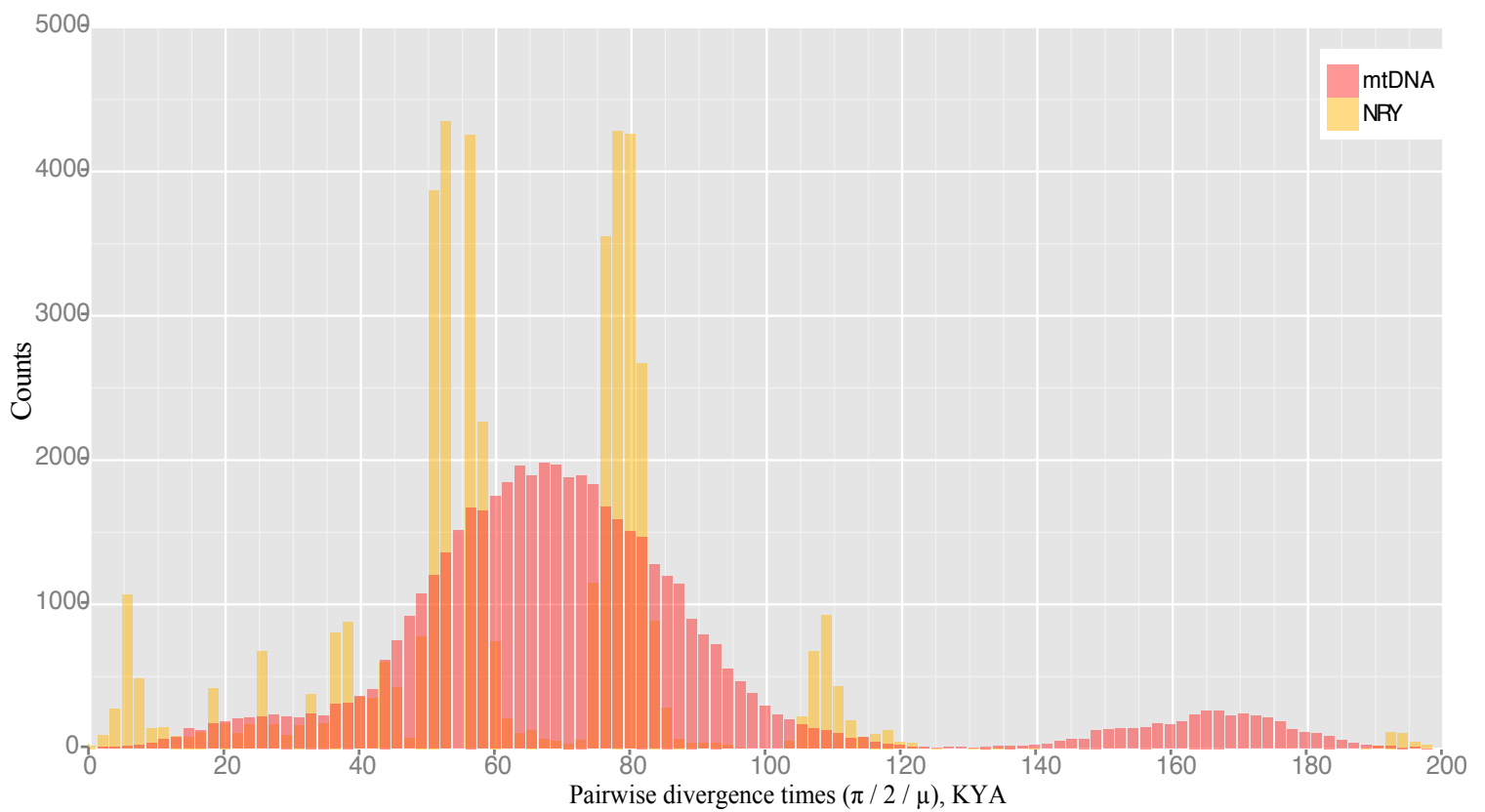


Figure S11. Distribution Y chromosome and mitochondrial DNA pairwise divergence times. The plot is based on the analysis of 320 mtDNA and Y chromosome sequences including African and non-African populations. The two highest peaks in the Y chromosome plot correspond to the two clusters of coalescent nodes in FigureS9, one at 43-48KYA involving non-African coalescent events within haplogroups C, D, and F, and the 62-64KYA cluster to the coalescent events among the non-African founders and between African E lineages and non-Africans.

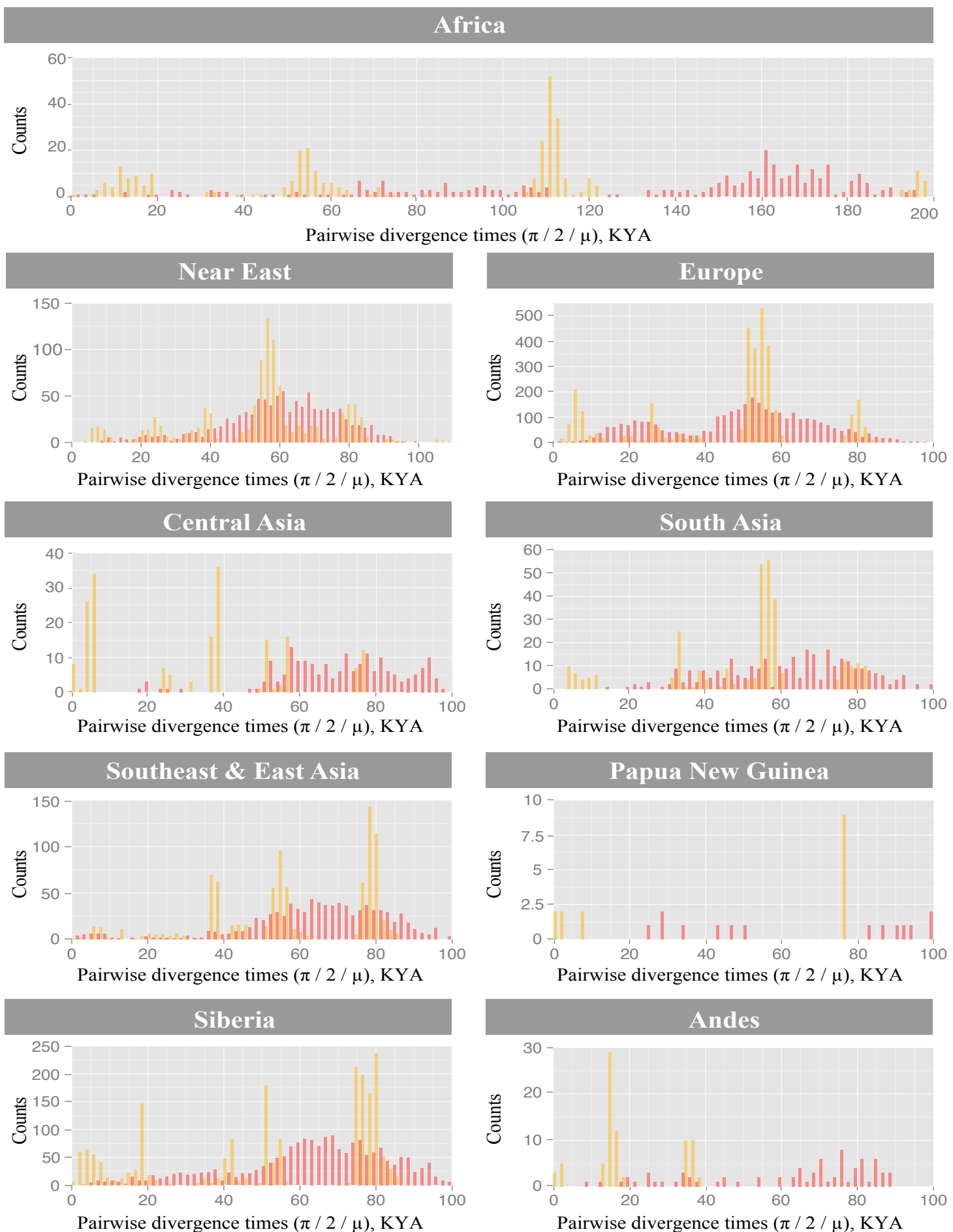


Figure S12. Regional plots of Y chromosome and mitochondrial DNA pairwise divergence times. The plot is based on the regional analysis of the 320 mtDNA and Y chromosome sequences as shown combined in Figure S11.

