OPEN ACCESS Freely available online

PLOS ONE

Y Chromosomes of 40% Chinese Descend from Three Neolithic Super-Grandfathers

Shi Yan[1,2]*, Chuan-Chao Wang[1], Hong-Xiang Zheng[1], Wei Wang[2], Zhen-Dong Qin[1], Lan-Hai Wei[1], Yi Wang[1], Xue-Dong Pan[1], Wen-Qing Fu[1,4], Yun-Gang He[2], Li-Jun Xiong[3], Wen-Fei Jin[2], Shi-Lin Li[1], Yu An[1], Hui Li[1], Li Jin[1,2]*

1 State Key Laboratory of Genetic Engineering, and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China, **2** Chinese Academy of Sciences Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, SIBS, CAS, Shanghai, China, **3** Epigenetics Laboratory, Institute of Biomedical Sciences, Fudan University, Shanghai, China, **4** Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America

Abstract

Demographic change of human populations is one of the central questions for delving into the past of human beings. To identify major population expansions related to male lineages, we sequenced 78 East Asian Y chromosomes at 3.9 Mbp of the non-recombining region, discovered >4,000 new SNPs, and identified many new clades. The relative divergence dates can be estimated much more precisely using a molecular clock. We found that all the Paleolithic divergences were binary; however, three strong star-like Neolithic expansions at ~6 kya (thousand years ago) (assuming a constant substitution rate of 1×10^{-9}/bp/year) indicates that ~40% of modern Chinese are patrilineal descendants of only three super-grandfathers at that time. This observation suggests that the main patrilineal expansion in China occurred in the Neolithic Era and might be related to the development of agriculture.

Citation: Yan S, Wang C-C, Zheng H-X, Wang W, Qin Z-D, et al. (2014) Y Chromosomes of 40% Chinese Descend from Three Neolithic Super-Grandfathers. PLoS ONE 9(8): e105691. doi:10.1371/journal.pone.0105691

Editor: Bing Su, Kunming Institute of Zoology, Chinese Academy of Sciences, China

Received December 17, 2013; **Accepted** July 24, 2014; **Published** August 29, 2014

Funding: This research was supported by the grants from the National Science Foundation of China (31271338 and 31071096), and from Ministry of Education (311016). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: yanshi@fudan.edu.cn (SY); lijin.fudan@gmail.com (LJ)

Introduction

Demographic change is one of the central questions in understanding human history, and strong population expansions may be linked to various events as climate changes, alteration of social structure, or technological innovations. The recent advent of next-generation sequencing technology enabled systematic analysis of the population history using the information from the whole genome with less ascertainment bias, so we can re-assess how the various factors have influenced the human population size and structure [1,2]. Recent analyses of mitochondrial genomes revealed that the expansions of female lineages of East Asians [3] and those of Europeans [4] started before the Neolithic Era, contradictory to the hypothesis that the agricultural innovation constitutes the primary driving force of population expansions [5]. These observations prompted this study to investigate expansions of male lineages.

The Y chromosome contains the longest non-recombining region (~60 Mbp, in which ~10 Mbp is unique sequence in the genome and easy to analyze) in the human genome [6,7], making it an informative tool for reconstructing genetic relationship of human populations and paternal lineages, and dating important evolutionary and demographic events [8,9,10,11]. However, the sequencing data of Y chromosomes of human populations were insufficient and biased even for those of current 1000-genome project for which coverage on Y chromosome was low (on average <1.4× in East Asian samples) [12].

According to the phylogenetic tree of Y chromosome, all the modern males could be categorized into 20 major monophyletic or paraphyletic groups (referred to as A to T) and their subclades [13,14]. Nearly all the Y chromosomes outside Africa are derivative at the SNP M168 and belong to any of its three descendent super-haplogroups – DE, C, and F [9,10,15], strongly supporting the out-of-Africa theory. The time of the anatomically modern human's exodus from Africa has yielded inconsistent results ranging from 39 kya [16], 44 kya [10], 59 kya [17], 68.5 kya [18] to 57.0–74.6 kya [19].

To achieve sufficiently high coverage in the non-recombining regions of Y chromosome (NRY) and an adequate representation of individual samples, we selected 110 males, encompassing the haplogroups O, C, D, N, and Q which are common in East Eurasians, as well as haplogroups J, G, and R which are common in West Eurasians (see Table S1), and sequenced their non-repetitive segments of NRY using a pooling-and-capturing strategy.

Results

Overall ~4,500 base substitutions were identified in all the samples from the whole Y chromosome, in which >4,300 SNPs that has not been publicly named before 2012 (ISOGG etc.). We designated each of these SNP a name beginning with 'F' (for Fudan University) (see Table S2). We obtained ~3.90 Mbp of sequences with appropriate quality (at least 1× coverage on >100

out of 110 samples, see Table S3), and identified ~3,600 SNPs in this region. A maximum parsimony phylogenetic tree of the 78 individuals with good coverage was reconstructed (Fig. 1 and Fig. S1), the topology of which is congruent with the existing tree of human Y chromosome [13,20]. The tree contained samples from haplogroups C, D, G, J, N, O, Q, and R, and thus represented all the three super-haplogroups out of Africa – C, DE and F. In addition to the known lineages, many new downstream lineages were revealed. All the earlier divergences were found to be bifurcations, except for three star-like structures, i.e. multiple lineages branching off from a single node, were observed under Haplogroup O3a-M324, indicating strong expansion events.

