Y-Chromosome Variation in Southern African Khoe-San Populations Based on Whole-Genome Sequences

Thijessen Naidoo[1,2,3,4,†], Jingzi Xu[1,†], Mário Vicente[1], Helena Malmström[1,5], Himla Soodyall[6,7,8], Mattias Jakobsson[1,3,5], and Carina M. Schlebusch[1,3,5,*]

[1]Human Evolution, Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Sweden

[2]Department of Archaeology and Classical Studies, Stockholm University, Sweden

[3]Science for Life Laboratory, Uppsala, Sweden

[4]Centre for Palaeogenetics, Stockholm, Sweden

[5]Palaeo-Research Institute, University of Johannesburg, Auckland Park, South Africa

[6]Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[7]National Health Laboratory Service, Johannesburg, South Africa

[8]Academy of Science of South Africa

*Corresponding author: E-mail: carina.schlebusch@ebc.uu.se.

Accepted: 12 May 2020

[†]These authors contributed equally to this work.

Data deposition: The complete Y-chromosome sequences were deposited on the European Genome Phenome Archive (https://www.ebi.ac.uk/ega/), accession number EGAS00001004459, and are available for research use under controlled access policies.

Abstract

Although the human Y chromosome has effectively shown utility in uncovering facets of human evolution and population histories, the ascertainment bias present in early Y-chromosome variant data sets limited the accuracy of diversity and TMRCA estimates obtained from them. The advent of next-generation sequencing, however, has removed this bias and allowed for the discovery of thousands of new variants for use in improving the Y-chromosome phylogeny and computing estimates that are more accurate. Here, we describe the high-coverage sequencing of the whole Y chromosome in a data set of 19 male Khoe-San individuals in comparison with existing whole Y-chromosome sequence data. Due to the increased resolution, we potentially resolve the source of haplogroup B-P70 in the Khoe-San, and reconcile recently published haplogroup A-M51 data with the most recent version of the ISOGG Y-chromosome phylogeny. Our results also improve the positioning of tentatively placed new branches of the ISOGG Y-chromosome phylogeny. The distribution of major Y-chromosome haplogroups in the Khoe-San and other African groups coincide with the emerging picture of African demographic history; with E-M2 linked to the agriculturalist Bantu expansion, E-M35 linked to pastoralist eastern African migrations, B-M112 linked to earlier east-south gene flow, A-M14 linked to shared ancestry with central African rainforest hunter-gatherers, and A-M51 potentially unique to the Khoe-San.

Key words: Y chromosome, next-generation sequencing, haplogroups, Khoe-San, southern Africa.

Introduction

The male-specific portion of the Y chromosome (MSY) has long been regarded as an effective tool in the study of human evolutionary history (Underhill and Kivisild 2007). It has proved useful mainly due to a lack of recombination along its length, making it the longest haplotypic block in the human genome (Scozzari et al. 2012); and its paternal mode of inheritance. The transmission of an intact haplotype from father to son, changing only through mutation, preserves a simpler record of its history and allows us to study the male contribution to the shaping of humanity.

The mutations found on Y chromosomes sourced from numerous human populations have been used to generate Y-chromosome phylogenies (Underhill et al. 2000; Hammer et al. 2001; Y-Chromosome Consortium 2002; Karafet et al. 2008), with well-defined and geographically informative haplogroups, that is, groups of haplotypes that share common ancestors and so present as clades in a phylogeny. While the first consensus phylogeny with standardized nomenclature was published in 2002 (Y-Chromosome Consortium 2002) with the last full version published in 2008 (Karafet et al. 2008) and a minimal reference version published in 2014 (van Oven et al. 2014), the International Society of Genetic Genealogy (ISOGG) maintains a comprehensive online version (https://isogg.org/tree/index.html) that is updated every year.

The utility of the Y-chromosome phylogeny notwithstanding, the ascertainment bias present when initially sourcing the mutations used to build it, resulted in limitations. The use of predefined sets of variants, limited the ability to generate unbiased estimates of global diversity and accurate estimates of the time to most recent common ancestor (TMRCA) (Poznik et al. 2013). As a solution, short tandem repeat polymorphisms (STRs) have been used for a long time, however, STRs had their own biases and issues (i.e., see Hallast et al. [2015] for a comparison between sequence-based and STR-based TMRCAs). The emergence of next-generation sequencing (NGS) technology, resulted in the discovery of thousands of unbiased variants (Cruciani et al. 2011; Francalacci et al. 2013; Poznik et al. 2013; Scozzari et al. 2014; Karmin et al. 2015; Barbieri et al. 2016), which allowed for substantially more accurate estimates of the age the Y-chromosome phylogeny and its haplogroups.

Although studies usually attempted to balance the global representation of populations in their samples, there did appear to be an underrepresentation of African samples in the final results, with this being especially true of haplogroups A and B. Barbieri et al. (2016) addressed this discrepancy by sequencing a portion of the Y chromosome in 547 Khoe-San- and Bantu-speakers; resulting in much older estimates of the ages of haplogroups A and B and their subclades.