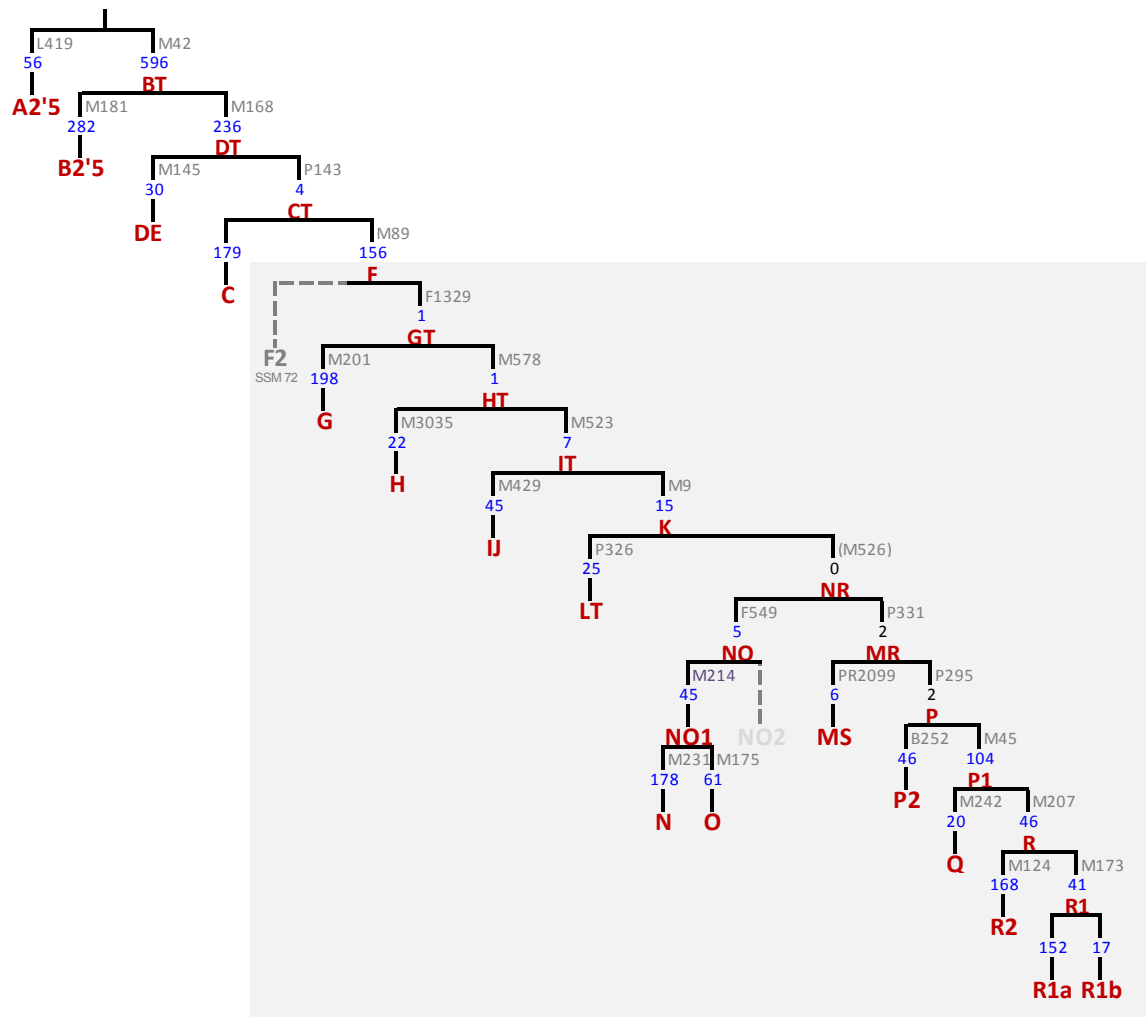


Figure S13. Refined topology of the basic branches of the Y-chromosome tree with an emphasis on the short branch lengths defining haplogroup F sub-clades. Mutation counts and branch defining markers (included in counts) are shown on the branches. If no previously named mutation existed on a given branch, a name from the B-series was given to one of the mutations discovered, and the rest were denoted with _eq in the annotations table. Branch lengths include recurrent mutations if such are present in no more than two (exceptionally three) locations in the Y-chromosome phylogeny. Given comparative data from Malaysian high coverage sequences (Wong et al. 2013), we map the position of the F2 clade represented by Malay sample SSM072 and redefine haplogroup NO by F549 and 4 equivalent mutations. NO splits further into NO1-M214 and the novel NO2-F2755, as defined by the SSM016 sequence.

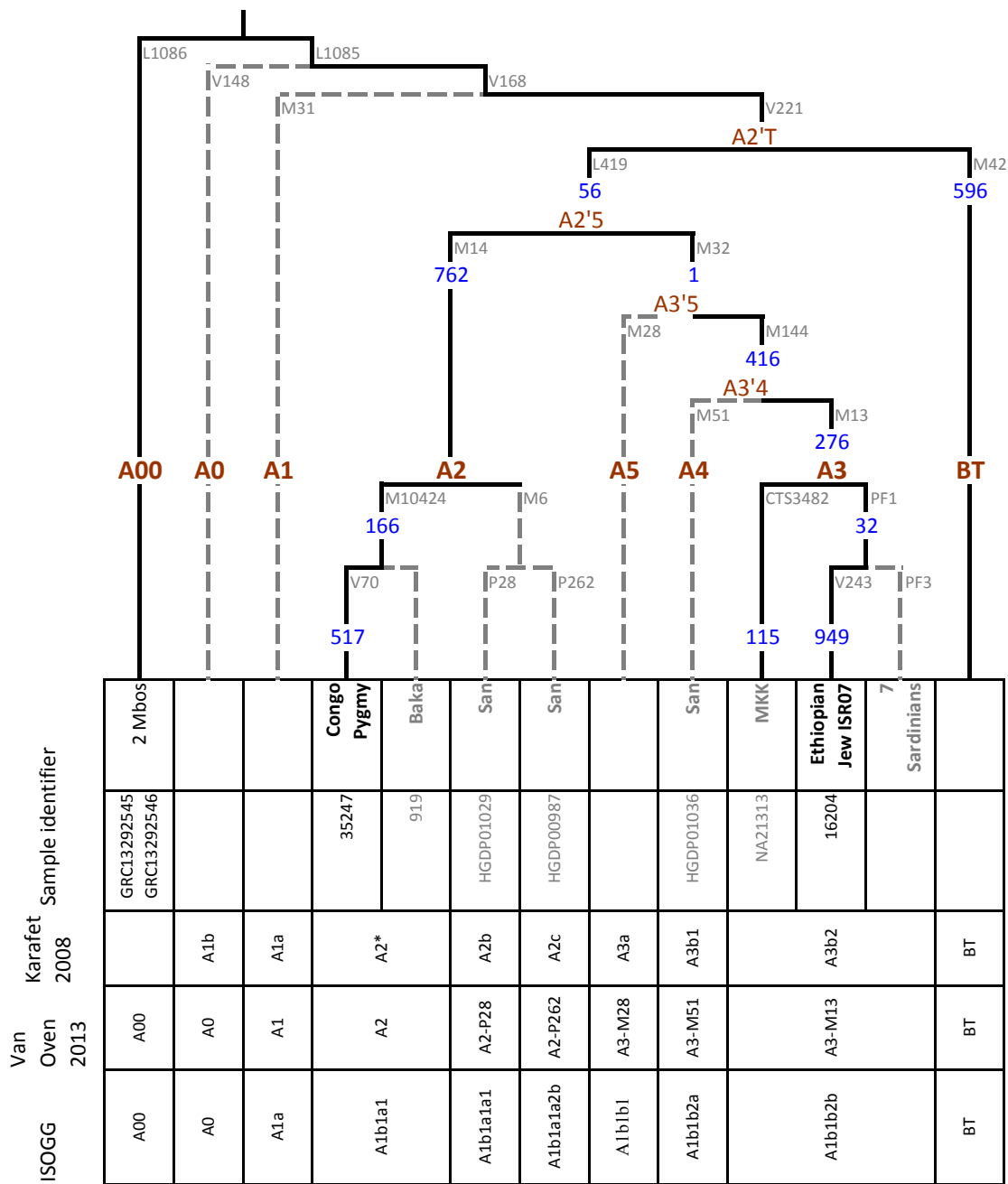


Figure S14. Refined topology of the Y-chromosome haplogroup A. Grey dashed lines denote branches defined by published sequences generated with Illumina technology.

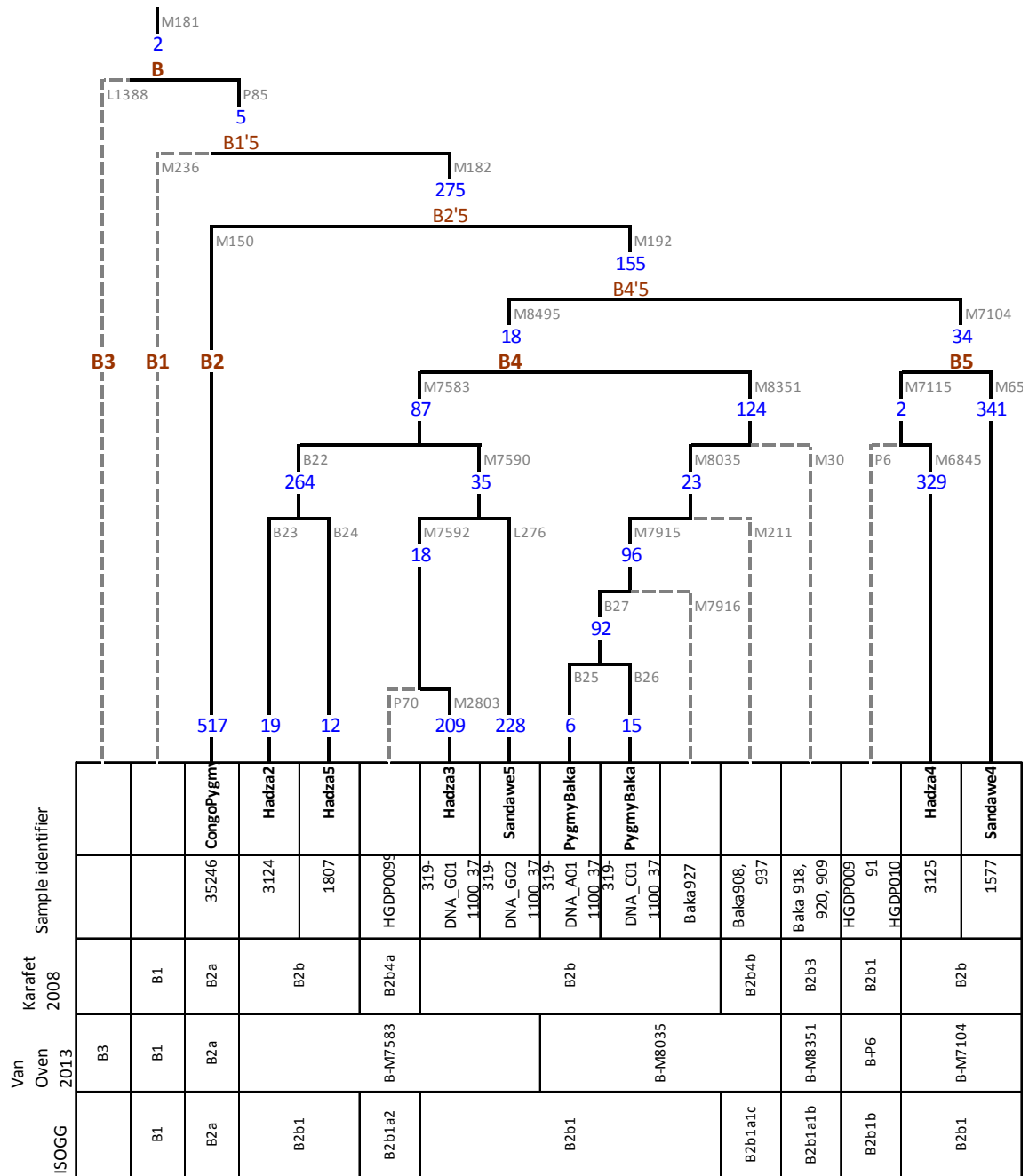


Figure S15. Refined topology of the Y-chromosome haplogroup B. Grey dashed lines denote branches defined by published sequences generated with Illumina technology.