By using Bayesian method [21] with a constant mutation rate of 1×10^{-9} substitution/base/year [7,19] (or one substitution per 256 years on 3.9 Mbp range, without considering the uncertainty of mutation rate), we calculated the date of each divergence event throughout the tree. The first divergence event out of Africa, i.e. between Haplogroup DE and the ancestor of C and F, is dated at 54.1 kya (95%CI 50.6–58.2), inside the range of previous estimations. Within the 3.9 Mbp range, only 3 SNPs were observed between the divergence events of DE/CF and C/F, indicating that DE, C, and F likely emerged subsequently in less than a thousand years. After diverged from Haplogroup C, no major split was observed in F for 18 thousand years, suggesting a strong bottleneck of F lineage. It should be noted that all the primary haplogroups (G, J, N, O, Q, and R) emerged before the last glacial maximum (LGM, ~20 kya), and most of the presently known East Eurasian clades have branched off in the late Upper Paleolithic Age (before 10 kya). All divergences on this tree before 7 kya were binary, suggesting that during the Paleolithic Age, slow population growth and bottlenecks or drift eradicated most of the ever existing clades [22].

The most surprising discovery in the tree is the three star-like expansions in Haplogroup O3-M324, i.e. under the M117 clade, the M134xM117 paragroup, and the 002611 clade. Here we denote the three star-like expansions as Oα, Oβ, and Oγ, respectively (see Discussion). Since the sample selection for high-throughput sequencing was intended for representing a wide variety of clades in East Asian populations, a star-like expansion indicates successful expansion of male lineages within a very short period (<500 years). These three clades are present with high frequency across many extant East Asian populations [23,24] and encompass more than 40% of the present Han Chinese in total (estimated 16% for Oα, 11% for Oβ, and 14% for Oγ) [20]. It is

