Y-Chromosomal sequences of diverse Indian populations and the ancestry of the Andamanese

Mayukh Mondal¹, Anders Bergström², Yali Xue², Francesc Calafell¹, Hafid Laayouni¹, Ferran Casals³, Partha P. Majumder⁴, Chris Tyler-Smith^{2*} and Jaume Bertranpetit^{1*}

1. Institute of Evolutionary Biology (CSIC-UPF), Universitat Pompeu Fabra, Doctor Aiguader 88 (PRBB), 08003 Barcelona, Catalonia, Spain.

2. Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

3. Genomics Core Facility, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.

4. National Institute of Biomedical Genomics, Kalyani 741 251, India.

Abstract

We present 42 new Y-chromosomal sequences from diverse Indian tribal and non-tribal populations, including the Jarawa and Onge from the Andaman Islands, which are analysed within a calibrated Y-chromosomal phylogeny incorporating South Asian (in total 305 individuals) and worldwide (in total 1286 individuals) data from the 1000 Genomes Project. In contrast to the more ancient ancestry in the South than in the North that has been claimed, we detected very similar coalescence times within Northern and Southern non-tribal Indian populations. A closest neighbour analysis in the phylogeny showed that Indian populations have an affinity towards Southern European populations and that the time of divergence from these populations substantially predated the Indo-European migration into India, probably reflecting ancient shared ancestry rather than the Indo-European migration, which had little effect on Indian male lineages. Among the tribal populations, the Birhor (Austro-Asiatic-speaking) and Irula (Dravidian-speaking) are the nearest neighbours of South Asian non-tribal populations, with a common origin in the last few millennia. In contrast, the Riang (Tibeto-Burman-speaking) and Andamanese have their nearest neighbour lineages in East Asia. The Jarawa and Onge shared haplogroup D lineages with each other within the last ~7 thousand years, but had diverged from Japanese haplogroup D Y-chromosomes ~53 thousand years ago, most likely by a split from a shared ancestral population. This analysis suggests that Indian populations have complex ancestry which cannot be explained by a single expansion model.

Introduction

The Y-chromosome is a powerful tool for analysing the paternal ancestry of human populations (Jobling and Tyler-Smith 2003). As most of the Y-chromosome does not recombine, reconstructing haplotypes is straightforward. Thus it is possible to build a phylogenetic tree for all human Y-chromosomal variation that, analyzed in a geographic context, allows a phylogeographic approach. Analysis of the whole sequence of the Y-chromosome will encompass not only the specific variants that defined the classical “haplogroups”, but all the nucleotide variation in the accessible regions of the chromosome. This provides higher resolution and more reliable time estimates of lineage divergences (Wei et al. 2013; Hallast et al. 2014; Karmin et al. 2015; Poznik et al. 2016; Willems et al. 2016).

Indian subcontinental population ancestry is complex and many attempts have been made to unravel this intricate history using both uniparental markers, mtDNA and the Y-chromosome (Kivisild et al. 2003; Cordaux et al. 2004; Metspalu et al. 2004; Sahoo et al. 2006; Thanseem et al. 2006; Sengupta et al. 2006; ArunKumar et al. 2012), and autosomes (Reich et al. 2009; Metspalu et al. 2011; Moorjani et al. 2013; Juyal et al. 2014; Basu et al. 2015). Till recently, the Y-chromosome analysis mostly relied on calculating frequencies of pre-defined haplogroups, and the time and place of origin have been estimated for some of these. Some high frequency haplogroups in the Indian subcontinent are also common outside, making interpretation of their ancestry challenging. Also, hypotheses to explain the origin of Indian Caste and Tribal paternal lineages are conflicting. Whereas some argued for similar origin of both Indian Caste and Tribal populations (Kivisild et al. 2003), an independent origin has also been argued for their paternal lineages (Cordaux et al. 2004). Furthermore, Thanseem et al. (2006) emphasized the genetic affinities between lower caste and tribal groups. Even the origin of the common Y-chromosomal haplogroup R (hereafter all haplogroups mentioned in this paper are from the Y-chromosome only) in India is unclear. Whereas some (Cordaux et al. 2004) argued for a non-indigenous origin for R-M17 (R1a1a1) and R-M124 (R2a)

1 resulting from Indo-European speaking migrations, others proposed that either R1a and R2
2 (Sahoo et al. 2006), or R1 and R2 (Thanseem et al. 2006), or R2 (Sengupta et al. 2006) were
3 not introduced into India by the Indo-European migration. A migration is likely to have been
4 graduated, without mass movements in an elite-dominance model (Thapar, 2014).

5 In addition, it was found that the Andamanese population carried the D haplogroup
6 (Thangaraj et al. 2003), which is phylogenetically closer to the common African haplogroup E
7 than to other Out-of-Africa (OOA) haplogroups (Poznik et al. 2016), which could suggest a
8 different origin for the settlers of the Andaman Islands than those from the continent. Apart
9 from the Andaman Islands, haplogroup D is found at high frequency among populations in
10 Tibet (and some Chinese minorities) and the Japanese archipelago, though curiously not in
11 continental India. In fact it was already demonstrated that this Andamanese D haplogroup
12 shares a more recent common ancestry with Asian haplogroup D individuals, than to African
13 haplogroup E individuals (Thangaraj et al. 2003; Thangaraj et al. 2005).

14 In the present study, we have investigated the male side of Indian history by reconstructing
15 sequence-based Y-chromosome phylogenies and analyzing them from a wide set of
16 populations, incorporating new non-tribal and tribal (including Andamanese) populations
17 (Mondalet al. 2016), together with non-tribal populations from South Asia from the 1000
18 Genomes Project (The 1000 Genomes Project Consortium 2015). We focused on elucidating
19 the ancestry of South Asian non-tribal populations, as well as the divergence time of the
20 Andamanese and Japanese haplogroup D chromosomes.