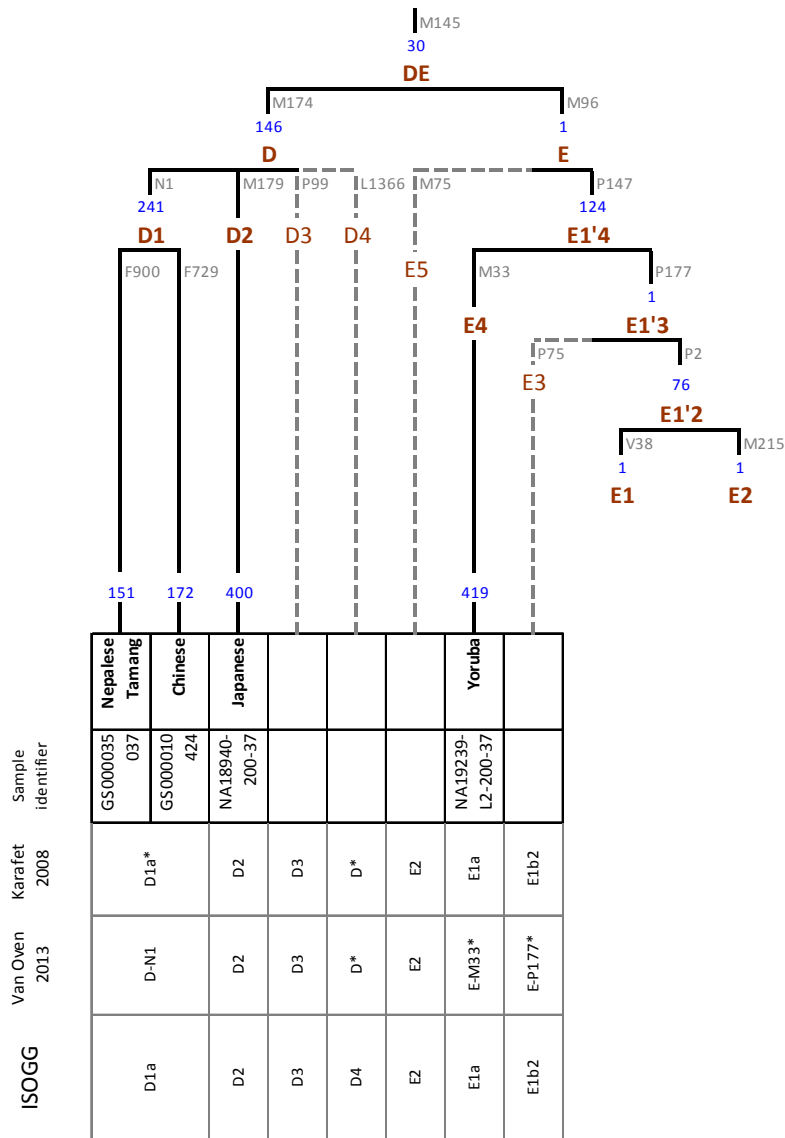


Figure S16. Refined topology of the Y-chromosome haplogroup DE. Grey dashed lines denote branches deriving from the respective nodes in previous phylogenetic reports but not found in our sample (Karafet et al 2008; van Oven et al 2013). Branches E1, E2, E1'3 are shown with length 1 because only one of the sub-branches was present in our sample and the rest of the identified mutations were counted to be equal with the sub-branch definition although some of them are actually defining the parent-clade.

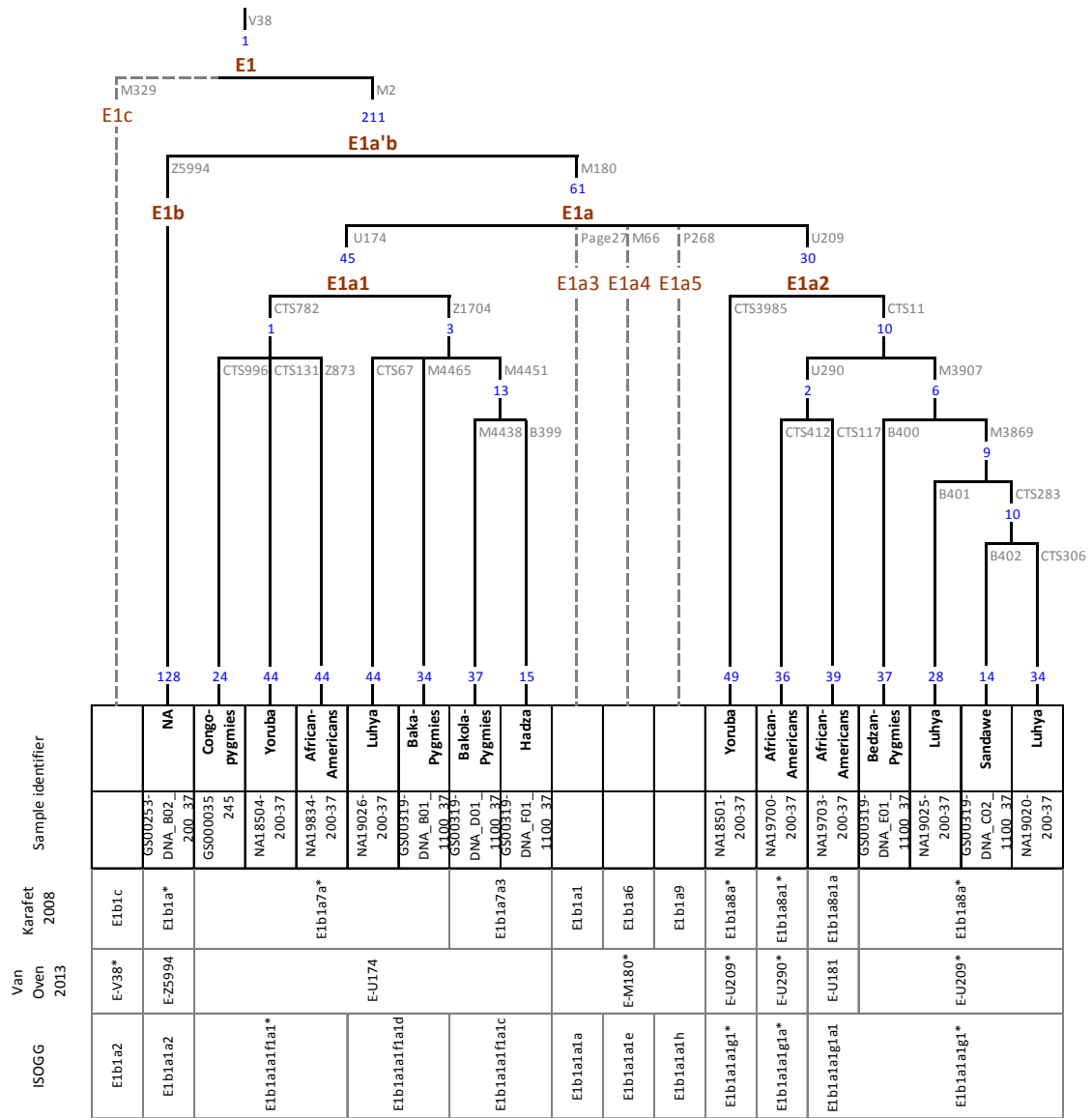


Figure S17. Refined topology of the Y-chromosome haplogroup E1. Grey dashed lines denote known branches with no sequence data and branches based on low coverage sequencing.