Following high-coverage sequencing of the full Y chromosome in a data set of 19 male Khoe-San individuals and comparison with existing full Y-chromosome sequence data, the present study describes the distribution of haplogroups found, in the wider African context. The high level of resolution afforded by a sequencing analysis allowed us to potentially resolve the source of haplogroup B-P70 in the Khoe-San. The use of two slightly differing phylogenies when assigning our haplogroups, allowed for the reconciliation of the haplogroup A-M51 data from Barbieri et al. (2016) with the most recent version of the ISOGG Y-chromosome phylogeny.

Results

We performed high-coverage whole-genome sequencing of 25 Khoe-San individuals from five different populations (Schlebusch et al. 2020). In this study, we discuss the results of the Y-chromosome sequences of the 19 male individuals that were included in the study. After processing and filtering the sequence data (see Materials and Methods), we obtained 5,783 variants; with an average depth of 31×. Once merged with the comparative data from seven additional populations (Drmanac et al. 2010; Auton et al. 2015; Lachance et al. 2012), our total data set contained 7,878 variants from 8.8 Mb (8,800,463 bases) of Y-chromosome sequence in 48 individuals.

Khoe-San Haplogroups

The major haplogroups found in our sequenced individuals were A-M14 (A1b1a1), A-M51 (A1b1b2a), B-M112 (B2b), E-M2 (E1b1a1), and E-M35 (E1b1b1); and were thus strongly concordant with previous surveys of Y-chromosome variation in Khoe-San populations (Underhill et al. 2000; Naidoo et al. 2010; Batini et al. 2011; Barbieri et al. 2016). The additional resolution provided by sequencing, however, uncovered several new variants especially supporting branches in haplogroups A and B, and allowed for greater clarity regarding the relationships of some subclades and markers (see table 1 and supplementary table S1, Supplementary Material online, for haplogroup assignments and population information).

The Internal Structure of A-M51

Haplogroup A1b1b2a (A-M51) has often been found to be the most common haplogroup A subclade in southern Africa (Barbieri et al. 2016), though usually found primarily in the Khoe-San or in populations with significant levels of Khoe-San admixture. Three previously reported subclades (Scozzari et al. 2012; Barbieri et al. 2016), haplogroups A1b1b2a1a (A-P71), A1b1b2a2 (A-V37), and A1b1b2a1b (A-V306), were found in our data set; and the branching structure was further refined and reconciled with the ISOGG Y phylogeny (ISOGG 2019-2020) (fig. 1 and table 1). Haplogroup A1b1b2a2 is the first to branch off, as reported by Barbieri et al. (2016), whereas haplogroups A1b1b2a1a and A1b1b2a1b are united by marker M239. Within haplogroup A1b1b2a1a, all three individuals were also derived at marker P102, with one of them ancestral at marker P291. This is at odds with the current ISOGG phylogeny, which has P291 basal to P102. The three subclades segregated independently among the populations, with A1b1b2a2 found in two !Xun individuals, A1b1b2a1a in three Nama individuals and A1b1b2a1b in two Ju|'hoansi individuals.

Gene Flow into Khoe-San Populations

Haplogroup E1b1b1b2b2a1 (E-M293) was found in two Khoe-San: one Nama and one !Xun. This E-M35 subclade was previously linked to the movement of pastoralist groups from eastern Africa to southern Africa ~2,000 years ago (Henn et al. 2008; Bajic et al. 2018). Moreover, the

Table 1
Population Information and Y-Chromosome Haplogroups of Individuals Used in the Study