21 **Methods**

22 **Samples**

23 In total, 42 new samples from ten different Indian populations (including five Andamanese)
24 were studied for the present analysis (Table 1 and Figure 1); for more information about the
25 populations, see Mondal et al. (2016). Data from South Asia from the 1000 Genomes Project
26 (Poznik et al. 2016) was integrated in the study, resulting in a total of 305 South Asian Y-
27 chromosome sequences.

28 **Sequencing**

29 The whole-genome sequencing of Indian populations was performed using the Illumina
30 technology (Mondal et al. 2016). We extracted Y-chromosome sequences from the BAM file
31 of whole genome sequences using samtools 1.1 (Li et al. 2009). The average coverage for the
32 uniquely mapped regions, suitable for short read sequencing (Poznik et al. 2013) is ~7x. A
33 total of 1244 individuals from phase 3 of the 1000 Genomes Project Y-chromosome BAM
34 files were downloaded from the project site (The 1000 Genomes Project Consortium 2015).

35 **Variant calling**

36 Variant calling on the BAM files of Indian samples and the 1000 Genomes Project was done
37 with the Genome Analysis Toolkit 3.5 (McKenna et al. 2010) using HaplotypeCaller and the
38 gvcf method, using the default parameters except for the ploidy of the genome, which was set
39 as 1. After generating gvcf files for all individuals, we called them together by
40 GenotypeGVCFs. We used dbSNP version 137 to get rsids for known SNPs (Sherry et al.
41 2001). Calling was restricted to the regions which are suitable for short-read calling on the Y-

chromosome (Poznik et al. 2013) using the -allSites flag in the GenotypeGVCFs. Other parameters were set to the default values.

Filtering

As the 1000 Genomes Project data have low coverage, we used several filters to focus on positions with more power for variant calling. We removed any position which had lower total coverage than 1502x of all individuals together (half of the average coverage for all sites) or more than 6006x (double the average coverage for all sites). If a polymorphic site had more than 200 individuals having at least one read different than the called genotype, it was also removed. Finally, positions with a ratio greater than of 0.1 for the number of reads with mapping quality 0 to the total number of reads (calculated using the samtools depth -Q flag) were also removed regardless of GATK's default parameter for variant calling. Any position having more than 30% of the sample with missing genotypes was also removed.

Phylogenetic inference and dating

After using the above filters and removing all the indels, we then removed all the monomorphic sites from the data set using vcftools (Danecek et al. 2011). We were left with 61,924 polymorphic sites for the whole data set. We next used PGDSpider-2.0.9.2 (Lischer and Excoffier 2012) to convert the vcf file to phylip format, which is required for RAxML input. We used RAxML 8.2.4 (Stamatakis 2014) for the phylogenetic analysis, with the ASC_GTRGAMMA model of nucleotide substitution (Poznik et al. 2016). We assigned HG02982 and HG01890 (haplogroup A0) from the 1000 Genomes Project as an out-group to all other lineages (Poznik et al. 2016). We used 100 bootstrap replicates to calculate the statistical support, and visualized the phylogeny with Tree of Life (Letunic and Bork 2011). The only major difference from Poznik et al. 2016 is that the HG02040 individual (from Kinh in Ho Chi Minh City, Vietnam), previously classified as haplogroup F*, is now within the haplogroup H cluster (Figure 3b). In the new data set, we obtained higher quality data for haplogroup H Y-chromosomes, which may account for this discrepancy; nonetheless, the bootstrap value for this branch is only 61 out of 100 and the branch is very basal within the H haplogroup in the tree, so the true branching pattern here remains uncertain.

In the RAxML output, distances were measured by the number of substitutions, which were transformed to the number of years to determine the time of divergence between lineages (as the number of substitutions is linearly related with time of divergence). Given the divergence time of A0 from the rest (Poznik et al. 2016), the genetic distance of A0 individuals from the root helped us to transform all the genetic distances to time, effectively calibrating by the same mutation rate (7.6×10^{-10} per base per year) (Fu et al. 2014). Caution is needed due to the lower coverage for the calculation of divergence times in the 1000 Genomes Project data, especially at the end of branches. We used "maps", "mapdata", "mapproj" from the R package to plot Indian subcontinent maps and "plotrix" and "Rcolorbrewer" to plot the pie charts.

Merging with the Karmin et al. (2015) data set

We also added one D1 haplogroup individual (NplTA1 from Nepal) from the Karmin et al. (2015) data set to analyze the D haplogroup phylogeny in more depth. As this data set was generated using the Complete Genomics platform, our variant calling approach deviated from that used for other individuals. We extracted all the positions from the file masterVarBeta_chrY-GS000035467-DID-ASM.tsv.bz2 for regions suitable for short read

calling on the Y-chromosome (Poznik et al. 2013). We only kept positions with power to detect reference or SNP("varType"="ref"/"snp"), and SNPs which passed the filter ("allele1VarFilter"="PASS"). Then we merged the data set with the previously filtered positions from the combined 1000 Genomes Project and Indian datasets. As the Complete Genomics data set has high coverage, we did not use any further filters. We next extracted two A0 individuals, all the haplogroup D Japanese from the 1000 Genomes Project, all the Andamanese, and the Nepalese from the Karmin et al. data set. We then reconstructed a phylogenetic tree as we did for 1000 Genomes Project and Indian data set (see above). We note that merging data from these two different sequencing technologies might create some unforeseen biases.