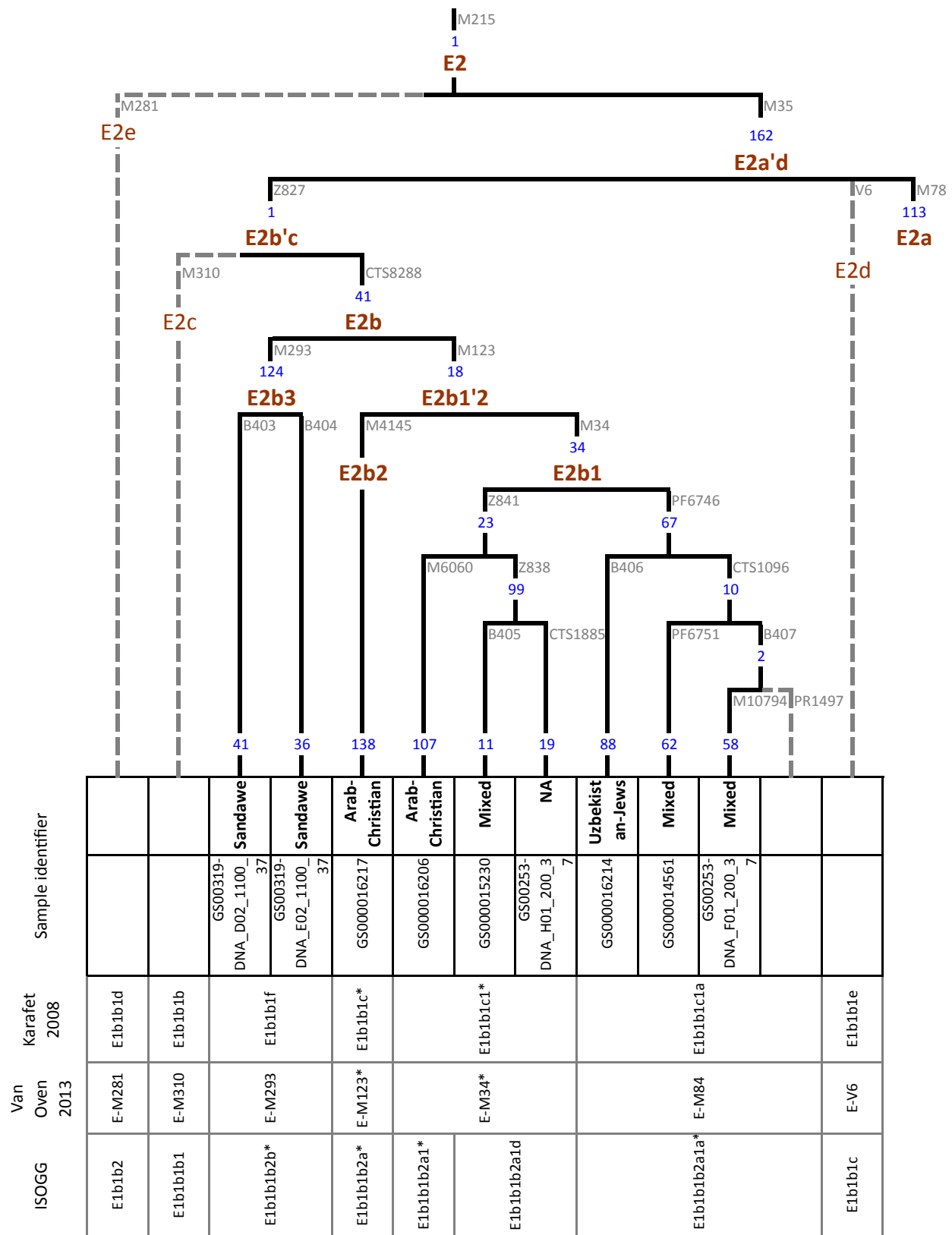


Figure S18. Refined topology of the Y-chromosome haplogroup E2. Grey dashed lines denote known branches with no sequence data and branches based on low coverage sequencing.

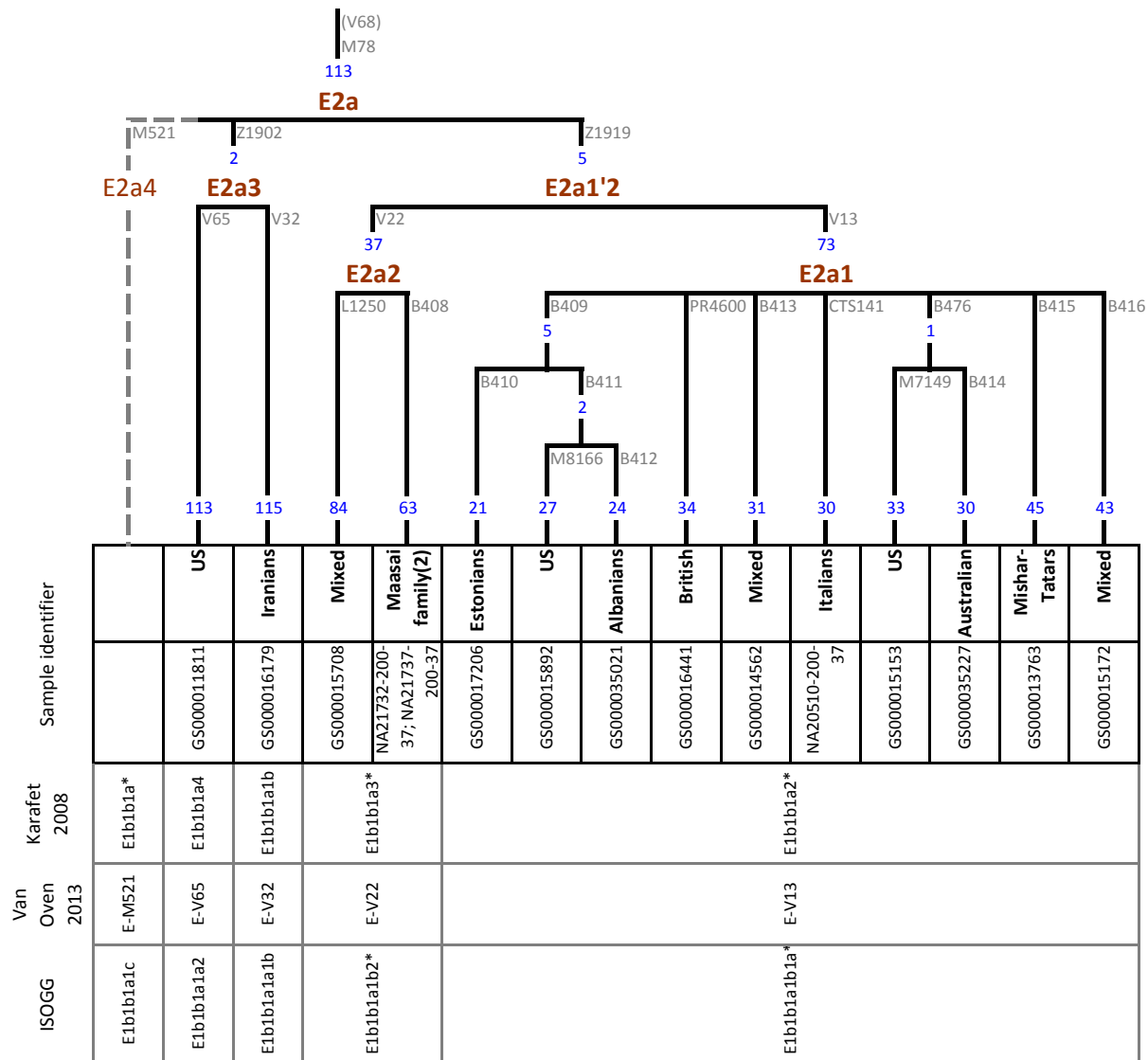


Figure S19. Refined topology of the Y-chromosome haplogroup E2a. Grey dashed lines denote known branches with no sequence data and branches based on low coverage sequencing.

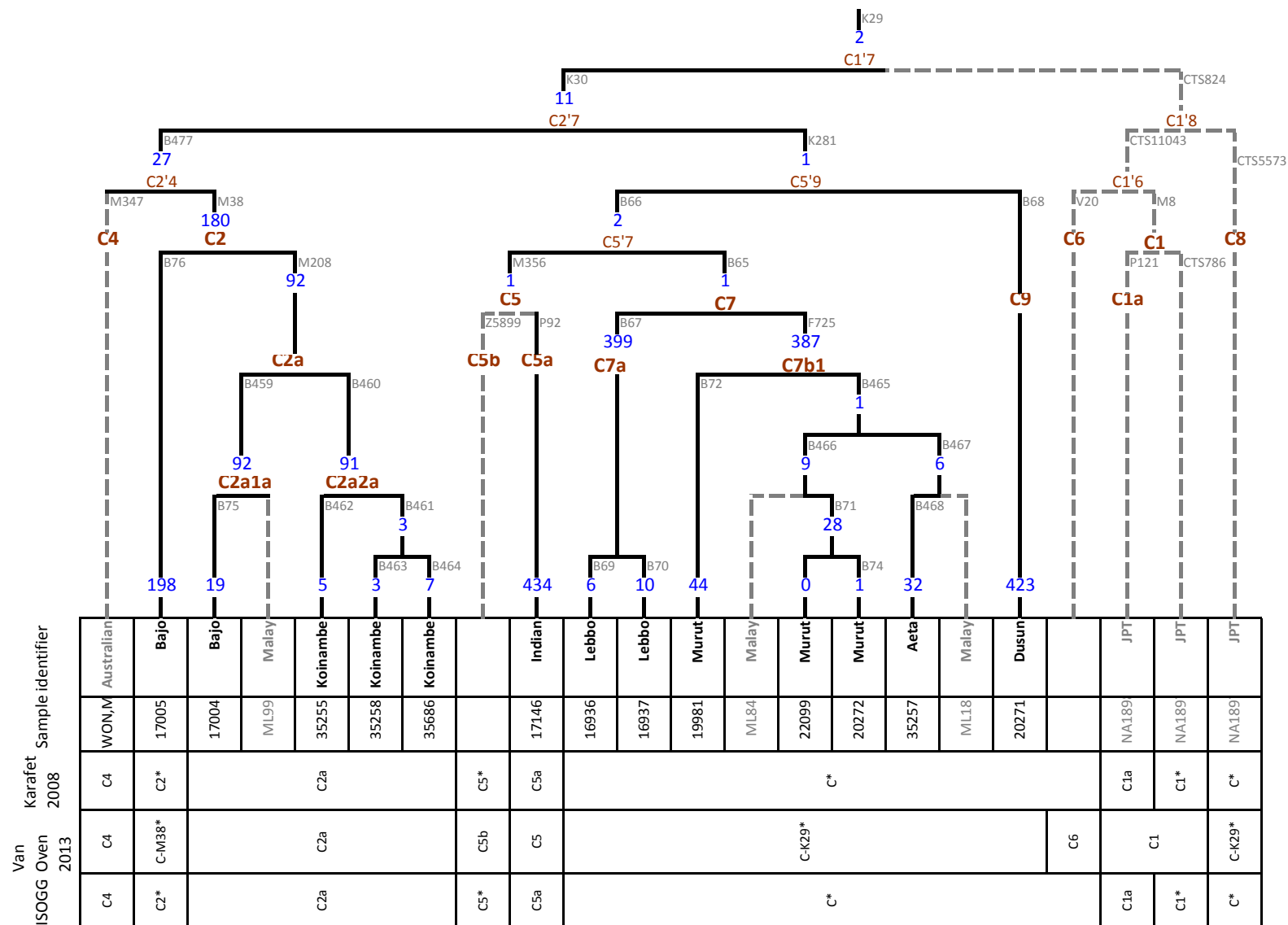


Figure S21. Refined topology of the Y-chromosome haplogroup C1'7. Grey dashed lines denote known branches with no sequence data and branches based on low coverage sequencing.