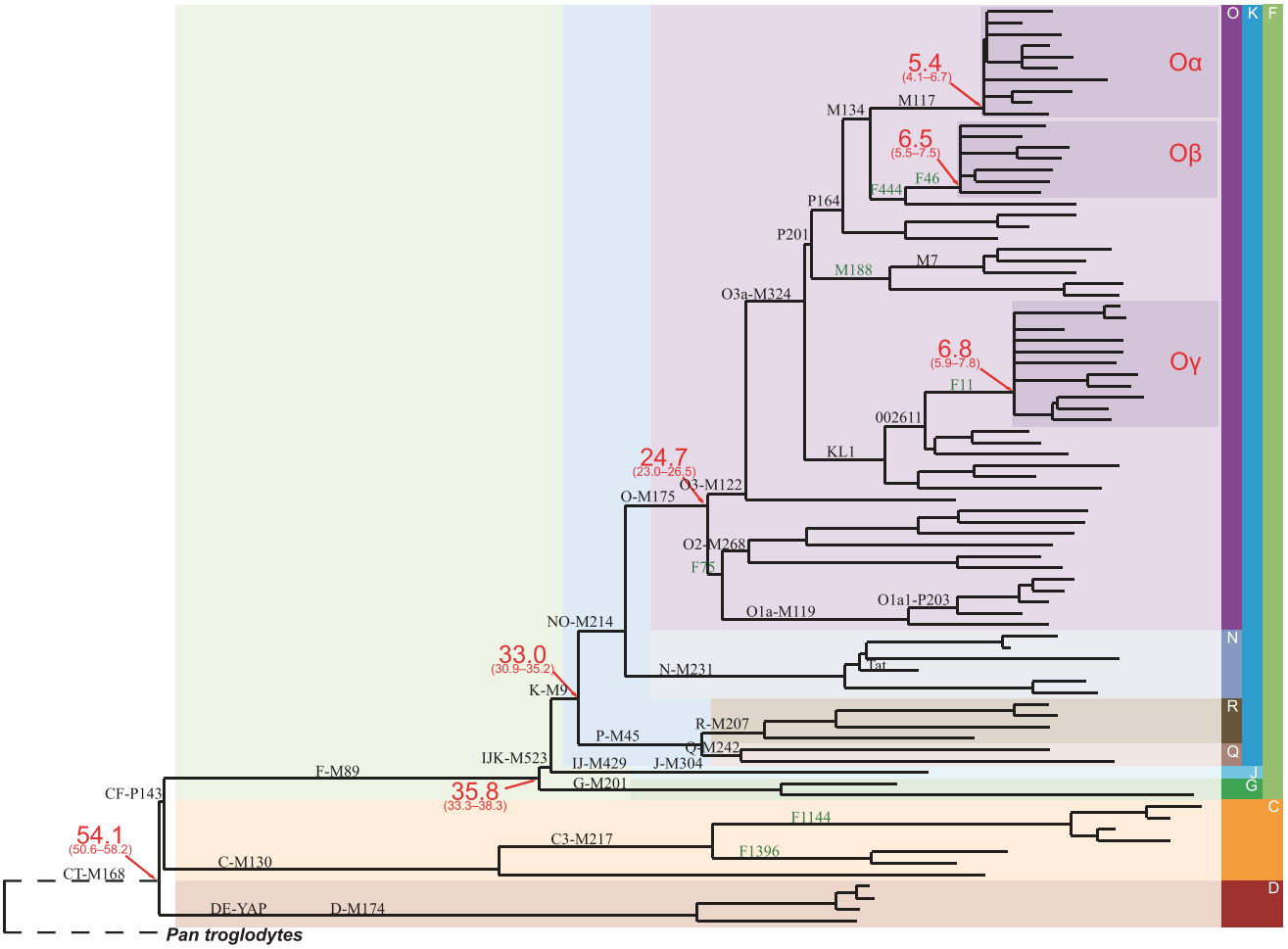

Figure 1. Phylogenetic tree of human Y chromosome, emphasizing the three star-like expansions (Oα, Oβ, Oγ). The tree was constructed from 78 samples sequenced in this study, together with three published East-Asian genomes and a chimpanzee genome. The branch lengths (horizontal lines) are proportional to the number of SNPs on the branch. Numbers in red indicate the coalescence time (in years, considering the variation in SNP counting, but ignoring uncertainty in mutation rate) and 95% confidence intervals of the node. For more details, see Fig. S1.
doi:10.1371/journal.pone.0105691.g001

conspicuous that roughly 300 million extant males are the patrilineal progenies of only three males in the late Neolithic Age.

The expansion dates are estimated 5.4 kya for Oα, 6.5 for Oβ, and 6.8 for Oγ (Fig. 1), after the shift to intensive agriculture in North China (since 6.8 kya) [25,26], in particular, during the Yangshao Culture (6.9–4.9 kya) in Central Yellow River Basin, Majiayao Culture (6.0–4.9 kya) in the Upper Yellow River Basin, and the Beixin (7.4–6.2 kya) – Dawenkou Culture (6.2–4.6 kya) in the Lower Yellow River Basin [27]. We therefore propose that in the late Neolithic Age, the three rapidly expanding clans established the founding patrilineal spectrum of the predecessors in East Asia. Since all the sequenced Han Chinese M117+ samples are under the Oα expansion, and M117+ subclade exists in moderate to very high frequency in many Tibeto-Burman ethnic groups [28,29,30], it would be of interest to know when the M117+ individuals in other ethnic groups diverged with the ones in Han Chinese, and whether they are also under the Oα expansion, in order to trace the origin and early history of Sino-Tibetan language family.

This study shows that all the strongly expanding Y chromosomal haplogroups (i.e. O-M175 or C-M130) had already migrated to East Asia more than 20 thousand years before their Neolithic expansion, thus supporting a boom of local farmers in China, which is consistent with the independent origin of agriculture [31], while differing from the case in Europe, where immigrant farmers from the Middle East contributed to the majority of modern Y chromosomes [32].

Discussion

Although most of the sequences in this study were obtained from individuals in China, the haplogroup representation (C, D, G, J, N, O, Q, and R) already enabled us to calculate the times of most of the major divergence events outside of Africa, like G/IJK, NO/P etc., since the times were achieved using the hypothesis of molecular clock, and the results of divergence time between haplogroups would not be affected by from whichever continent or country the individuals were sampled. One good sequence from each of two haplogroups is enough for calculating their divergence time, and more sequences could only help to enhance the precision but would not greatly change the result.

The significant improvement of accuracy of dating in this study comparing to former East Asian studies is attributed to the large number of newly discovered SNPs. It is noted that the relative standard deviation of calculated divergence time is in inverse proportion to the square root of observed SNP occurrence (see Discussion S1). Furthermore, the average counts of SNPs from the common ancestor of CF/DE to a modern individual is 210 in this study, limiting the theoretical 95%CI to only ±13.6%, comparing to 9 SNPs on average in the previous study with the 95%CI over ±60% [16]. Considering that 3.9 Mbp range constitutes only less than half of 10 Mbp non-repetitive region in Y chromosome [7], the time resolution of east Asian Y chromosome phylogeny is expected to be doubled in the near future.