Clades and closest neighbour analysis

The phylogenetic tree was used to define and calculate the age of the clades of interest, and for the closest neighbour analysis. A clade was defined when all the sequences in a subtree came from individuals from the same (super) population (Figure 2). In our case, the minimum number is two to form a clade. Super-populations were defined as three different categories: North Indian [Indo-European Speakers] non-tribal, South Indian [Dravidian Speakers] non-tribal and Indian in general. For the Indian super-population, we included both North and South Indian non-tribal individuals with Dravidian (Irula) and Austro Asiatic (Birhor) individuals, but not the Andamanese and Tibeto-Burman individuals as they have a different ancestry from other Indian populations (Mondal et al. 2016). We searched for all the biggest clades in the phylogenetic tree for each super-population, regardless of the haplogroup classification. The algorithm stops the search when one individual does not belong to that super-population. Then we calculated the Time to the Most Recent Common Ancestor (TMRCA) of such clades (Figure 2), which were used to calculate the divergence time of internal clusters in Figure 4a.

The closest neighbours were those sequences which were closest to a specific clade of a super-population (just outside of the clade) containing at least one Y-chromosome from a different super-population. In this case we also identified the specific population to which the closest neighbour belongs, and the time depth of the joint cluster (TMRCA of the clade and the closest neighbour together). Depending on the tree structure, the closest neighbour of a single clade can consist of a single or multiple individuals (Figure 2). The divergence times of such neighbours were calculated from the average TMRCA of the joint cluster to every individual of that cluster (essentially the average height of the joint cluster). The analysis of closest neighbours provides information about the time and location of the most recent migrations between the target populations and other populations represented in the tree. In Figure 4a, the blue distribution shows the divergence time of all such neighbours from specific super-population clades. Figure 4b shows the time depth of the closest neighbour for each sequence, separated by population of origin (horizontal axis), and differentiating the three super-populations where the closest neighbour is found (North India, South India and India). In Figure 4c, we only concentrated on Europeans, who are the closest neighbours of the Indian super-population (essentially a subset of Figure 4b).

All the phylogenetic and clade analyses was done with the “ape” R package (Paradis et al. 2004). As clade-specific analysis can be biased because of sampling effects, we also looked for the closest European for every Indian individual regardless of their clade or haplogroup, with similar results (not shown).

Whole-genome sequence analysis

A total of ten Japanese (JPT) autosomal whole-genome BAM files (five haplogroup D and five haplogroup O) were downloaded from the 1000 Genomes Project site. Andamanese and Dai BAM files were accessed from a previous project (Mondal et al. 2016). Variant calling was done in a similar way to that of the Y-chromosome, except that we changed the ploidy option to 2 for HaplotypeCaller and did the variant calling for only the polymorphic SNPs in our data set. We used the VariantRecalibrator from GATK using dbsnp137, HapMap 3.3, the 1000 Genomes Project Omni 2.5 and the 1000 Genomes Project Phase 1 SNPs with high confidence downloaded from the Broad Institute ftp site ([ftp.broadinstitute.org](ftp://ftp.broadinstitute.org), 11/05/2013). After VariantRecalibrator, we only kept SNPs which passed the filter and those without missing information. We added ancestral information from the 1000 Genomes Project website (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/). We then used ADMIXTOOLS 1.1 (Patterson et al. 2012) to calculate Dstat for autosomal data.

Simulations

We built a simple model where JPT-O (Japanese with haplogroup O Y-chromosomes) separated from AND (Andamanese) and JPT-D (Japanese with haplogroup D) 77 kya, and AND and JPT-D separated from each other 53 kya, following the Y-chromosome analysis. We used a mutation rate (μ) of 1.25×10^{-8} per site per generation (Scally and Durbin 2012), recombination rate (r) of 1.3×10^{-8} per site per generation and a generation time of 29 years. As Dstat is neither affected by the effective population size nor by time of admixture (Patterson et al. 2012), we set the effective population sizes of all these populations to 10,000 and the time of admixture from JPT-O to JPT-D around 400 generations ago. We simulated 30,000 regions of 50 kb each using ms (Hudson 2002):

```
ms 40 30000 -t 50 -r 52 100000 -I 3 20 10 10 -es 0.0175 2 <VAR> -ej 0.0175 4 3 -ej 0.0457 2 1 -ej 0.0655 3 1
```

Where $\langle \text{VAR} \rangle = 0-0.99$ with step of .01.

Dstat values were calculated from the simulated data. The fitting of the data was done by “lm” from the R package.

Accession numbers

Sequences have been deposited in the European Nucleotide Archive. PRJEB11455: Andamanese whole-genome sequences (BAM and FASTQ files); PRJEB16019: continental Indian whole-genome sequences; PRJEB19598: Y-chromosome vcf files.

Files are also accessible through an ftp server. User: US3Pdczl (case sensitive)
Password: KMRmDQQv (case sensitive). Server: sit-web.upf.edu Transfer mode: active

A high quality copy of the Y-chromosome tree in Figure 3a is accessible in:
<http://biologiaevolutiva.org/jbertranpetit/scientific-publications/> and also in
<https://www.dropbox.com/sh/ldpkqtc6yxqnqsk/AAA2062ovKULGmsWbvbYDzXVa?dl=0>

Results and Discussion

Indian Non-Tribal Y-chromosome ancestry

The haplogroup frequency distribution in South Asia (the 42 new Indian sequences and the 263 from 1000 Genomes Project) (Figure 1) is similar to that documented earlier (Poznik et al. 2016) in Indian populations. It is interesting to note that the well-recognized genetic cline from North to South for non-tribal Indian populations using autosomal data (Reich et al. 2009; Juyal et al. 2014; Basu et al. 2015) is not evident in the Y-chromosome haplogroup distribution (Figure 1 and Table 2): there is no specific haplogroup frequency correlated with the North to South distribution, a cline that could have been produced by a single population migration and admixture between two main original populations (represented now by Indo-European and Dravidian speakers) as proposed by some researchers (Reich et al. 2009). It should be pointed out, however, that the total number of populations analysed here is small, some sample sizes are very small, and three of the populations (Gujarati Indian from Houston, Texas [GIH], Indian Telugu from the UK [ITU] and Sri Lankan Tamil from the UK [STU]) from the 1000 Genomes Project data set were collected from emigrant individuals. Nevertheless, previous studies that focused on classical Y haplogroup frequencies (Kivisild et al. 2003; Cordaux et al. 2004; Sengupta et al. 2006) have similarly found that such a cline is not present.