Ref.	Sample Code	Population	AMY-Tree Haplogroup	AMY-Tree Marker	ISOGG Haplogroup
1	KSP103	Ju|'hoansi	A2a1b1	A-M114	A1b1a1a2a
1	KSP154	!Xun	A2a1b1	A-M114	A1b1a1a2a
1	KSP116	Ju|'hoansi	A2a1b2	A-P262	A1b1a1a2b
1	KSP139	Nama	A3b1a	A-P71	A1b1b2a1a
1	KSP124	Nama	A3b1a	A-P71	A1b1b2a1a1
1	KSP140	Nama	A3b1a	A-P71	A1b1b2a1a1
1	KSP105	Ju|'hoansi	A3b1c	A-V306	A1b1b2a1b
1	KSP106	Ju|'hoansi	A3b1c	A-V306	A1b1b2a1b
1	KSP146	!Xun	A3b1b	A-V37	A1b1b2a2
1	KSP150	!Xun	A3b1b	A-V37	A1b1b2a2
4	PygmyBaka1	Baka	B2b1a1b	B-M8035	B2b1a1c2a1~
4	PygmyBaka3	Baka	B2b1a1b	B-M8035	B2b1a1c2a2~
4	Hadza2	Hadza	B2b1a2	B-M7583	B2b1a2a1~
4	Hadza5	Hadza	B2b1a2	B-M7583	B2b1a2a2~
4	Sandawe5	Sandawe	B2b1a2	B-M7583	B2b1a2b1~
4	Hadza3	Hadza	B2b1a2	B-M7583	B2b1a2b2a~
1	KSP111	Ju|'hoansi	B2b1a2	B-M7583	B2b1a2b2b~
1	KSP155	!Xun	B2b1a2	B-M7583	B2b1a2b2b~
4	Sandawe4	Sandawe	B2b1b*	B-M7104	B2b1b~
4	Hadza4	Hadza	B2b1b*	B-M7104	B2b1b1b~
2	NA19429	Luhya	E1b1a1a1a	E-M58	E1b1a1a1a1a3a~
4	Hadza1	Hadza	E1b1a1a1f1a1c*	E-P116	E1b1a1a1a1c1a1a3a1a1a2b~
4	PygmyBakola1	Bakola	E1b1a1a1f1a1c*	E-P116	E1b1a1a1a1c1a1a3a1a1b~
1	KSP063	Karretjie	E1b1a1a1f1a1d	E-CTS8030	E1b1a1a1a1c1a1a3a1c1
2	NA19443	Luhya	E1b1a1a1f1a1*	E-U174	E1b1a1a1a1c1a1a3a1c1a1a~
2	NA19428	Luhya	E1b1a1a1f1a1d	E-CTS8030	E1b1a1a1a1c1a1a3a1d1b1a1~
4	PygmyBaka2	Baka	E1b1a1a1f1a1d	E-CTS8030	E1b1a1a1a1c1a1a3a1g~
2	NA18504	Yoruba	E1b1a1a1f1a1*	E-U174	E1b1a1a1a1c1a1a3c1b1a1~
2	NA18507	Yoruba	E1b1a1a1f1a1*	E-U174	E1b1a1a1a1c1a1a3d3~
3	NA19834	Afr. American	E1b1a1a1f1a1*	E-U174	E1b1a1a1a1c1a1a3d6a
1	KSP067	Karretjie	E1b1a1a1f1a1*	E-U174	E1b1a1a1a1c1a1a3e1~
2	NA18501	Yoruba	E1b1a1a1g1*	E-U209	E1b1a1a1a2a1a1b~
2	NA18498	Yoruba	E1b1a1a1g1*	E-U209	E1b1a1a1a2a1a2a1~
1	KSP092	|Gui and ||Gana	E1b1a1a1g1a1	E-U181	E1b1a1a1a2a1a3b1a1
1	KSP096	|Gui and ||Gana	E1b1a1a1g1a1	E-U181	E1b1a1a1a2a1a3b1a1
2	NA18871	Yoruba	E1b1a1a1g1a*	E-U290	E1b1a1a1a2a1a3b1a10a~
3	NA19703	Afr. American	E1b1a1a1g1a1	E-U181	E1b1a1a1a2a1a3b1a1b~
1	KSP069	Karretjie	E1b1a1a1g1a2	E-Z1725	E1b1a1a1a2a1a3b1a2a2
2	NA19397	Luhya	E1b1a1a1g1a2	E-Z1725	E1b1a1a1a2a1a3b1a2a2a1a~
3	NA19700	Afr. American	E1b1a1a1g1a*	E-U290	E1b1a1a1a2a1a3b1a5~
4	Sandawe1	Sandawe	E1b1a1a1g1*	E-U209	E1b1a1a1a2a1a3b1d1c1a1b~
4	PygmyBedzan1	Bedzan	E1b1a1a1g1*	E-U209	E1b1a1a1a2a1a3b1d1d~
3	NA21732	Maasai	E1b1b1a3c	E-AM00003	E1b1b1a1b2a3
3	NA21737	Maasai	E1b1b1a3c	E-AM00003	E1b1b1a1b2a3
1	KSP137	Nama	E1b1b1d*	E-M293	E1b1b1b2b2a1
1	KSP152	!Xun	E1b1b1d*	E-M293	E1b1b1b2b2a1a1c~
4	Sandawe2	Sandawe	E1b1b1d*	E-M293	E1b1b1b2b2a1a1c~
4	Sandawe3	Sandawe	E1b1b1d*	E-M293	E1b1b1b2b2a1a1d~

1, this study; 2, Auton et al. (2015); 3, Drmanac et al. (2010); 4, Lachance et al. (2012).
*designates a Y chromosome paragroup

Y-chromosome from the Nama individual belonged to the basal E1b1b1b2b2a1 clade, whereas the !Xun individual belonged to a subclade further derived for markers CTS2104, CTS2297, CTS2553, and Y17343, which were also found in two Sandawe individuals (fig. 1 and supplementary table S1, Supplementary Material online).