The determination of mutation rate is a crucial question in calculation of the absolute divergent times, which caused the most dating differences among the studies[7,33]. As revealed by previous studies, this inconsistency of mutation rate was resulted from two aspects: among different regions of the chromosome, and between older and younger time scales. The former has been disclosed in a study of autosomes, that the base substitution rate of CpG bases is 9.5-fold that of non-CpG bases [34], as well as for mitochondrial DNA, the substitution rate was not only differentiated between coding and control regions, but also in a base-by-base manner [35]. It is worth to point out that recently, Wei et al. published a similar study about Y chromosome sequencing of 36 individuals (mainly Haplogroup R1b and E1b), in which 3.15 or 8.83 Mbp range was sequenced [19], and they achieved a time of out-of-Africa at 57–74 kya using various methods, which is slightly older than our result (54 kya), although the same mutation rate of 1×10^{-9} substitution/base/year were employed. The difference could be ascribed to the regions chosen for date estimation; we compared the regions that Wei et al. and we studied, and found that in their study, the SNP density in the region that was sequenced only in their study is significantly higher than that in the region that both studies have sequenced ($P<0.005$) (Table S4).

The difference between long-term (evolutionary) and short-term (genealogical) mutation rates has also been observed before. For calculating the divergence time using Y-chromosomal STR (short tandem repeat), the father-son mutation rate is about three times the "evolutionary" [36]; similar rate difference was also observed for mitochondrial nucleotide substitution rate [35,37]. This controversy is usually explained by selection on deleterious mutations [38]. The autosomal genealogical substitution rate was estimated at 1.2×10^{-8} substitution/base/generation [34,39], which is less than half of the rate we used in this study. However, due to that 80–85% of de novo mutations are attributed to the father's side [34], and that the Y chromosome contains the least genes among the chromosomes and thus underwent lessened purifying selection [40], the mutation rate used in this study is still compatible with the previous studies.

We also compared human and chimpanzee Y chromosomes (see SI Methods), and found ~45,800 substitutions between the two species which fall into the range that we compared for human samples; roughly 1/4 intra-human SNPs have no homologous loci on chimpanzee Y chromosome. Assuming the divergence between human and chimpanzee was at ~6,000 kya [41] and a constant substitution rate, the divergence time for DE and CF would be only 40 kya, which is younger than our result. This suggests that base-substitution rate between human and chimpanzee is higher than the rate inside human species, which can be explained with the huge interspecies difference of the Y-chromosome structures, and the observation that the chimpanzee Y chromosomal genes decayed faster than human [40].

To overcome the factors for uncertainty of mutation rate, a calibration with series of samples of comparable time scales might be used. For the case of mitochondrial DNA, a recent study, in which several C-14 calibrated ancient complete sequences (4–40 kya) were incorporated into the tree, made the absolute dates much more convincing [42], and we expect a parallel calibration for the Y chromosome in the near future.

Despite of the mutation rate uncertainty, we evaluate our calculation of absolute divergence time as acceptable. Firstly, our out-of-Africa date (54.1 kya) is still within the range of previous estimations (39–74.6 kya). Secondly, the out-of-Africa date is similar to the recent estimation of two great mitochondrial expansions outside Africa – M (49.6 kya) and N (58.9 kya) [43]. Thirdly, it is not contradictory to the emergence of earliest modern human fossil out of Africa (e.g. ~ 50 kya in Australia) [44].

The accumulative substitution count from the DE/CF divergence to a modern individual varies from to 168 (YCH113) to 241 (YCH198). Despite of this variation, by testing the assumption of molecular clock for the tree, the null hypothesis of a molecular clock could not be rejected ($P>0.05$) using PAML package v4.4 [45] with the GTR model, unlike the mitochondrial tree from complete sequences, which showed violation to the clock assumption [43]. Part of the branch length variation may come from the false negative detection of SNPs, especially on a long

terminal branch; however, this effect was mostly eliminated that we chose only the sequences with good quality for time estimation, so the branch length difference for these sequences should mainly reflects the real variation, and should have little effect in time estimation.

The current Y haplogroups were named according to the rule of Y Chromosome Consortium (YCC) [14]. Along with increasing clades being discovered, the present nomenclature became cumbersome in some cases, e.g. 'R1b1b2a1a2d3a' in the ISOGG tree 2010 (http://www.isogg.org), which is prone to frequent name changes and hard to remember. Another commonly used nomenclature such as 'O-M117' or 'O-F46' is also not suitable for a determined star-like expansion, since there are many SNPs found ancestral to the star point, while these SNPs may be found not all equivalent in the future, e.g. some individuals might be found M117+ but not belonging to the star expansion, then the name of the star point must be renewed. Therefore, here we propose a modification to the current nomenclature system: for any important star-like expansion that leads to large population (e.g. several millions) and multiple lineages ($\geq$5) in short time as revealed by long-range sequencing (>1 Mbp was needed in order to limit the expansion within 1,000 years), a lineage name with lowercase Greek letter is applied directly after the Latin capital letter of the first-class haplogroup name. For example, the star-like expansions under M117, F46, and F11 are now named as Oα, Oβ, and Oγ, respectively, and their downstream lineages should still be named following the rule of YCC 2002, with Arabic number succeeding the Greek letter, e.g. Oα1a1. These names of the star-like expansions are not bound to any single defining SNP (e.g. M117), but to the expansion itself, i.e. the expansion names should always keep unchanged despite new side clades would be found to its upstream, in order to keep the nomenclature stable. For the currently equivalent SNPs on the branch leading to the expansion, we will know the occurring order only after vast amount of samples being genotyped for those SNPs.