We next analyzed whole Y-chromosomal sequences according to their position in the calibrated phylogeny (Figure 3a and 3b; see Figure 3c for codes). The initial analysis then considered all clades (made of at least of two adjacent sequences) that belong to the same super-population (North India [mainly Indo-European speaking populations] or South India [mainly Dravidian speaking populations]), and including only the caste and not the tribal populations (see Methods). When calculating the oldest TMRCA of such clades for the two groups, we found that North Indian and South Indian-specific clades have a similar time of divergence, without an older distribution for the South (Figure 4a in red, with strong overlap), as could be expected if the South harbored an older population structure than the North. More specifically, under the simple model of Ancestral North Indian (ANI) and Ancestral South Indian (ASI) populations in different proportions and creating a north to south cline of ANI-ASI ancestry, the South Indian cluster would be expected to show a higher number of older TMRCAs than the North Indian cluster (as it is a more recent migration with a likely bottleneck and thus would have a lower TMRCAs), which we failed to detect in our data set. Interestingly, both of these populations have their oldest TMRCAs at ~18 kya, as reported earlier although using micro-satellite (Sengupta et al. 2006). Nonetheless, it has to be stressed that the time depth of a cluster may not be a good estimator of the population expansion time, as the Y-chromosomes in the migrant population will have pre-existing diversity that may remain in the new location (Barbujani et al. 1998). This is the reason for analysing not just the age of the clades, but also including the closest neighbour, as explained below and in Methods.

From the same tree (Figure 3a), it is possible to calculate the divergence time of the closest neighbour of each specific clade that has at least one individual not belonging to the same super-population (North or South India); this analysis will tell us where the closest sequence has been found outside the region, and the time depth of such neighbours. This analysis may thus provide interesting information about the neighbouring clade that was “left behind” in the migratory process, and so inform about the common time and maybe place of origin of the ancestors of both populations; nonetheless, caution has to be taken because of the

dependence on sampling size of this analysis. Results (Figure 4a in blue) show that the distributions for North and South India are similar, stressing that the neighbours of North and South Indian Y-chromosomes are mostly of similar age, although in South India there is a small proportion of deeper clusters with ages >30 kya, the only difference detected between the two super-populations.

We then concentrated on the origin and distribution of such neighbours in different worldwide populations (Figure 4b). When looking at these closest neighbours, we found that all Indian populations (both North and South) are the closest neighbours of each other till very recently, ~5 kya (Figure 4b; note the effect of the smaller sample size of the new Indian sequences as compared to the 1000 Genomes Project). No differences between North and South could be discerned within India, which had a similar reciprocal sharing of neighbours except that some South Indian neighbours are found with large time depth (in Figure 4b, four points above 30kya, one each in ILA and STU, and two in BEB) as the only tendency towards more ancient times, as seen also in Figure 4a.

Interestingly, none of the closest neighbours of South Indian non-tribal clades was found outside India (except some American populations from the 1000 Genome Project, that are known to be admixed), while North India (and India, of course) had a high frequency of neighbours among Europeans, suggesting that South Indian-specific clades have deeper roots in the Indian subcontinent. Nearly no neighbours were detected among East Asians or Africans. Surprisingly, the two South European populations (Toscani in Italia, TSI, and Iberian Population in Spain, IBS) are the closest neighbours of North Indian populations outside India (Figure 4b); unfortunately in this data set there are no data available for West Asia to indicate a more plausible place where the two groups (India and South Europe) could have some partial common origin; future work in the regions will allow a more precise analysis. The distribution of time depths for the closest neighbours of Indians demonstrated two different clusters for these two South European populations (Figure 4c). One is common to all Europeans and close to 38.6 kya (± 7.4 kya), while the second is more specific to South Europeans (TSI and IBS) and around 13.9 kya (± 4.6 kya). However, we need to stress that the absence of a relevant sample (likely from Western Asia) in the closest neighbour analysis can lead a higher time of divergence than the true divergence.

When looking at the haplotype composition of the Y-chromosomes that have their closest neighbours in Europe in recent times (less than 25 kya), a high proportion (~72%), belong to haplogroup J2, which is well distributed in India and of clear ancestry in West Asia, likely related to the demic diffusion during the Neolithic (Singh et al. 2016), and a much lower proportion to haplogroup R1a (~10%) which is reported to be associated with the Indo-European language migration (Semino 2000). Interestingly, the divergence times of these closest European neighbours are ~17.3 (± 3.8) kya for J2 and 7.3(± 2.2) kya for R1a.

Indian Tribal Y-chromosome ancestry

In our previous study of autosomal data (Mondal et al. 2016), we showed that Birhor (BIR, Austro-Asiatic-speaking tribal population) and Irula (ILA, Dravidian-speaking tribal population) are similar to each other, and have negligible amounts of European ancestry, suggesting they are a good proxy for the ASI, even if geographically the BIR live in the north; both are tribal. Surprisingly, in the present analysis (Figure 4b) Birhor Y-chromosome sequences were the closest neighbour of North Indian clades and Irula were the closest neighbours of South Indian clades only, suggesting that there might have been major exchange of Y-chromosomes (male driven) between Tribal and Non-tribal populations with

their geographic neighbours that depended on their geographic positions in recent times (around 10 kya). Nonetheless, it should be stressed that we have a low number of tribal Y-chromosomes in our data set (4 Birhor and 3 Irula), and more sequences of tribal individuals are needed to complete the picture of the relative importance of North-South in relation to Caste-Tribal populations.