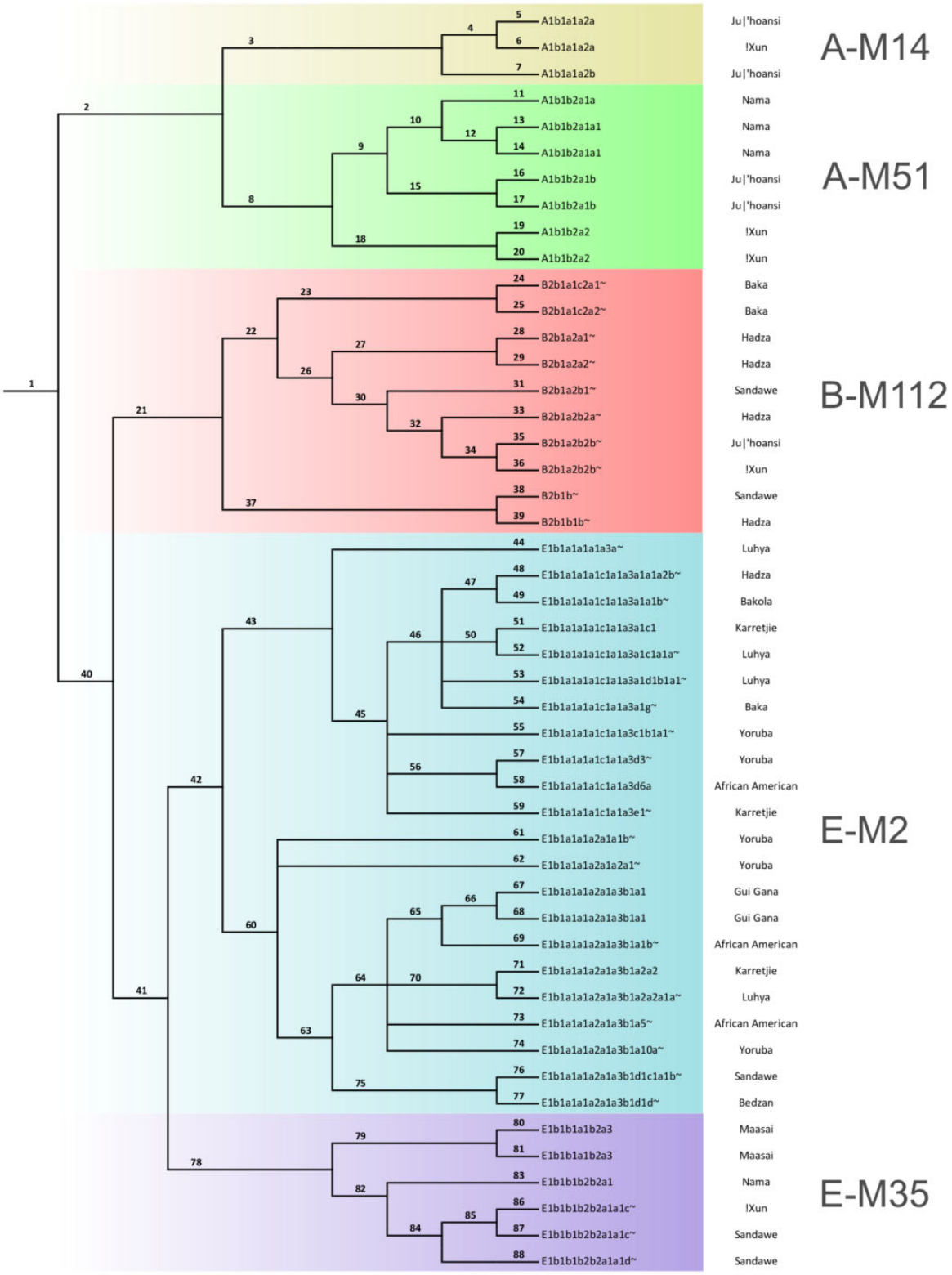

Fig. 1.—Y-chromosome phylogeny of the 48 individuals in the data set. Populations and haplogroups are shown on the right of the tree. Refer to supplementary table S1, Supplementary Material online, for additional information for each branch (numbered) in the phylogeny, including numbers of variants per branch and the lists of variants defining each branch. Branch lengths in the figure are representative of topology, and do not reflect TMRCA estimates nor the number of variants per branch.

The two haplogroup B2b (B-M112) chromosomes, found within one Ju|'hoansi and one !Xun, fell within a subclade of haplogroup B2b1a2b2∼ (B-M7592), defined by marker M7591. The closest related Y chromosome belonged to a sister clade within haplogroup B2b1a2b2∼ (B2b1a2b2a∼) and was from a Hadza individual (fig. 1 and supplementary table S1, Supplementary Material online). Although these Khoe-San B2b chromosomes are known to be derived at marker P8 (unpublished data), this marker has since been removed from recent versions of the ISOGG Y-chromosome phylogeny. The finding that marker M7591 is also derived and appears to be phylogenetically equivalent to markers P70 and P8, allows for the reintroduction of the lineage to the phylogeny.

The E1b1a1 (E-M2) haplogroup occurs in individuals from many populations (fig. 1). Khoe-San Y chromosomes within haplogroup E1b1a1 are likely to have been introduced by surrounding Bantu-speaker populations. The rapid expansion of E1b1a1 across sub-Saharan Africa (de Filippo et al. 2011) makes it difficult to pinpoint its exact sources into the Khoe-San without taking into account historical data.

Branch Length Heterogeneity

Several earlier studies (Scozzari et al. 2014; Hallast et al. 2015; Barbieri et al. 2016) found evidence of branch length heterogeneity among Y-chromosome haplogroups, and provided possible reasons for its occurrence. We also noted significant differences in branch length heterogeneity among the major African haplogroups (supplementary tables S2 and S3, Supplementary Material online). A reduced mean branch length for haplogroup A, noted previously by Scozzari et al. (2014), was again apparent from our data. Although most major haplogroups differed significantly (with the exception of the E1b1a subclades), we found that haplogroup B did not appear to have as reduced a mean branch length, relative to haplogroup E, as found previously (Hallast et al. 2015; Barbieri et al. 2016). Within haplogroup E, E1b1b1 was found to have the highest mean branch length; though this may have been due to a lower sample size compared with haplogroup E1b1a.