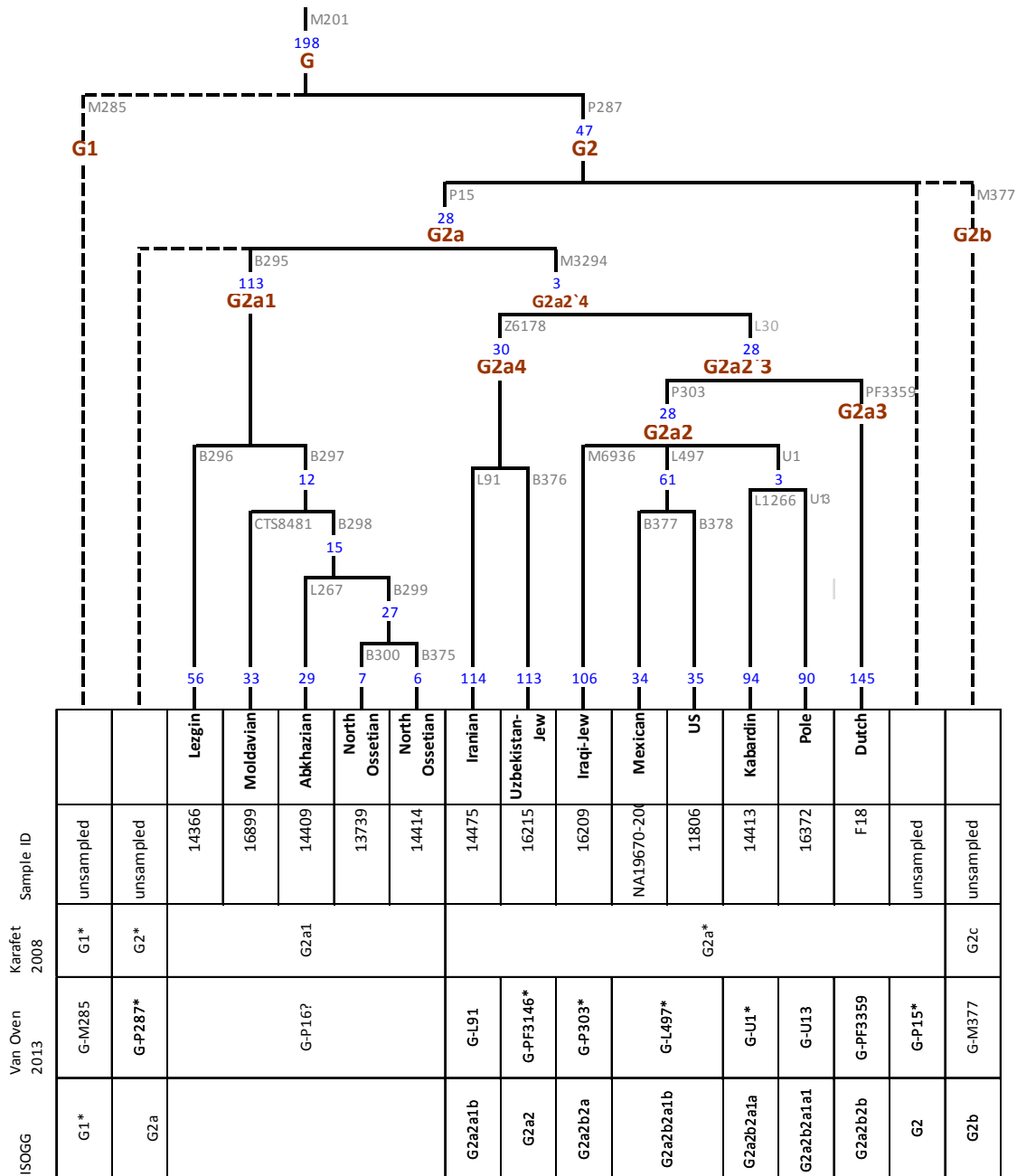


Figure S22. Refined topology of the Y-chromosome haplogroup G. Grey dashed lines denote known branches with no sequence data in our dataset. In the case of haplogroup G all samples generated for this project belonged to clade G2a. Within this clade we distinguish four sub-clades, including G2a1-B295, which was earlier defined by a recurrent marker P16.

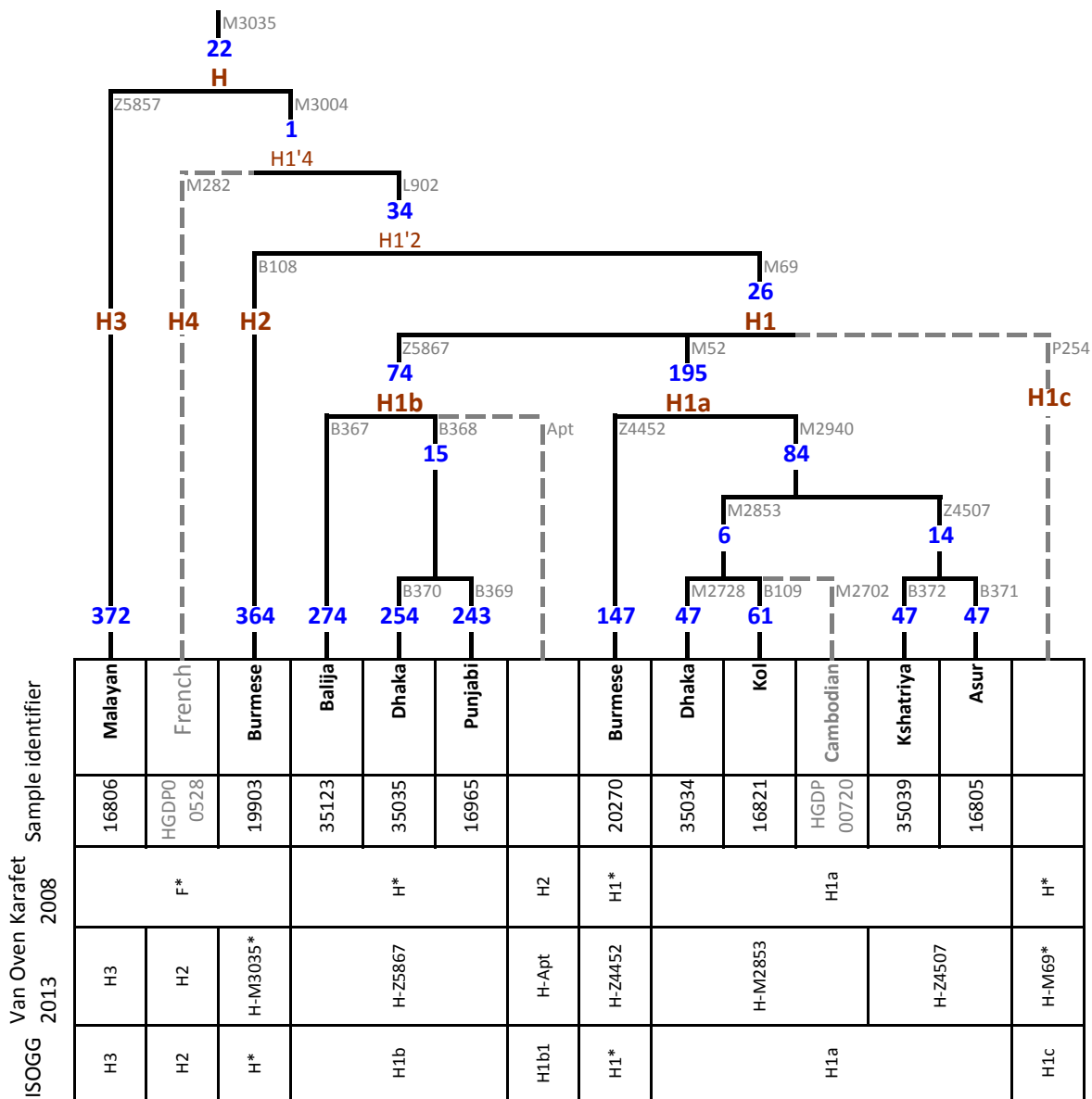


Figure S23. Refined topology of the Y-chromosome haplogroup H. Grey dashed lines denote known branches with no sequence data and branches based on low coverage sequencing.