The Y-chromosomes of the Tibeto-Burman population (Riang, RIA) show a different pattern, as expected from previous studies (Mondal et al. 2016). All their Y-chromosomes belong to haplogroup O, as also previously reported (Ning et al. 2016). Haplogroup O is the main haplogroup in East Asia and is also found in 10% of Bengalis from Bangladesh (BEB), being thus a haplogroup characteristic of, and exclusive to, the East of the region considered here. When calculating the TMRCA of each RIA chromosome to its neighbour in the tree (outside the RIA population, which in most cases are Far East Y-chromosomes), we find a time of around ~8.7 kya (± 1 kya). The geographic distribution and frequency of Eastern lineages in India suggest a relatively recent migration from East Asia that entered India via the North-Eastern border.

Andamanese ancestry

The Andamanese stand out in our data set because of their haplogroup D, which in India has been reported only in the Andaman Islands, and is present in all of the five individuals we sequenced. This haplogroup is especially interesting as it has a deep genealogy, sharing a ~4 kya more recent ancestry with the African haplogroup E than all other haplogroups found in OOA populations (Poznik et al. 2016). This relationship with African populations has been a basis for postulating a distinct early OOA migration, differentiated from the later and more widespread Asian expansion (Shi et al. 2008). This proposal is in stark contrast to findings from autosomal data, where the Andamanese showed a common ancestry with other Asian populations without a trace of contribution from a putative earlier OOA (Mondal et al. 2016). In addition, this haplogroup has been found at high frequencies in Japan (Poznik et al. 2016) and in Tibet (Lu et al. 2016).

We found that this haplogroup in the Andamanese has a deep substructure (mean of divergence time of closest neighbours/mean of oldest TMRCA for population specific clades=6) which is a much higher ratio than for other Indian populations (Table 2). This suggests that Andamanese Y-chromosomes have had a long-lasting separation from the other populations examined, and that male gene flow into the Andaman Islands has been limited (Table 2).

As this haplogroup is also found at high frequencies in Japanese individuals from Tokyo (JPT) from the 1000 Genomes Project, it is possible to analyse the divergence times between the two populations for both the haplogroup D Y chromosomes and the autosomes. The divergence time between Andamanese and JPT haplogroup D Y chromosomes is ~53 kya (± 2.7 kya), while the divergence time between haplogroup D and O chromosomes is ~77.3 kya (± 3.3 kya). In contrast, the autosomal divergence time is ~52 kya between Andamanese and East-Asians (Mondal et al. 2016, Supplementary information, using the mutation rate described in Scally and Durbin 2012). This possible discrepancy between Y-chromosome (divergence between D and O haplogroup) and autosomal (divergence between Andamanese and East Asians) data could potentially be explained by two different hypotheses: a) It could be caused by two different settlements: people with haplogroup D first populated Asia, including the most external localities, like the Andaman Islands and Japan; later in Japan this haplogroup was partially replaced by haplogroup O individuals (JPT-O). According to this

model, haplogroup D is more frequent in the two extremes of the archipelago as a consequence of this later replacement b) If a single main OOA event happened, haplogroup D along with other haplogroups in other OOA populations (i.e. C and macrohaplogroup F) were already present in the initial OOA population, but later haplogroup D was lost from other populations by random drift except in the Andaman Islands, Japan (Poznik et al. 2016), and Tibet. If the first hypothesis were true, even if the genomes of JPT-O and JPT-D had homogenized, we would expect the Andamanese and JPT-D to share more derived autosomal alleles with each other than with other East Asian populations (for example Dai from China). We found that this is not the case (Table 3). In fact, as seen from the autosomes (Mondal et al. 2016), the Andamanese are an out group to East Asian populations (at least for the Dai and Japanese as shown in this paper), diverging in a trichotomy with East Asians and Indians.

In order to investigate whether the Andamanese D lineages form a clade with the Nepalese/Chinese D1 or the Japanese D2 lineages, as the 1000 Genomes Project data do not have any haplogroup D1 individual, we added one Nepalese individual from Karmin et al. (2015; see Methods for more details). Although the Andamanese are within the D2 haplogroup phylogeny, the time divergence of D1, D2 and Andamanese haplogroup are shown to be quite similar, making a trichotomy (not shown).

We also built a simple simulation model starting from Y-chromosome data, where the Andamanese and JPT-D first separated from other OOA populations around 77 kya, and then separated each other around 53 kya. Later, JPT-O individuals started to admix with JPT-D individuals without affecting Andamanese. We need 99% ($\pm 3\%$) admixture from JPT-O to JPT-D to match the empirical Dstat, which suggests a complete replacement of the autosomal genetic material of JPT-D individuals without affecting the Y-chromosome at all. This proves that the first scenario is highly unlikely, leaving us with the second hypothesis. The high frequency of haplogroup D in Tibet (Gayden et al. 2007 and references therein) may be part of a different expansion, although new studies with sequence data are needed to investigate this further.

Within the five Andamanese Y-chromosomes, there is no structure between the two ethnic groups (Jarawa and Onge), and they are intermingled in the Y phylogeny. The mean time depth within the Andamanese D haplogroup is 7 kya (± 2.7 kya).

Conclusions

The comparative analysis of Y-chromosome sequences of many Indian populations - with different locations, languages and social structure - has demonstrated that Indian populations have a complex ancestry which cannot be explained by a single expansion from the North (or North-West) to the Indian subcontinent on a uniform pre-existing population. It is interesting to note that Y-chromosome data displayed less diversification between the North and South than has been described in autosomal studies (Basu et al. 2003; Reich et al. 2009; Moorjani et al. 2013; Juyal et al. 2014), with no clear clines of frequency of haplogroups in the present dataset.

The analysis of full Y-chromosomal sequences allows a deeper analysis, with assessment of the time depth of population-specific clades and, of particular interest, of the time depth and place of origin of clusters of Indian sequences compared with the closest sequence found outside India in a worldwide survey. The lack of a North to South cline of frequency of any haplogroup and similar time of divergence for both North and South India non-tribal populations suggest a similar paternal ancestry for both of these populations.