Placement of Branches

Markers have been added to the Y-chromosome phylogeny at a very rapid pace in the last few years. As a result, the positions of many branches have not been finalized. Our results indicate the need to reposition a few of these tentatively placed branches, especially within haplogroup A. For instance, all three A1b1a1 chromosomes were also derived for several markers tentatively assigned to haplogroups A1b1a1a2b∼ and A1b1a1a2a1a∼ (∼ denotes unconfirmed placement on ISOGG phylogeny). Further, although the ISOGG phylogeny places Y chromosomes derived at marker P262 into a subclade of A1b1a1a2a, our results reveal that this is incorrect. Marker P262-derived Y chromosomes are better placed in a sister clade to haplogroup A1b1a1a2a. This branch is supported by over 60 additional variants that segregate with P262, and we have renamed it A1b1a1a2b (fig. 1 and supplementary table S1, Supplementary Material online).

Again within haplogroup A, all A1b1b2a chromosomes were derived at markers tentatively assigned to haplogroups A1b1b2b∼ and A1b1b2b2∼. The placement of these markers is more likely to be basal to both A1b1b2a and A1b1b2b.

Discussion

In this study, we present the Y chromosomes of 19 Khoe-San individuals, following whole-genome sequencing (Schlebusch et al. 2020). We used AMY-tree version 2.0 (Van Geystelen et al. 2013) to identify the major haplogroups in the data set, based on their "Updated tree version 2.0" and known haplogroup-defining variants. Notably, some of the variants used to define haplogroups in this earlier phylogeny were not present in more recent versions of the ISOGG Y-chromosome phylogeny, but were still being used to explore variation in African populations (Barbieri et al. 2016). Once we identified the placement of our samples on the ISOGG Y-chromosome phylogeny, we were able to reposition these haplogroup-defining variants onto the ISOGG Y-chromosome phylogeny; specifically, in the case of haplogroup A1b1b2a chromosomes. As well, hundreds of new variants were discovered; most notably those found to populate the relatively sparse branches of haplogroups A1b1 and B2b. Their discovery in this study and in previous studies (Barbieri et al. 2016) should be taken into account and used to build out the branches of the oft-neglected haplogroups A and B.

Even in this minimal data set of 19 individuals, the haplogroup distribution reflected previous findings in Khoe-San populations (Underhill et al. 2000; Naidoo et al. 2010; Batini et al. 2011; Barbieri et al. 2016). The associations of the Khoe-San with other populations in the same major haplogroups also gave some indication of historical interactions and shared ancestry. While haplogroups A1b1a1, A1b1b2a, B2b, and E1b1b1 have long been associated with Khoe-San populations (Underhill et al. 2000), at this point, it appears haplogroup A1b1b2a may be the only surviving haplogroup in the Khoe-San to have originated autochthonously. The presence of haplogroup A1b1a1 in the Khoe-San in this study and others (Batini et al. 2011) has been characterized primarily by the more terminal lineages of the haplogroup and lineages ancestral to these have been found in central Africa (Batini et al. 2011).

The majority of haplogroup B2b lineages found so far in Khoe-San populations have fallen into the subclades B-P6 and B-P70 (Batini et al. 2011; Barbieri et al. 2016), whereas other B2b lineages have been found in central and eastern Africa. Previously, Batini et al. (2011) linked the presence of B-P70 in the Khoe-San to possible gene flow from central African Rainforest hunter-gatherer populations. This was based on

Table 2
TMRCA Estimates of Y-Chromosome Haplogroups Found in the Study

	TMRCA Estimate (ka)				Barbieri et al. (2016)
Mutation rate (mutations/bp/year)	**0.74×10^{-9}**		**0.87×10^{-9}**		**0.82×10^{-9}[a]**
Haplogroup	**Median**	**95% HPD**	**Median**	**95% HPD**	**Median**
A1b	169,018	133,711–216,880	146,952	116,496–185,147	193,000
A1b1a1a2	12,507	7,829–18,812	10,745	7,119–15,375	33,000
A1b1b2a	47,966	33,936–66,832	40,981	29,632–54,679	64,000
A1b1b2a1	35,572	23,600–52,078	30,342	20,532–41,834	
A1b1b2a1a	6,933	4,685–9,705	5,917	3,998–8,207	
A1b1b2a1b	8,267	4,620–13,098	7,045	4,018–10,887	
A1b1b2a2	1,669	763–2,996	1,410	655–2,415	
B2b1	50,073	40,106–61,847	42,908	34,338–52,537	62,000
B2b1a	46,948	37,249–58,643	40,260	32,103–49,891	
B2b1a1c2a	143	2–524	120	2–368	
B2b1a2	34,798	26,491–44,885	29,731	22,676–37,405	
B2b1a2a	734	245–1,502	622	217–1,240	
B2b1a2b	28,981	20,801–38,357	24,766	18,283–31,989	
B2b1a2b2	25,557	17,459–34,967	21,819	15,397–28,878	
B2b1b	44,411	33,525–57,120	38,086	28,982–48,180	
E1b1a1a1a1c1a1a3	14,332	11,338–18,158	12,253	10,133–14,921	22,000
E1b1a1a1a2a1a	9,358	7,418–11,582	8,078	6,512–9,769	15,000
E1b1b	30,063	19,813–42,157	25,468	17,560–35,183	21,000
E1b1b1b2b2a1	7,376	5,088–10,121	6,268	4,390–8,472	