Since all the Paleolithic divergences of Y chromosome lineages are binary, the three roughly contemporaneous star-like expansions revealed in this study indicate a remarkable demographic change in the late Neolithic Age. The earliest agriculture in North China emerged before 10 kya [46], however, no distinct Y chromosomal expansion could be related to this event. The three star-like expansions happened several thousand years later, thus are likely linked to middle Neolithic cultures such as Yangshao (6.9–4.9 kya) and Dawenkou Culture (6.2–4.6 kya) in the Yellow River Basin [27]. During this period, agriculture became mature and intensive, and the majority of human diet shifted from food collection into production [47,48]. Crop harvest constituted a more stable food source than hunting and gathering, and enabled nourishing population at higher density. In addition, liberation of males from hazardous hunting might have enhanced male viability into adulthood, thus the effective population size of Y chromosome increased. Besides the progress in agriculture, changes in social structure might also contribute to the patrilineal expansion. In the middle and late phases of Yangshao and Dawenkou culture, the burial customs showed a gradual transition from an egalitarian matrilineal society into a hierarchical patrilineal one [49,50]. Interestingly, the major maternal expansions in China shown by mitochondrial tree (among which are also several star-shaped expansions) occurred much earlier, at the late Paleolithic Age [3]. This immense non-synchrony between maternal and paternal expansion suggests a possible transition of social structure, that in the late Neolithic Age, a few paternal lineages achieved greater advantage on the existing basis of the population that started expansion since the Paleolithic Age. After the strong Neolithic expansions, the reproductive advantage of the farmers lasted for 4,000 years, until most of the gatherer-and-hunter tribes in the Yellow River Basin were absorbed by the farming societies of Huaxia, from which the Han ethnicity was formed.

Although without ancient DNA proofs, we cannot yet confirm the initial expanding regions of these three clans, whether they were original in the middle or lower reach of Yellow River Valley or migrated from the vicinity, we are now at least certain that a majority of Han Chinese did derive from just a few patrilineal ancestors in the Neolithic Age. Whether each of them could be related to the legendary Emperors *Yan* and *Huang* or their tribes, is to be solved with more prudence and with the help of interdisciplinary genetic, archeological, ethnical, and documentary studies.

Methods

Ethics Statement

The study was under the approval of the Ethics Committee of Biological Research at Fudan University. All the samples were collected with the informed consent signed by the sample donors.

Samples

We collected whole blood from ~800 Chinese male volunteers. Genomic DNA was extracted using QIAamp DNA Blood Mini Kit (QIAGEN, Hilden, NRW, Germany). SNaPshot multiplex kit (ABI, Carlsbad, CA, US) was used for typing Y chromosomal SNPs according to the most recent phylogenetic tree [13,20], and 17 Y-STRs were determined with Y-filer kit (ABI, Carlsbad, CA, US). We selected 110 samples for next-generation sequencing, considering Y haplogroup, STR haplotype, as well as ethnic origin, in order to represent a broad spectrum of Y chromosome lineages of Chinese populations (Table S1). The selected samples covered most sublineages of Haplogroup O (72 samples), as well as Haplogroup C, D, G, J, N, Q, and R.

Library preparation

Genomic DNA of the selected samples were sheared using Bioruptor UCD-200 (Diagenode, Liège, Belgium) to 200–250 bp length, then were fixed to blunt-end, added 3′-A tail, and ligated with barcode-linked Illumina paired-end adaptors (Table S1). Ligation products were amplified by PCR, and 300 – 350 bp sections were extracted through agarose gel electrophoresis. Except for one sample (YCH53), the others were pooled into 8 pools, with 10–15 samples in equal amount in each pool (Table S1). NRY was enriched using custom designed bait library (see below) of G3360-90000 SureSelect kit for Illumina paired-end (Agilent, Santa Clara, CA, US) (the baits were listed in Table S5). After another round of amplification, the pools went through single-end or paired-end sequencing with either GAIIx or HiSeq2000 sequencer for 100 or 2×100 cycles (Illumina, San Diego, CA, US).

Bait design

For Agilent SureSelect enrichment, bait library was designed with the following procedures: we first simulated reads mapping by generating 70-bp fragments of reference Y chromosome (hg18 or NCBI build36) (http://hgdownload.cse.ucsc.edu/downloads.html#human) for each 10 bp, e.g. chrY:1-70, chrY:11-80 etc. The fragments were then mapped on the complete hg18 genome using soap2 aligner (http://soap.genomics.org.cn/)[51]. All the match results with 0–2 mismatch bases on all chromosomes were summed up, and only the fragments without any repetitive matches (on either Y or other chromosome) were kept as unique

fragments. The range of those unique fragments was combined, and the combined ranges that are at least 240 bp long were selected for bait design on Agilent eArray website (https://earray.chem.agilent.com/earray/). Totally 40,379 baits covering 4,292,864 bp were successfully designed and ordered for production. The ranges (on hg18) of the generated baits are listed in Table S5.