Our time divergence estimate matches previous studies which argued that most of the haplogroups present in India (C5, F*, H, L1 and R2) arose inside India rather than being brought from outside (Sengupta et al. 2006; Carvalho-Silva et al. 2006). The puzzling point is that the well-recognized later Indo-European migration, which strongly affected the northern regions, did not produce detectable major changes in the Y-chromosome gene-pool. This could be explained either by a similar Y-chromosome background in the two main migrations (for example, Neolithic and Indo-European) or that the second one did not have a major genetic impact, at least for the Y-chromosome (Sengupta et al. 2006). In general, these results support the view of Thapar (2014) that Dravidian speakers were spread geographically throughout India and that the incursion of Indo-Aryan speakers in the Northwest dominated pre-existing Dravidian tribal populations, with an élite-dominance model and small genetic impact.

Even more interesting, the closest neighbours of Indian clades in our dataset are generally from Southern Europe (and not other European populations), a place known to have had more influence from the first Neolithic expansion from the Levant through Anatolia and less from the steppe migration which was perhaps responsible for the Indo-European expansion of languages in Europe (Haak et al. 2015); the future availability of ancient Y-chromosome sequences and reanalysis after merging available data from Western Asia (Hallast et al. 2014; Batini et al. 2015; Karmin et al. 2015) will help to better interpret this finding.

The time divergence between Indian and European Y-chromosomes, based on the closest neighbour analysis, shows two different distinctive divergence times for J2 and R1a, suggesting that the European ancestry in India is much older (>10 kya) than what would be expected from a recent migration of Indo-European populations into India (~4-5 kya). Also the proportions suggest the effect might be less strong than generally assumed for the Indo-European migration. Interestingly, the ANI ancestry was recently suggested to be a mix of ancestries from early farmers of western Iran and people of the Bronze Age Eurasian steppe (Lazaridis et al. 2016). Our results agree with this suggestion. In addition, we also show that the divergence time of this ancestry is different, suggesting a different time to enter India.

These results suggest that the European-related male ancestry in Indian populations might be much older and more complex than anticipated, and might originate from the first wave of agriculturists or even earlier, giving stronger support to the very old ages for Indian Y-chromosomes; it also downplays the importance of migration related to the Indo-Aryan linguistic expansion. It is interesting to note that both the closest neighbours of North and South Indian clusters have a lowest minimum around ~4-5 kya coinciding with the Indo-European migration, suggesting that the real stratification of these two populations started around that time, and may be related to a retraction to the South of Dravidian speakers (Thapar, 2014). Due to the low number of sequences from Dravidian and Austro-Asiatic tribes, it is difficult to be certain about the ancestry of these tribal populations. Nonetheless, both of these populations have shared recent ancestry (~10 kya) with their non-tribal neighbouring populations (BIR with North Indian and ILA with South Indian) which contrasts with that seen from autosomal data (Mondal et al. 2016).

The Andamanese and some Japanese (and also Tibetans, not present in this data set) showed the presence of haplogroup D Y-chromosomes, which provided one of the main reasons for previous genetic studies to postulate that the Andamanese represented an out-group to all other OOA populations, descendants of an early modern human expansion, following a coastal route migration from which only a few populations survive to the present day; we have been able to dismiss this hypothesis. We have shown that Andamanese and Japanese

1 individuals carrying haplogroup D separated around ~53 kya, coinciding with most of the
2 other major OOA haplogroup divergences. We have also demonstrated, using autosomal data,
3 that the Andamanese are indeed an out group to East Asian populations, which was also
4 shown before (Chaubey and Endicott 2013; Aghakhanian et al. 2015). This strongly suggests
5 that haplogroup D does not indicate a separate ancestry for Andamanese populations. Rather,
6 haplogroup D was part of the standing variation carried by the OOA expansion, and later lost
7 from most of the populations except in Andaman and partially in Japan and Tibet.

References

- Aghakhanian F, Yunus Y, Naidu R, et al (2015) Unravelling the genetic history of Negritos and Indigenous populations of Southeast Asia. *Genome Biol Evol* 7:1206–1215. doi: 10.1093/gbe/evv065
- ArunKumar G, Soria-Hernanz DF, Kavitha VJ, et al (2012) Population Differentiation of Southern Indian Male Lineages Correlates with Agricultural Expansions Predating the Caste System. *PLoS One*. doi: 10.1371/journal.pone.0050269
- Barbujani G, Bertorelle G, Chikhi L (1998) Evidence for Paleolithic and Neolithic gene flow in Europe. *Am J Hum Genet* 62:488–92. doi: 10.1086/301719
- Basu A, Mukherjee N, Roy S, et al (2003) Ethnic India: A genomic view, with special reference to peopling and structure. *Genome Res* 13:2277–2290. doi: 10.1101/gr.1413403
- Basu A, Sarkar-Roy N, Majumder PP (2015) Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Pnas* 113:201513197. doi: 10.1073/pnas.1513197113
- Batini C, Hallast P, Zadik D, et al (2015) Large-scale recent expansion of European patrilineages shown by population resequencing. *Nat Commun* 7:152. doi: 10.1038/ncomms8152
- Carvalho-Silva DR, Zerjal T, Tyler-Smith C (2006) Ancient Indian roots? *J Biosci* 31:1–2. doi: 10.1007/BF02705228
- Chaubey G, Endicott P (2013) The Andaman Islanders in a Regional Genetic Context: Reexamining the Evidence for an Early Peopling of the Archipelago from South Asia. *Hum. Biol.* 85:
- Cordaux R, Aunger R, Bentley G, et al (2004) Independent Origins of Indian Caste and Tribal Paternal Lineages. *Curr Biol* 14:231–235. doi: 10.1016/S0960-9822(04)00040-5
- Danecek P, Auton A, Abecasis G, et al (2011) The variant call format and VCFtools. *Bioinformatics* 27:2156–2158. doi: 10.1093/bioinformatics/btr330
- Fu Q, Li H, Moorjani P, et al (2014) Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514:445–9. doi: 10.1038/nature13810
- Gayden T, Cadenas AM, Regueiro M, et al (2007) The Himalayas as a directional barrier to gene flow. *Am J Hum Genet* 80:884–94. doi: 10.1086/516757
- Haak W, Lazaridis I, Patterson N, et al (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522:207–211. doi: 10.1038/nature14317
- Hallast P, Batini C, Zadik D, et al (2014) The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. *Mol Biol Evol* 32:661–673. doi: 10.1093/molbev/msu327
- Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338. doi: 10.1093/bioinformatics/18.2.337
- Jobling MA, Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet* 4:598–612. doi: 10.1038/nrg1124
- Juyal G, Mondal M, Luisi P, et al (2014) Population and genomic lessons from genetic analysis of two Indian populations. *Hum Genet* 133:1273–1287. doi: 10.1007/s00439-014-1462-0
- Karmin M, Saag L, Vicente M, et al (2015) A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res* 25:459–466. doi: 10.1101/gr.186684.114
- Kivisild T, Rootsi S, Metspalu M, et al (2003) The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am J Hum Genet* 72:313–332. doi:

10.1086/346068

Lazaridis I, Nadel D, Rollefson G, et al (2016) Genomic insights into the origin of farming in the ancient Near East. *Nature* 536:1–22. doi: 10.1038/nature19310

Letunic I, Bork P (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* 39:W475–8. doi: 10.1093/nar/gkr201

Li H, Handsaker B, Wysoker A, et al (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. doi: 10.1093/bioinformatics/btp352

Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 28:298–9. doi: 10.1093/bioinformatics/btr642

Lu D, Lou H, Yuan K, et al (2016) Ancestral Origins and Genetic History of Tibetan Highlanders. 1–15. doi: 10.1016/j.ajhg.2016.07.002

McKenna A, Hanna M, Banks E, et al (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303. doi: 10.1101/gr.107524.110

Metspalu M, Kivisild T, Metspalu E, et al (2004) Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet* 5:26. doi: 10.1186/1471-2156-5-26

Metspalu M, Romero IG, Yunusbayev B, et al (2011) Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am J Hum Genet* 89:731–744. doi: 10.1016/j.ajhg.2011.11.010

Mondal M, Casals F, Xu T, et al (2016) Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nat Genet* 48:1066–70. doi: 10.1038/ng.3621

Moorjani P, Thangaraj K, Patterson N, et al (2013) Genetic evidence for recent population mixture in India. *Am J Hum Genet* 93:422–438. doi: 10.1016/j.ajhg.2013.07.006

Ning C, Yan S, Hu K, et al (2016) Refined phylogenetic structure of an abundant East Asian Y-chromosomal haplogroup O*-M134. *Eur J Hum Genet* 24:307–9. doi: 10.1038/ejhg.2015.183

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–290. doi: 10.1093/bioinformatics/btg412

Patterson N, Moorjani P, Luo Y, et al (2012) Ancient admixture in human history. *Genetics* 192:1065–1093. doi: 10.1534/genetics.112.145037

Poznik GD, Henn BM, Yee M-C, et al (2013) Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* 341:562–5. doi: 10.1126/science.1237619

Poznik GD, Xue Y, Mendez FL, et al (2016) Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat Genet* 12:809–809. doi: 10.1038/ng.3559

Reich D, Thangaraj K, Patterson N, et al (2009) Reconstructing Indian population history. *Nature* 461:489–494. doi: 10.1038/nature08365

Sahoo S, Singh A, Himabindu G, et al (2006) A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proc Natl Acad Sci U S A* 103:843–8. doi: 10.1073/pnas.0507714103

Scally A, Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet* 13:745–53. doi: 10.1038/nrg3295

Semino O (2000) The Genetic Legacy of Paleolithic Homo sapiens sapiens in Extant Europeans: A Y Chromosome Perspective. *Science* (80-) 290:1155–1159. doi: 10.1126/science.290.5494.1155

Sengupta S, Zhivotovsky LA, King R, et al (2006) Polarity and temporality of high-resolution

1 y-chromosome distributions in India identify both indigenous and exogenous expansions
2 and reveal minor genetic influence of Central Asian pastoralists. *Am J Hum Genet*
3 78:202–21. doi: 10.1086/499411

4 Sherry ST, Ward MH, Kholodov M, et al (2001) dbSNP: the NCBI database of genetic
5 variation. *Nucleic Acids Res* 29:308–311. doi: 10.1093/nar/29.1.308

6 Shi H, Zhong H, Peng Y, et al (2008) Y chromosome evidence of earliest modern human
7 settlement in East Asia and multiple origins of Tibetan and Japanese populations. *BMC*
8 *Biol* 6:45. doi: 10.1186/1741-7007-6-45

9 Singh S, Singh A, Rajkumar R, et al (2016) Dissecting the influence of Neolithic demic
10 diffusion on Indian Y-chromosome pool through J2-M172 haplogroup. *Sci Rep* 6:19157.
11 doi: 10.1038/srep19157

12 Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of
13 large phylogenies. *Bioinformatics* 30:1312–3. doi: 10.1093/bioinformatics/btu033

14 Thapar R. (2014) Can genetics help us understand Indian social history? *Cold Spring Harb*
15 *Perspect Biol.* 26; 6(11):a008599.