[a]Mutation rate from Poznik et al. (2013).

the presence and diversity of haplogroup B-P7 (the ancestral lineage to B-P70) in the Rainforest hunter-gatherer populations and low levels of it elsewhere. Our findings, however, place the Khoe-San-specific lineage within a clade (B2b1a2b2∼) together with a Hadza lineage. The TMRCA of haplogroup B2b1a2b2∼ has been estimated to ∼21–26 ka (table 2). Given this relatively deep age, we cannot rule out the possibility that B2b1a2b2∼ was also present in central African Rainforest hunter-gatherer populations, and entered the Khoe-San population as postulated by Batini et al. (2011). However, our findings together with multiple lines of evidence pointing to gene flow from eastern Africa into southern Africa (Henn et al. 2008; Schlebusch et al. 2012, 2017; Breton et al. 2014; Macholdt et al. 2014; Pickrell et al. 2014; Skoglund et al. 2017; Bajic et al. 2018; Schlebusch and Jakobsson 2018) indicate that B2b1a2b2∼ in the Khoe-San may have come from eastern Africa. Notably, this B2b1a2b2∼ subclade has mainly been found in San populations such as the Ju|'hoansi and !Xun, and not in Khoekhoe herder populations such as the Nama (Schlebusch 2010). This is in contrast to the E1b1b1b2b2a1 haplogroup, which is found quite commonly in Nama individuals. As the E1b1b1b2b2a1 haplogroup is a marker of the movement of pastoralism from eastern Africa to southern Africa (Henn et al. 2008), this would indicate that the arrival of B2b1a2b2∼ in southern Africa was separate from the arrival of pastoralism. Evidence of a gradient of relatedness among eastern African and southern African hunter-gatherers was demonstrated by Skoglund et al. (2017); and while most of the eastern African genomic component present in southern African hunter-gatherers was attributed to the arrival of pastoralism, it appeared that the ancient southern African individuals who did not show the strong signal of eastern African admixture still shared more alleles with eastern Africans than with western Africans; possibly indicating some level of isolation-by-distance. Evidence of gene flow between the Hadza and the Ju|'hoansi, a population shown to have been affected minimally by the recent gene flow from eastern Africa (Schlebusch et al. 2017; Skoglund et al. 2017), has also been noted (Schlebusch et al. 2012). These factors may indicate that the presence of haplogroup B2b1a2b2∼ in southern Africa may predate the arrival of pastoralism.

The subclades of haplogroup A1b1b2a showed independent segregation in three Khoe-San populations. Several more individuals from these populations need to be screened to elucidate the distribution of these haplogroups in Khoe-San populations.

Barbieri et al. (2016) obtained estimates of the TMRCA for the Y-chromosome phylogeny and the major African haplogroups with 547 individuals, using counts of mutations and BEAST analysis. With fewer individuals (48) and two mutation rates, 0.74×10^{-9} (Karmin et al. 2015) and 0.87×10^{-9} (Helgason et al. 2015), our TMRCA estimates for the A1b root (A2-T in Barbieri et al. [2016]) were lower (table 2); though still within the HPD intervals. Our TMRCA estimates of the major African haplogroups were also usually lower than the (Barbieri

et al. 2016) estimates (likely due to the much lower sample size and lower haplogroup diversity), with the exception of haplogroup E1b1b1. This was reflected by haplogroup E1b1b1 also displaying the longest branch lengths in our data set (supplementary tables S2 and S3, Supplementary Material online). Although we corroborated the findings of other studies (Scozzari et al. 2014; Hallast et al. 2015; Barbieri et al. 2016) which also observed branch length heterogeneity, the patterns we observed differed slightly, with no clear sign of long E1b1a branches (Barbieri et al. 2016). This may have been due to differences in our sample sizes, or possibly due to the higher resolution of deep sequencing allowing us to uncover more variants from a larger proportion of the Y chromosome (8.8 Mb vs. ~965 kb in Barbieri et al. 2016).

The discovery of new markers on the Y chromosome and the expansion of the phylogeny have gathered pace since DNA sequencing using next-generation methods has become more commonly used to analyze this portion of the genome. The current phylogeny, however, is one with several new branches, together with some uncertainty regarding their precise placement. The screening of more samples for the already discovered markers would help in finalizing the placement of branches with unconfirmed placement.

Materials and Methods

Samples and Sequencing Pipeline

The present study is a subset of a larger study (Schlebusch et al. 2020) that sequenced high-coverage full genomes (average depth: 56×) from 25 Khoe-San individuals. The individuals were selected from a set of samples that were previously genotyped on an Illumina 2.5 M chip (Schlebusch et al. 2012). We selected five individuals from five different Khoe-San populations to represent southern, central, and northern Khoe-San groups for the full genome study, 19 of these individuals were male (3 Karretjie People; 4 Nama; 5 Ju|'hoansi; 2 |Gui and ||Gana; and 5 !Xun). Selection criteria for the individuals for the full genome study, included low amounts of admixture with Bantu-speakers or Europeans, based on autosomal data. Although males were preferentially selected, we did not consider information on Y chromosomes and mitochondrial DNA before selecting the individuals. Thus, in terms of Y chromosomes, these individuals constitute random draws from the populations.