Processing of next-generation sequencing data

The barcodes were removed and the reads were assigned to each sample. For paired-end sequencing, the reads were assigned only when the both barcodes were the same. The reads were mapped to hg18 using *bwa* aligner (version 0.5.8) [52], and sam files were generated. Reads that were uniquely mapped on Y chromosome were extracted and transformed into bam file with *samtools* (version 0.1.8) [53]. Duplicates were removed by either Picard's *MarkDuplicate* (http://picard.sourceforge.net) (for single-end) or *samtools rmdup* (for paired-end). Indels were re-aligned using GATK [54,55], and after *samtools mpileup*, variations were called under the following criteria: for one sample, the position where the alternative allele (compared to hg18) must be $\geq 2\times$ coverage and at the same time $\geq 3/4$ of total coverage. All the variance candidates were collected, and genotypes were called on all the sequenced samples. Out of those candidates, SNPs were semi-manually filtered considering consistency to the Y chromosomal phylogeny, coverage (especially for the private SNPs, a minimum of $4\times$ was required), and flanking sequence (to avoid those included or next to a homopolymer or an STR). Three other publicly available East Asian genomes, YanHuang (YH) (O1a1-P203) [56], KoRef (SJK) (O2b-M176) [57], and GMIAK1 (O3a2c*-P164xM134) [58] were also included in analysis.

Time estimation of the nodes in the phylogenetic tree

A coverage filter was applied for time estimation, i.e., only the loci with good coverage among the sequenced samples, i.e., with more than 100 out of 110 SNP calling results with an unambiguous 0 or 1 were selected for phylogenetic reconstruction (0 for same as reference, 1 for mutation, question mark "?" when neither reference or alternative counts for more than 3/4 for this sample, and a minus mark "-" for no coverage. Beside these, "x" for tested 1 but should be 0, and "@" for tested ? but should be 1. The "x" and "@" were manually determined according to tree topology and the pattern of barcode confusion. See Table S6). SNPs were extracted into pseudo-sequences, and a maximum parsimony tree of only good- and moderate-quality sequences was calculated using ARB program [59].

To avoid uncertainty in downstream branches that might influence branch lengths, we used only good-quality sequences for Bayesian time estimation (with on average $>6\times$ coverage at targeted regions and no obvious mislabeling, see Table S1). We used BEAST [21] for calculating the divergence time of each node in the phylogenetic tree. All the 47 high-quality sequences together with YH and SJK were used for time estimation. We generated pseudo-sequences from these 49 individuals, with the above described loci (>100 out of 110 with unambiguous result) that have polymorphism among the 49 individuals, 3823 bases in total. For missing or ambiguous data, we imputed the result following the tree topology (for the very rare case that imputation doesn't work, we assigned a random base from the two possible allele). We determined appropriate DNA substitution model with MrModeltest 2.3 [60] for subsequent Bayesian MCMC analysis. For Bayesian MCMC analysis, the times of each cluster were estimated using BEAST1.6.1 [21,61]. Each MCMC sample was based on a run of 20 million generations sampled every 10,000 steps with the first 2 million generations regarded as burn-in. To test the assumption of molecular clock for the tree, we used PAML package v4.4 with the GTR model. The null hypothesis of a molecular clock cannot be rejected ($P>0.05$) by comparison between the models. We used the GTR model of nucleotide substitution determined with MrModeltest 2.3 with a strict clock. The single nucleotide substitution rate was set as 1×10^{-9}/nucleotide/year. The effective sample size of the coalescent prior was above 900. A relaxed clock was also employed for comparison and the results were similar.

Chimpanzee genome (panTro3) [62] (http://hgdownload.cse.ucsc.edu/downloads.html#chimp) was used as comparison for time estimation. All the base substitutions of chimpanzee genome comparing to hg18 were discovered using the similar method as for bait design: the simulated 100-bp-long reads at each 10 bp were generated and mapped onto hg18 using *bwa*, and SNPs were discovered. The SNPs intra human beings were also called for the chimp reads. The root for the human samples in this study was thus determined.