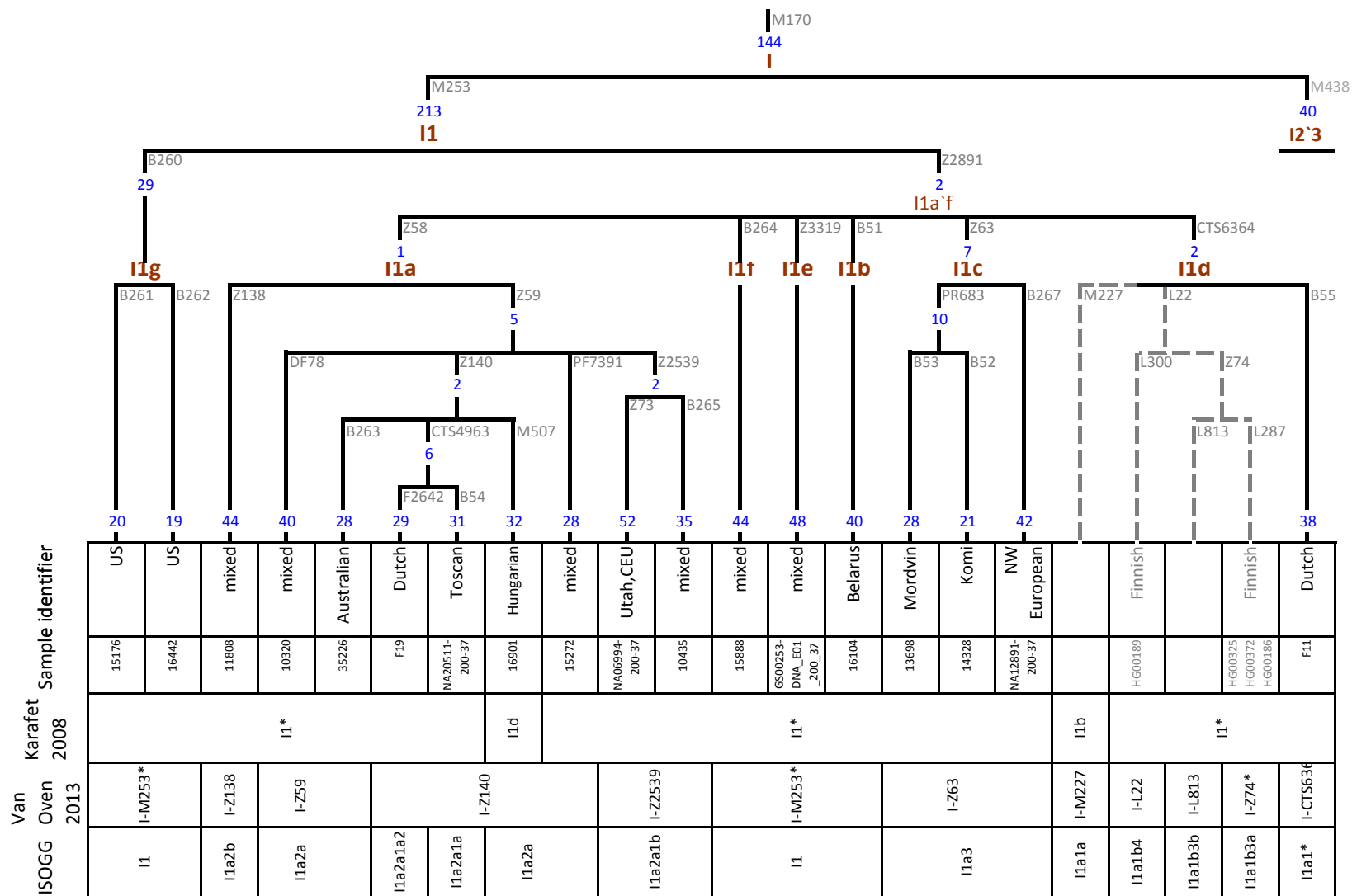


Figure S24. Refined topology of the Y-chromosome haplogroup I1. Grey dashed lines denote known branches with no sequence data and branches based on low coverage sequencing.

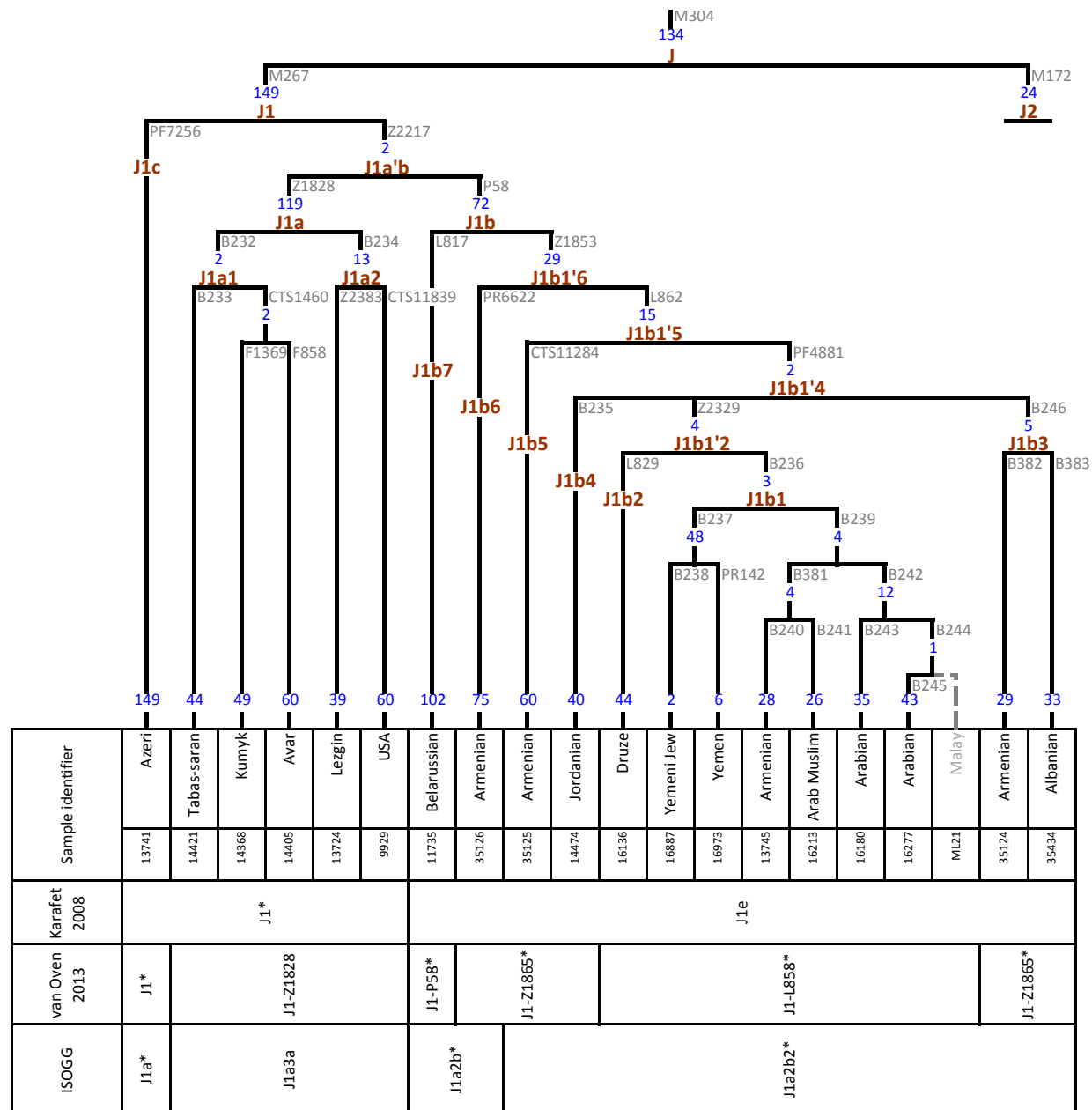


Figure S26. Refined topology of the Y-chromosome haplogroup J1.

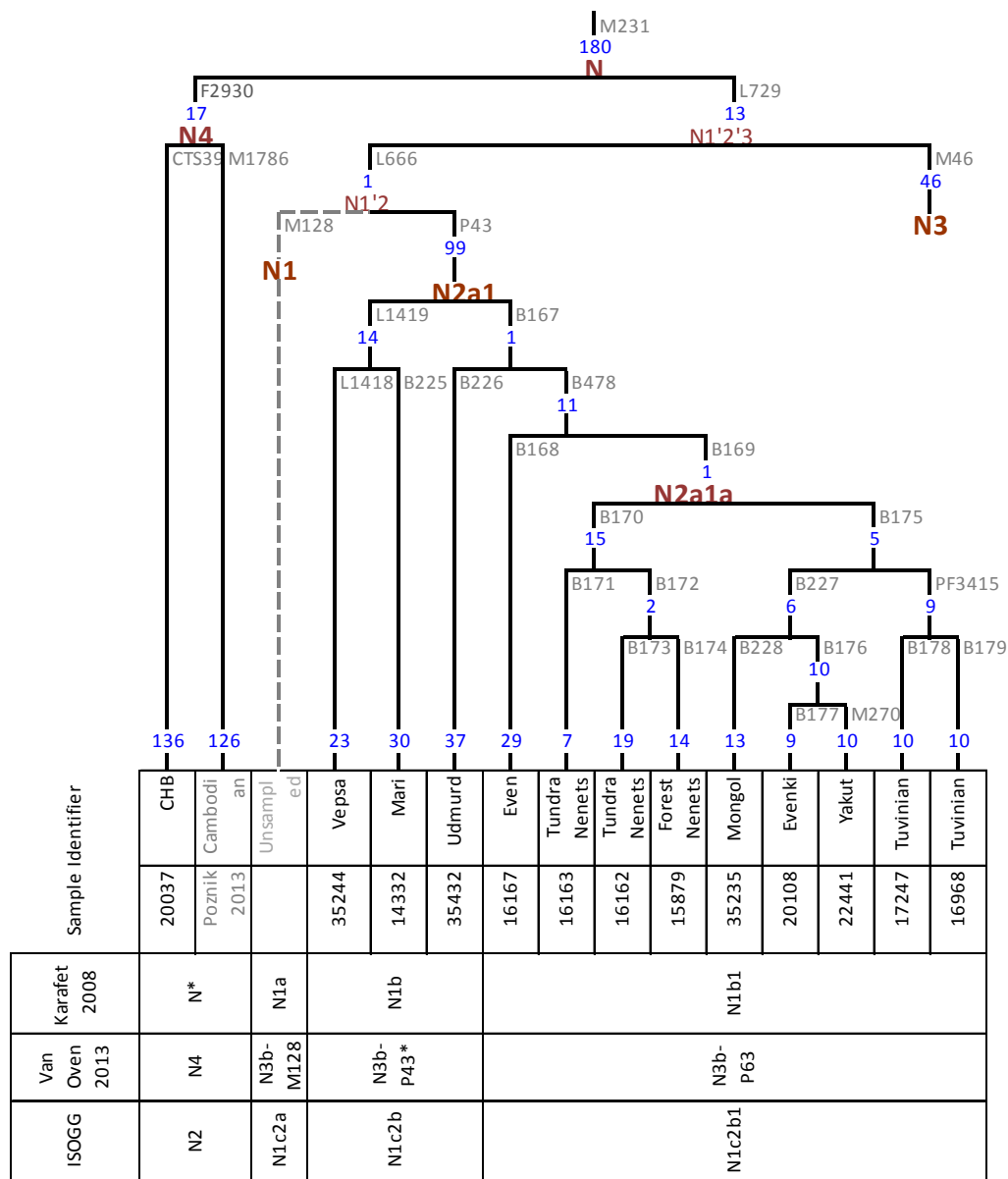


Figure S29. Refined topology of the Y-chromosome haplogroup N. Grey dashed lines denote known branches with no sequence data and branches based on low coverage sequencing. Haplogroup N 1'3 is further split into N 1'2-L166 (former N3b, PhyloTree) and N3-M46 (former N3a, PhyloTree).