DNA samples from individuals were collected with the subjects' informed consent, and the project was approved by the Human Research Ethics Committee (Medical) at the University of the Witwatersrand, Johannesburg (Protocol Number: M1604104 and M180654), the Working Group of Indigenous Minorities in Southern Africa (WIMSA), and the South African San Council (SASC). The study was also approved by the Swedish ethical review authority, Reference Number: Dnr 2019-05174.

DNA libraries of the Khoe-San samples were prepared with TrueSeq DNA Sample preparation kit v2 (Cat No. FC-121-2001/2002, Illumina Inc.). These were sequenced on an Illumina HiSeq Sequencing System (Illumina Inc., San Diego, CA) at the SciLifeLab SNP&SEQ Technology Platform in Uppsala. BAM files were generated by mapping the reads to the 1000 genomes phase 2 reference assembly (hs37d5) using BWA 0.6.2 (BWA-MEM algorithm) (Li and Durbin 2010), and further processed with GATK v.2.5.2 (McKenna et al. 2010), Picard v.1.92, and Samtools (Li et al. 2009). This involved duplicate marking in Picard, realignment around indels in GATK, calculating the MD flag with Samtools calmd and Base Quality Score Recalibration (BQSR) in GATK.

We used the UnifiedGenotyper module of GATK to call Y-chromosome variants from BAM files of the samples according to GATK Best Practices recommendations (DePristo et al. 2011; Van der Auwera et al. 2013). The default SNP genotype likelihoods calculation mode was used, we specified the ploidy argument as 1 for the haploid Y-chromosome data, and we did not consider indels. Both a variant-sites call set and the complete sequence VCF file containing invariant sites were generated for downstream analysis. The QD, DP, MQ, and FS information were used to determine hard-filtering thresholds, which were set at no less than 5%; with minimum QD = 1.16, minimum DP = 120.0, minimum MQ = 10.0, and maximum FS = 70.0. Both raw variant and nonvariant sites were filtered, based on these thresholds, using GATK's VariantFiltration and VCFtools (Danecek et al. 2011).

For comparative data, we selected males from seven additional African (or of African descent) populations from the 1000 Genomes Project (Auton et al. 2015)—5 Yoruba and 4 Luhya, the Complete Genomics diversity panel (Drmanac et al. 2010)—3 African Americans and 2 Maasai, and from the Lachance et al. (2012) data set—5 Hadza, 5 Sandawe, 3 Baka, 1 Bakola, and 1 Bedzan. Variants were filtered using in-house scripts with the following parameters: a minimum depth (DP) of 6×, a minimum genotype quality (GQ) of 50, and excluding records that were marked "VQLOW."

As is recognized by many studies, a large portion of human Y chromosome is ill-suited for NGS (Poznik et al. 2013; Wei et al. 2013; Karmin et al. 2015); and so we applied an additional regional filter for all Y-chromosome sequences, in order to restrict further analysis to regions of Y chromosome from which we could obtain reliable sequence data. The filter was defined by Karmin et al. (2015) (filter a + b+d) on the basis of analyses of Illumina HiSeq data with human reference genome GRCh37.

We merged the final variant call sets using the vcf-merge function from the VCFtools package (Danecek et al. 2011) and removed sites with >5% missingness.

Haplogroup Assignment and Branch Length Analysis

Haplogroup assignment was performed using AMY-tree v. 2.0 (Van Geystelen et al. 2013). Additionally, variants were

assigned to Y-chromosome phylogeny branches based on allele sharing and clade formation among individuals. Only variants <5% missingness (a maximum of two individuals with missing data per variant) were used in subsequent steps. Reference-derived variants were identified and were either removed or correctly placed on the phylogeny. In order to root the phylogeny, we used A00 sequences (Mendez et al. 2013; Karmin et al. 2015) to confirm the status of known A1b-defining variants in our data set. Variants defining the A1b1 and BT branches were differentiated from each other by checking against the A00 outgroup sequences (see supplementary tables S1 and S4, Supplementary Material online). Branch-defining variants were then matched against the ISOGG 2019-2020 Y-chromosome phylogeny (http://www.isogg.org/tree/index.html) to assign the most recent haplogroup names to the branches; including the internal branches, to confirm phylogenetic congruency. Scripts are available on request to the authors.

We also assessed whether branch length differed among the major haplogroups, by counting variants from the A1b root till the ends of the terminal branches. Haplogroups were compared using a Mann–Whitney *U* test.

Phylogenetic Analysis

We reconstructed a phylogenetic tree and dated the nodes using BEAST V1.8.2 (Drummond et al. 2012). The Y-chromosome variant-sites VCF file was converted to FASTA format using a script (vcf-tab-to-fasta; https://code.google.com/archive/p/vcf-tab-to-fasta/, last accessed May 28, 2020), before importing to BEAUti, a graphical user interface application included in BEAST package, to generate the BEAST input XML file. We chose the best-fitting substitution model by conducting test runs in jModelTest V2.7.1 (Darriba et al. 2012). The tree model was set to Coalescent: constant size with a piecewise-linear skyline model. We applied general time reversible substitution model for Y-chromosome data, using a log-normal relaxed clock with a mutation rate 0.74×10^{-9} mutations/bp/year (Karmin et al. 2015) and 0.87×10^{-9} mutations/bp/year (Helgason et al. 2015).