Supporting Information

Discussion S1 Supporting Discussions.
(DOCX)

Figure S1 Phylogenetic tree of human Y chromosome. The tree is constructed from 78 samples sequenced in this study, together with three published East-Asian genomes (YH, SJK, GMIAK1) and a chimpanzee genome (Pan), which are labeled with '*'. Except for YCH145 (Spanish), SJK and GMIAK1 (both Korean), all the human samples are Chinese. The branch lengths (horizontal lines) are proportional to the number of SNPs on the branch, and the SNP numbers are labeled under the branches). The SNPs labeled on the horizontal lines are only representative. The SNPs labeled in green represent newly recognized clades in this study. The estimated coalescence time (in years) for the nodes are calculated only from good-quality ($>6\times$ coverage) human sequences (in bold italic) by BEAST with relaxed clock (see SI Methods), and the numbers in brackets are for 95% confidence intervals (ignoring uncertainty in mutation rate).
(EPS)

Figure S2 Revised migration routes of modern human. (a) Split of the first out-of-Africa ancestor and early migration in Asia. (b) The emergence of the main haplogroups before and during the LGM. (c) Foundation of present haplogroup distribution before 8 kya. (d) Major population expansion events in East Asia (shaded) in the Neolithic Age and their probable relationship with modern language families.
(EPS)

Table S1 Samples list.
(XLS)

Table S2 Newly named SNPs.
(XLS)

Table S3 Ranges that were covered by at least 100 of all the 110 samples.
(XLS)

Table S4 Comparison of SNPs discovered in this study and Wei et al. (2012).
(XLS)

Table S5 Bait regions designed for Agilent SureSelect capturing (positions are for chrY of hg18).
(XLS)

Table S6 Genotyping results.
(XLS)

Author Contributions

Conceived and designed the experiments: SY LJ. Performed the experiments: SY CCW ZDQ. Analyzed the data: SY HXZ WW LHW XDP WQF YGH WFJ LJ. Contributed reagents/materials/analysis tools: YW LJX SLL YA HL. Wrote the paper: SY LJ.