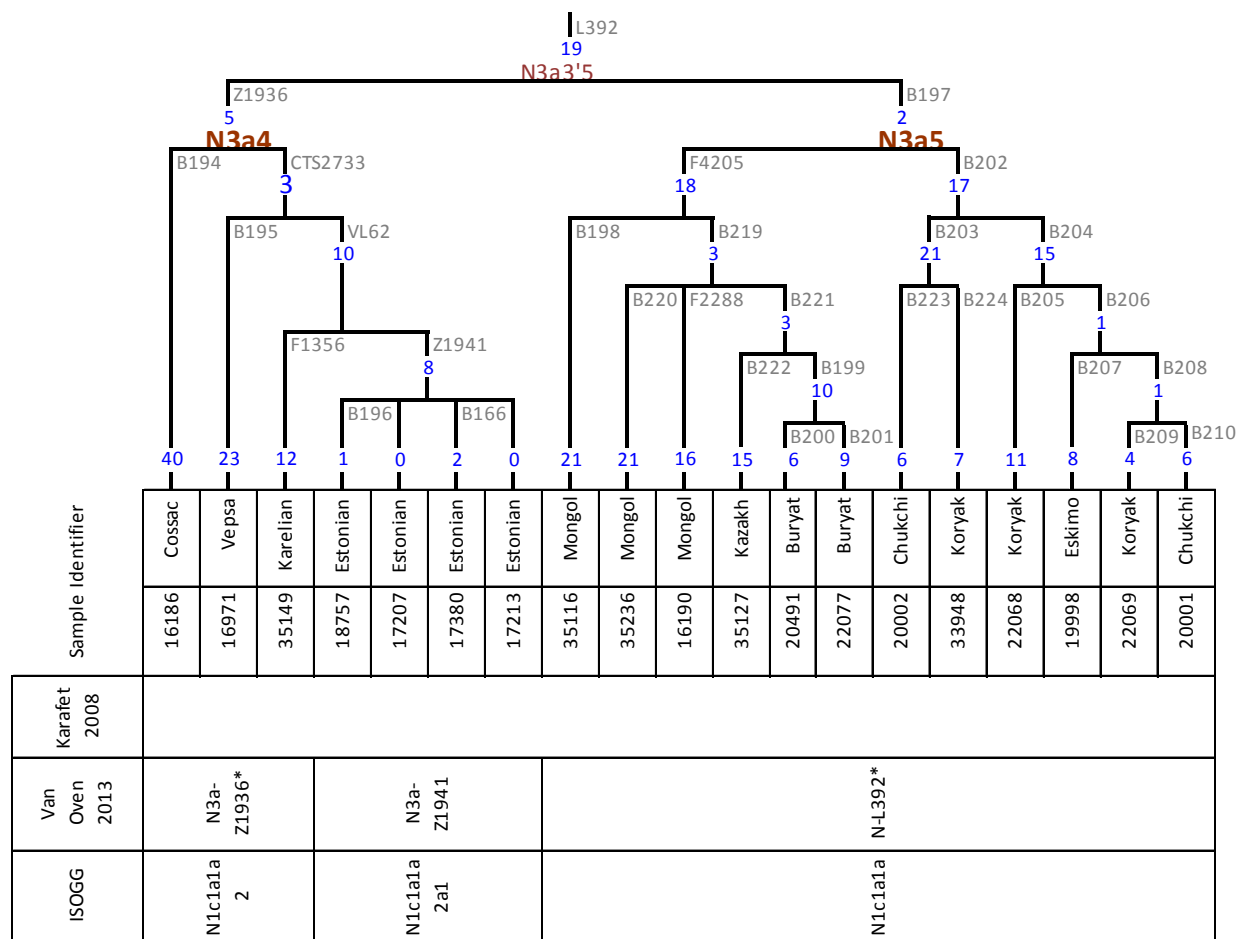
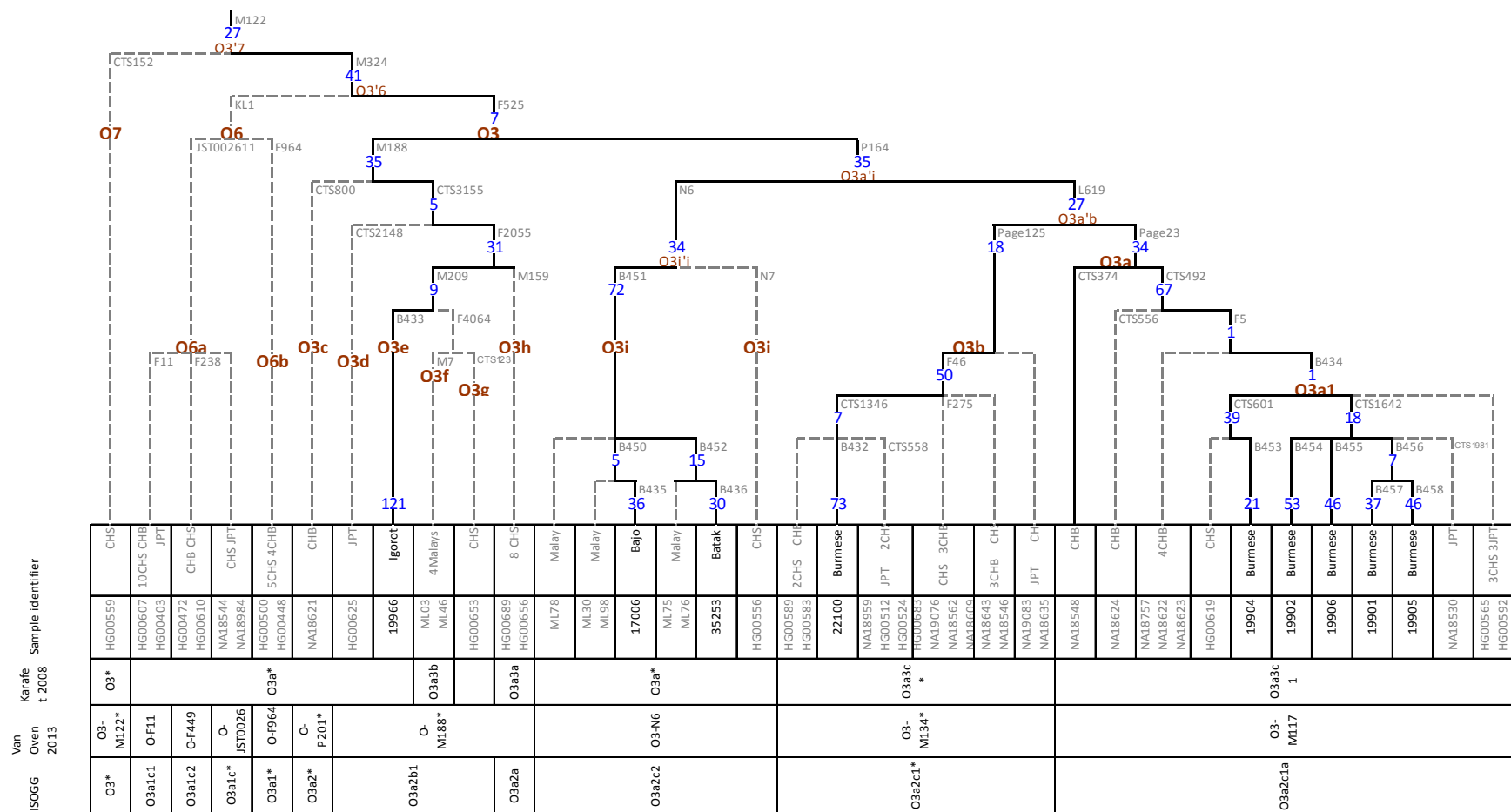


Figure S31. Refined topology of the Y-chromosome haplogroups N3a4 and N3a5.