16 Thangaraj K, Singh L, Reddy AG, et al (2003) Genetic Affinities of the Andaman Islanders, a
17 Vanishing Human Population. *Curr Biol* 13:86–93. doi: 10.1016/S0960-9822(02)01336-
18 2

19 Thanseem I, Thangaraj K, Chaubey G, et al (2006) Genetic affinities among the lower castes
20 and tribal groups of India: inference from Y chromosome and mitochondrial DNA.
21 *BMC Genet* 7:42. doi: 10.1186/1471-2156-7-42

22 The 1000 Genomes Project Consortium (2015) A global reference for human genetic
23 variation. *Nature* 526:68–74. doi: 10.1038/nature15393

24 Wei W, Ayub Q, Chen Y, et al (2013) A calibrated human Y-chromosomal phylogeny based
25 on resequencing. *Genome Res* 23:388–395. doi: 10.1101/gr.143198.112

26 Willems T, Gymrek M, Poznik GD, et al (2016) Population-Scale Sequencing Data Enable
27 Precise Estimates of Y-STR Mutation Rates. *Am J Hum Genet* 98:919–933. doi:
28 10.1016/j.ajhg.2016.04.001

29

Acknowledgements

Funding was provided by the joint Spain-India bilateral grant PRI-PIBIN-2011-0942 and BFU2016-77961-P (AEI/FEDER, UE) both awarded by the Ministerio de Economía y Competitividad (Spain) and with the support of Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2014 SGR 866). Anders Bergström, YaliXue and Chris Tyler-Smith were supported by The Wellcome Trust (Grant 098051).

Figure Legends

Figure 1: Approximate positions of the Indian and 1000 Genomes Project populations analysed, with a haplogroup composition pie chart, both for the newly-studied and the 1000 Genomes Project samples. Size of the pies: small, one individual; intermediate, 4-8 individuals; large, 42-60 individuals from the 1000 Genomes Project.

Figure 2: Schematic representation of clades and closest neighbours. P1 and P2 are two different super-populations. The red square defines the clade for P1 and the blue square shows the closest neighbours of this clade. The time divergence of the oldest TMRCA from such clades is defined as the TMRCA for the internal cluster (t_i), and the time divergence of the closest neighbours for such a clade is defined as the TMRCA for the external cluster (t_0).

Figure 3: Maximum likelihood phylogeny for the 1000 Genomes Project and the new Indian data (a) all individuals, (b) collapsing all major branches except the new Indian Y-chromosome samples and (c) colour codes for haplogroups and populations.

Figure 4: Phylogenetic neighbours of Indian Y chromosomes. a) In red, distribution of the time to the most recent common ancestor of North Indian- and South Indian-specific clades (internal clusters). In blue, distribution of the divergence time of the closest neighbours of such clusters that do not belong to the specific super-population (external clusters). Left: North Indian-specific clades; and right: South Indian-specific clades. Y axis is kya. b) Closest neighbours for North Indian-specific clades (red plus sign), South Indian-specific clades (blue circle) and Indian-specific clades (black triangle) that are external to each super-population. The X axis shows the population of these external neighbours whereas the Y axis shows the time divergence. c) Histogram of divergence time of the closest external neighbours of Indian-specific clades that are found among Europeans (CEU, GBR, FIN, TSI, IBS). A bimodal distribution is clear for the South European populations. X axis in kya and Y axis is the frequency of such external neighbours.

1 **Figure 1.**

2

3

4

1 **Figure 2**
2

3
4

1 **Figure 3a**

2
3

1 **Figure 3b**

2

3

1 Figure 3c

2

3

Figure 4:

a)

b)

1 c)

2
3

1 **Table 1:** Sample sizes of Indian males sequenced for this work

Population	Geographical Region	Latitude	Longitude	Linguistic Affiliation	Social Category	No. of Males Sequenced
Punjabi (PUN)	North	31	75.8	Indo-European	Middle Caste	1
Brahmin (UBR)	North	26.8	80.8	Indo-European	Upper Caste	7
Rajput (RAJ)	North	26.2	73	Indo-European	Middle Caste	7
Bengali (BEN)	North	22.7	88.1	Indo-European	Lower Caste	1
Vellalar (VLR)	South	13	79.9	Dravidian	Middle Caste	8
Irula (ILA)	South	8.1	77.3	Dravidian	Tribe	3
Birhor (BIR)	Central	23.6	83.4	Austro-Asiatic	Tribe	4
Andamanese (AND) (Jarawa [JAR] and Onge [ONG])	Andaman Islands	11.6	92.6	Unclassified	Tribe	5 (3 and 2)
Riang (RIA)	North-east	24	92	Tibeto-Burman	Tribe	6

2

1 **Table 2** Haplogroup counts in Indian population samples. ti is the oldest most recent
2 common ancestor for a population-specific clade, and to is the time divergence of the closest
3 neighbour of such clades in kya.

Name	C	D	H	J	L	N	O	R	mean(t0)/mean(ti)
PUN (1)	0	0	0	0	0	0	0	1	NA
UBR (7)	0	0	1	2	0	0	0	4	5.85/2.9(2.0)
RAJ (7)	1	0	2	2	0	0	0	2	NA
BEN (1)	0	0	0	1	0	0	0	0	NA
VLR (8)	0	0	1	3	4	0	0	0	8.4/6.5 (1.3)
ILA (3)	0	0	2	0	1	0	0	0	NA
BIR (4)	0	0	4	0	0	0	0	0	6.6/4.5 (1.5)
AND (5)	0	5	0	0	0	0	0	0	53.0/8.9(6)
RIA (6)	0	0	0	0	0	0	6	0	11.1/11.2 (1.0)

4

1 **Table 3:** Dstat values of Andamanese, DAI, JPT-D, JPT-O, YRI and Ancestral

W	X	Y	Z	D score	Z score
DAI	JPT-D	AND	YRI	0.0028	0.718
DAI	JPT-O	AND	YRI	0.0059	1.521
JPT-O	JPT-D	AND	YRI	-0.0033	-1.475
DAI	JPT-D	AND	Ancestral	0.0043	1.214
DAI	JPT-O	AND	Ancestral	0.0059	1.647
JPT-O	JPT-D	AND	Ancestral	-0.0017	-0.856

2
3