References

1. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475: 493–U484.
2. Cavalli-Sforza LL, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. Nat Genet 33: 266–275.
3. Zheng H-X, Yan S, Qin Z-D, Wang Y, Tan J-Z, et al. (2011) Major Population Expansion of East Asians Began before Neolithic Time: Evidence of mtDNA Genomes. PLoS One 6.
4. Pereira L, Richards M, Goios A, Alonso A, Albarran C, et al. (2005) High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. Genome Res 15: 19–24.
5. Gignoux CR, Henn BM, Mountain JL (2011) Rapid, global demographic expansions after the origins of agriculture. Proc Natl Acad Sci USA 108: 6044–6049.
6. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, et al. (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423: 825–U822.
7. Xue YL, Wang QJ, Long Q, Ng BL, Swerdlow H, et al. (2009) Human Y Chromosome Base-Substitution Mutation Rate Measured by Direct Sequencing in a Deep-Rooting Pedigree. Curr Biol 19: 1453–1457.
8. Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, et al. (2001) Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol 18: 1189–1203.
9. Jin L, Su B (2000) Natives or immigrants: Modern human origin in East Asia. Nat Rev Genet 1: 126–133.
10. Underhill PA, Shen PD, Lin AA, Jin L, Passarino G, et al. (2000) Y chromosome sequence variation and the history of human populations. Nat Genet 26: 358–361.
11. Jobling MA, Tyler-Smith C (2003) The human Y chromosome: An evolutionary marker comes of age. Nat Rev Genet 4: 598–612.
12. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
13. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, et al. (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18: 830–838.
14. The Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res 12: 339–348.
15. Ke YH, Su B, Song XF, Lu DR, Chen LF, et al. (2001) African origin of modern humans in East Asia: A tale of 12,000 Y chromosomes. Science 292: 1151–1153.
16. Cruciani F, Trombetta B, Massaia A, Destro-Bisol G, Sillito D, et al. (2011) A Revised Root for the Human Y Chromosomal Phylogenetic Tree: The Origin of Patrilineal Diversity in Africa. Am J Hum Genet 88: 814–818.
17. Thomson R, Pritchard JK, Shen PD, Oefner PJ, Feldman MW (2000) Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. Proc Natl Acad Sci USA 97: 7360–7365.
18. Hammer MF, Zegura SL (2002) The human Y chromosome haplogroup tree: Nomenclature and phylogeography of its major divisions. Annu Rev Anthropol 31: 303–321.
19. Wei W, Ayub Q, Chen Y, McCarthy S, Hou Y, et al. (2013) A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res 23: 388–395.
20. Yan S, Wang CC, Li H, Li SL, Jin L, et al. (2011) An updated tree of Y-chromosome Haplogroup O and revised phylogenetic positions of mutations P164 and PK4. Eur J Hum Genet 19: 1013–1015.
21. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7: 214.
22. Jobling MA, Hurles ME, Chris T-S (2003) Human Evolutionary Genetics: Origins, Peoples and Disease. Garland Science.
23. Karafet TM, Hallmark B, Cox MP, Sudoyo H, Downey S, et al. (2010) Major East-West Division Underlies Y Chromosome Stratification across Indonesia. Mol Biol Evol 27: 1833–1844.
24. Su B, Xiao JH, Underhill P, Deka R, Zhang WL, et al. (1999) Y-chromosome evidence for a northward migration of modern humans into eastern Asia during the last Ice Age. Am J Hum Genet 65: 1718–1724.
25. Barton L, Newsome SD, Chen FH, Wang H, Guilderson TP, et al. (2009) Agricultural origins and the isotopic identity of domestication in northern China. Proc Natl Acad Sci USA 106: 5523–5528.
26. Bettinger RL, Barton L, Morgan C (2010) The Origins of Food Production in North China: A Different Kind of Agricultural Revolution. Evol Anthr 19: 9–21.
27. The Institute of Archaeology Chinese Academy of Social Sciences (2010) Chinese Archaeology - Neolithic. Beijing: China Social Sciences Press.
28. Gayden T, Cadenas AM, Regueiro M, Singh NB, Zhivotovsky LA, et al. (2007) The Himalayas as a directional barrier to gene flow. Am J Hum Genet 80: 884–894.
29. Shi H, Dong YL, Wen B, Xiao CJ, Underhill PA, et al. (2005) Y-chromosome evidence of southern origin of the East Asian - Specific haplogroup O3-M122. Am J Hum Genet 77: 408–419.
30. Xue YL, Zejal T, Bao WD, Zhu SL, Shu QF, et al. (2006) Male demography in East Asia: A north-south contrast in human population expansion times. Genetics 172: 2431–2439.
31. Diamond J, Bellwood P (2003) Farmers and their languages: The first expansions. Science 300: 597–603.
32. Balaresque P, Bowden GR, Adams SM, Leung H-Y, King TE, et al. (2010) A Predominantly Neolithic Origin for European Paternal Lineages. PLoS Biol 8.
33. Kuroki Y, Toyoda A, Noguchi H, Taylor TD, Itoh T, et al. (2006) Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. Nat Genet 38: 158–167.
34. Campbell CD, Chong JX, Malig M, Ko A, Dumont BL, et al. (2012) Estimating the human mutation rate using autozygosity in a founder population. Nat Genet 44: 1277–1281.
35. Soares P, Ermini L, Thomson N, Mormina M, Rito T, et al. (2009) Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. Am J Hum Genet 84: 740–759.
36. Zhivotovsky LA, Underhill PA, Cinnioğlu C, Kayser M, Morar B, 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: 50–61.
37. Heyer E, Zietkiewicz E, Rochowski A, Yotova V, Puymirat J, et al. (2001) Phylogenetic and familial estimates of mitochondrial substitution rates: Study of control region mutations in deep-rooting pedigrees. Am J Hum Genet 69: 1113–1126.
38. Penny D (2005) Evolutionary biology - Relativity for molecular clocks. Nature 436: 183–184.
39. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. Nature 488: 471–475.
40. Hughes JF, Skaletsky H, Pyntikova T, Minx PJ, Graves T, et al. (2005) Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. Nature 437: 100–103.
41. Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D (2006) Genetic evidence for complex speciation of humans and chimpanzees. Nature 441: 1103–1108.
42. Fu Q, Mittnik A, Johnson PLF, Bos K, Lari M, et al. (2013) A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. Curr Biol 23: 1–7.
43. Behar DM, van Oven M, Rosset S, Metspalu M, Loogvali E-L, et al. (2012) A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from its Root. Am J Hum Genet 90: 675–684.
44. Roberts RG, Jones R, Smith MA (1990) Thermoluminescence Dating of a 50,000-Year-Old Human Occupation Site in Northern Australia. Nature 345: 153–156.
45. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.
46. Yang X, Wan Z, Perry L, Lu H, Wang Q, et al. (2012) Early millet use in northern China. Proc Natl Acad Sci USA 109: 3726–3730.
47. Zhao Z (2010) New data and new issues for the study of origin of rice agriculture in China. Archaeol Anthropol Sci 2: 99–105.
48. Fuller DQ (2011) Pathways to Asian Civilizations: Tracing the Origins and Spread of Rice and Rice Cultures. Rice 4: 78–92.
49. Zhang Z-H (2004) Zhongguo Xinshiqi Shidai Kaogu (Neolithic Archaeology of China). Nanjing: Nanjing University Press.
50. Jiao T (2001) Gender Studies in Chinese Neolithic Archaeology. In: Arnold B, Wicker NL, editors. Gender and the archaeology of death: AltaMira Press. pp.51–61.
51. Li RQ, Yu C, Li YR, Lam TW, Yiu SM, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966–1967.
52. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.
53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.
54. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491–498.
55. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303.
56. Wang J, Wang W, Li RQ, Li YR, Tian G, et al. (2008) The diploid genome sequence of an Asian individual. Nature 456: 60–U61.

57. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, et al. (2009) The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. Genome Res 19: 1622–1629.
58. Kim JI, Ju YS, Park H, Kim S, Lee S, et al. (2009) A highly annotated whole-genome sequence of a Korean individual. Nature 460: 1011–1015.
59. Ludwig W, Strunk O, Westram R, Richter L, Meier H, et al. (2004) ARB: a software environment for sequence data. Nucleic Acids Res 32: 1363–1371.
60. Nylander JAA (2004) MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.
61. Drummond AJ, Ho SY, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4: e88.
62. The Chimpanzee Sequencing and Analysis Consortium2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69–87.