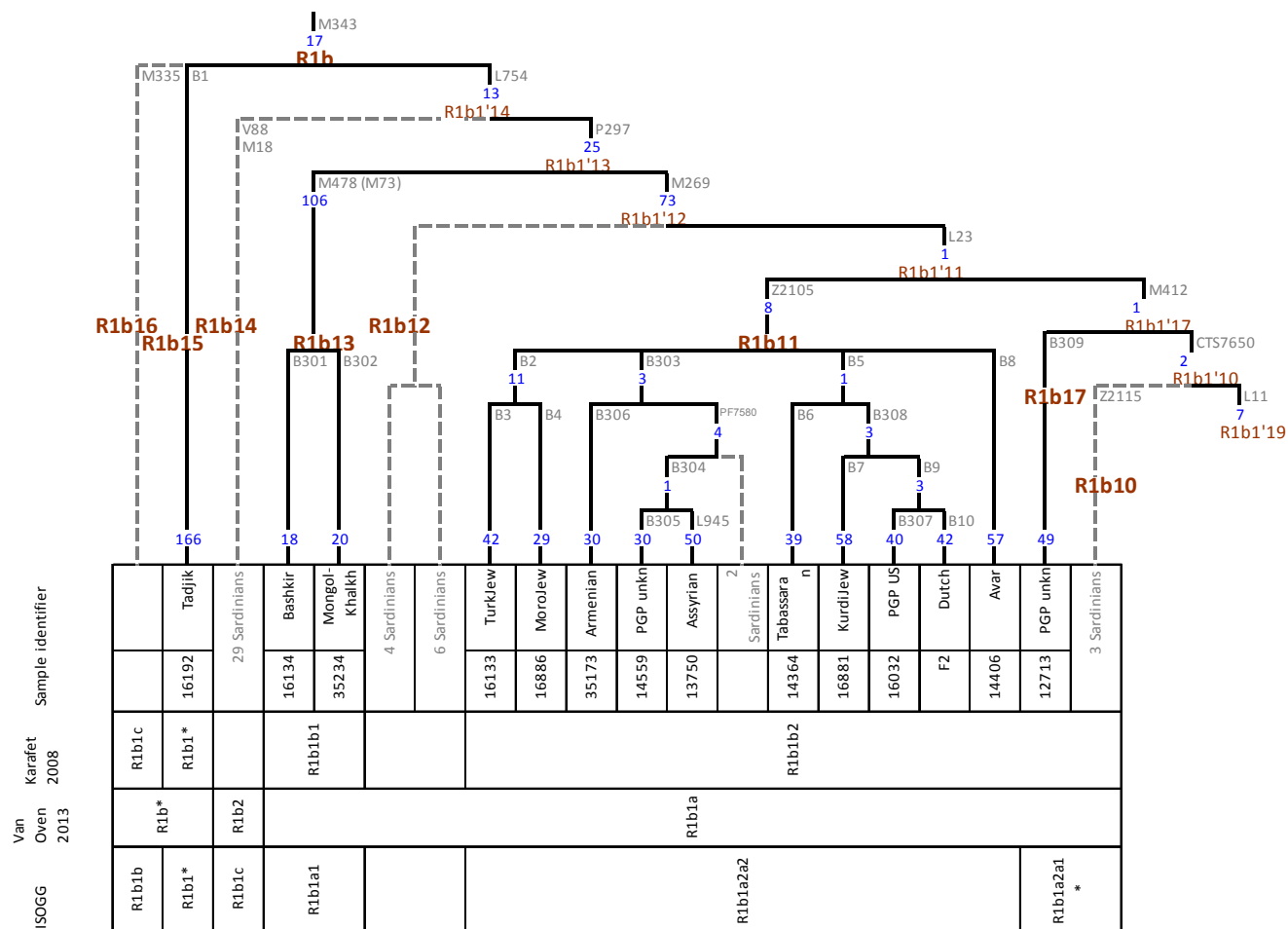


Figure S38. Refined topology of the Y-chromosome haplogroup R1b10-17. Grey dashed lines denote known branches with no sequence data and branches based on low coverage sequencing.

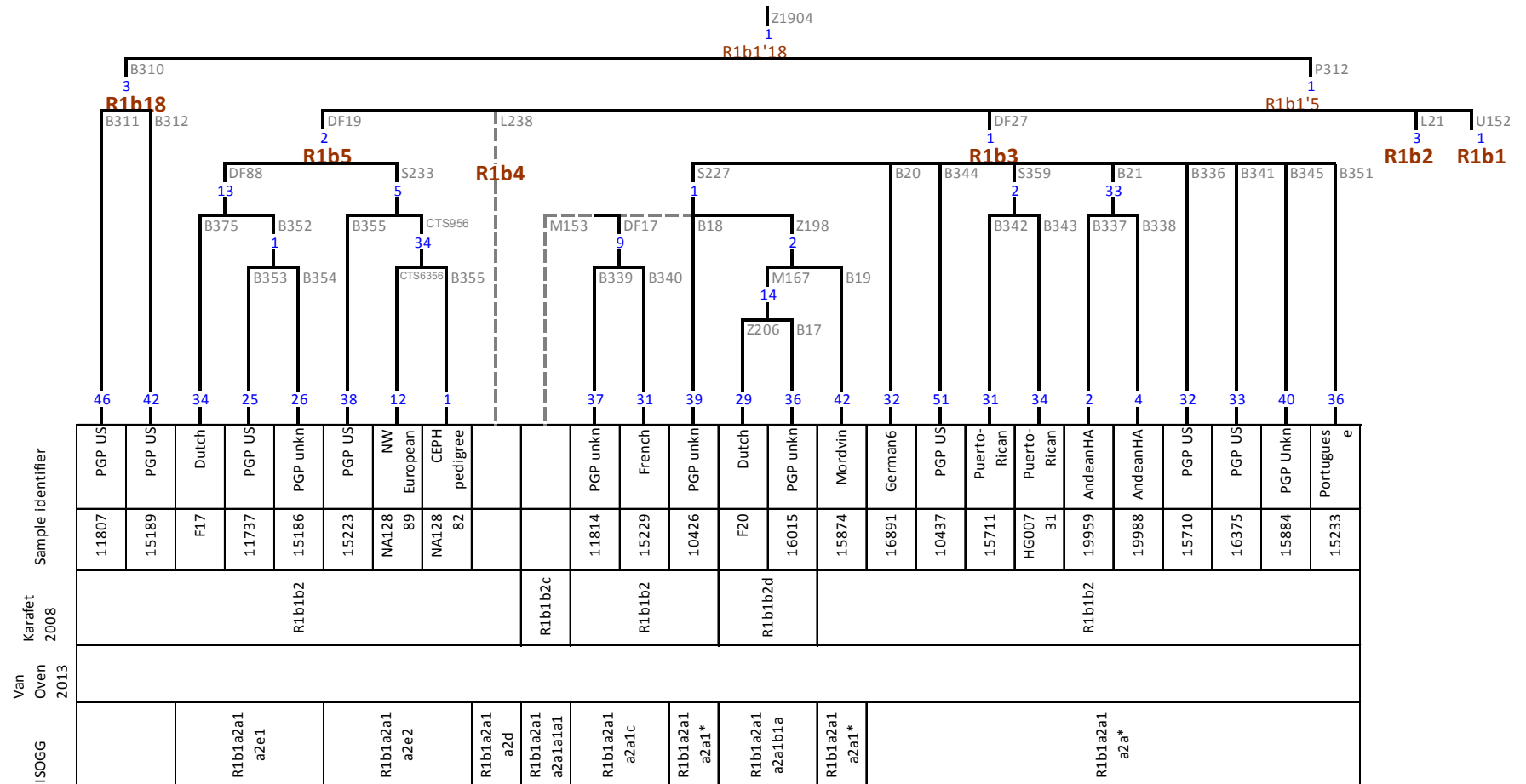


Figure S40. Refined topology of the Y-chromosome haplogroup R1b3-5. Grey dashed lines denote known branches with no sequence data and branches based on low coverage sequencing.

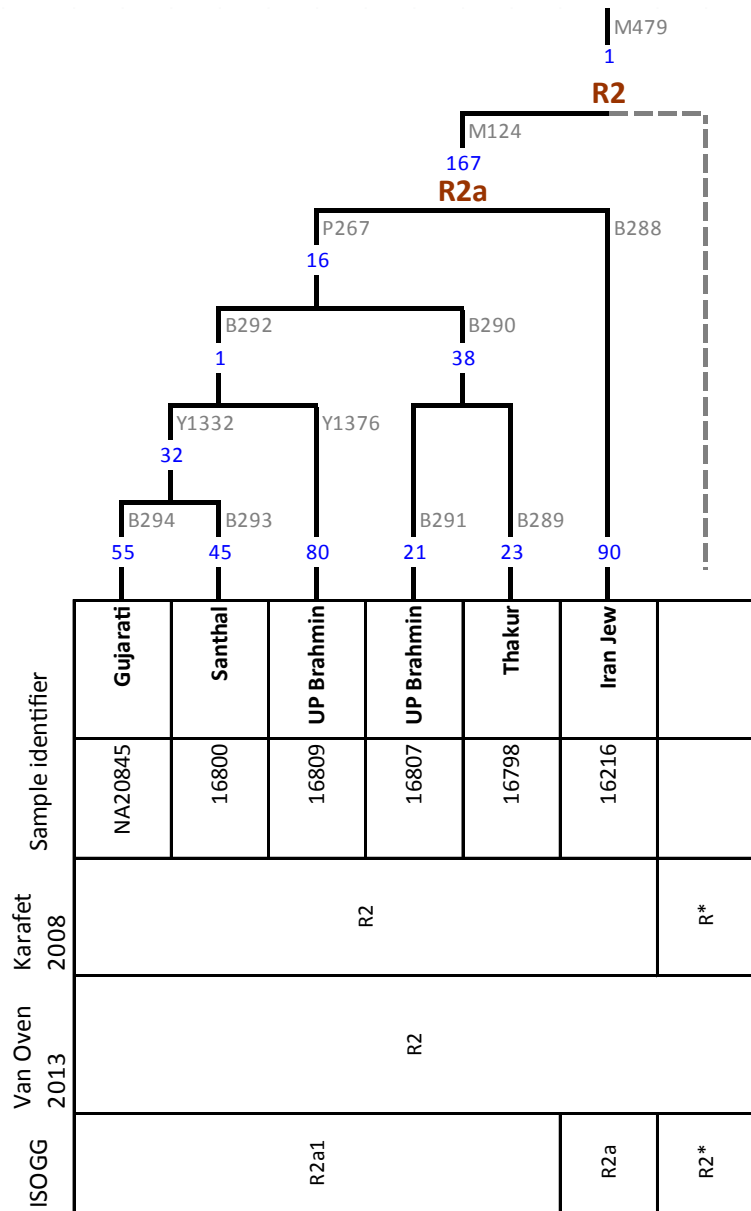


Figure S42. Refined topology of the Y-chromosome haplogroup R2.
 Grey dashed lines denote known branches with no sequence data.