To reduce the computational load, we used only variant sites for the Y-chromosome BEAST analysis. However, we incorporated the information of invariant sites by specifying the nucleotide composition in the BEAST input XML file (Karmin et al. 2015). To determine the nucleotide composition of invariant sites, we compared the variant-sites VCF with the all-sites VCF and counted the number of A, T, C, and G nucleotides for invariable sites. For each mutation rate, we performed two independent runs of 100 million MCMC iterations with a sampling in every 1,000 steps. The initial 10% of each run was discarded as burn-in. The output was inspected in Tracer v1.6, confirming that all EES values were >200 and the two runs were combined with LogCombiner (Rambaut et al. 2018).

We annotated maximum clade credibility trees in TreeAnnotator (Drummond et al. 2012) and extracted the mean, median, and 95% HPD intervals of the node heights for dating.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We are grateful to all subjects who participated in this research. We thank the Working Group of Indigenous Minorities in Southern Africa (WIMSA), the South African San Council and Michael de Jongh (University of South Africa, UNISA) for their support and for assisting and facilitating during fieldwork. We thank Johanna Lagensjö and the Uppsala SNP&Seq Platform for use of laboratory space and reagents. Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. The facility is part of the National Genomics Infrastructure supported by the Swedish Research Council for Infrastructures and Science for Life Laboratory, Sweden. The SNP&SEQ Technology Platform is also supported by the Knut and Alice Wallenberg Foundation. The computations were performed at the Swedish National Infrastructure for Computing (SNIC-UPPMAX). We thank Joseph Lachance and Sarah Tishkoff for sharing the data published of Lachance et al. (2012) and Monika Karmin for providing the NRY A00 sequences published in Mendez et al. (2013) and Karmin et al. (2015). We thank Carolina Bernhardsson for help with the data upload. This work was supported by the Swedish Research Council (No. 621-2014-5211 to C.M.S. and No. 642-2013-8019 to M.J.), the European Research Council (No. 759933 to C.M.S), the Lars Hierta Foundation (C.M.S.), the Nilsson-Ehle Endowments (C.M.S.), and the Knut and Alice Wallenberg Foundation (M.J.).

Literature Cited

Auton A, et al. 2015. A global reference for human genetic variation. Nature 526(7571):68–74.

Bajic V, et al. 2018. Genetic structure and sex-biased gene flow in the history of southern African populations. Am J Phys Anthropol. 167(3):656–671.

Barbieri C, et al. 2016. Refining the Y chromosome phylogeny with southern African sequences. Hum Genet. 135(5):541–553.

Batini C, et al. 2011. Signatures of the preagricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. Mol Biol Evol. 28(9):2603–2613.

Breton G, et al. 2014. Lactase persistence alleles reveal partial East African ancestry of southern African Khoe pastoralists. Curr Biol. 24(8):852–858.

Cruciani F, 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(6):814–818.

Danecek P, et al.. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.
Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 9(8):772–772.
de Filippo C, et al.. 2011. Y-Chromosomal Variation in Sub-Saharan Africa: Insights Into the History of Niger-Congo Groups. Mol Biol Evol. 28(3):1255–1269.
DePristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491–498.
Drmanac R, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327(5961):78–81.
Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 29(8):1969–1973.
Francalacci P, et al. 2013. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 341(6145):565–569.
Hallast P, et al. 2015. The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. Mol Biol Evol. 32(3):661–673.
Hammer MF, et al. 2001. Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol. 18(7):1189–1203.
Helgason A, et al. 2015. The Y-chromosome point mutation rate in humans. Nat Genet. 47(5):453–457.
Henn BM, et al. 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. Proc Natl Acad Sci U S A. 105(31):10693–10698.
Karafet TM, et al. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res. 18(5):830–838.
Karmin M, et al. 2015. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res. 25(4):459–466.
Lachance J, et al. 2012. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell 150(3):457–469.
Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26(5):589–595.
Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078–2079.
Macholdt E, et al. 2014. Tracing pastoralist migrations to southern Africa with lactase persistence alleles. Curr Biol. 24(8):875–879.
McKenna A, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297–1303.
Mendez FL, et al. 2013. An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. Am J Hum Genet. 92(3):454–459.
Naidoo T, et al. 2010. Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. Invest Genet. 1(1):6.
Pickrell JK, et al. 2014. Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci U S A. 111(7):2632–2637.
Poznik GD, et al. 2013. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science 341(6145):562–565.
Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. Syst Biol. doi:10.1093/sysbio/syy032.
Schlebusch CM, et al. Forthcoming 2020. Khoe-San genomes reveal unique variation and confirm deepest population divergence in Homo sapiens. Mol Biol Evol.
Schlebusch CM. 2010. Genetic variation in Khoisan-speaking populations from southern Africa [PhD thesis]. [Johannesburg]: Division of Human Genetics, University of the Witwatersrand.
Schlebusch CM, et al. 2012. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science 338(6105):374–379.
Schlebusch CM, et al. 2017. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science 358(6363):652–655.
Schlebusch CM, Jakobsson M. 2018. Tales of human migration, admixture, and selection in Africa. Annu Rev Genomics Hum Genet. 19(1):405–428.
Scozzari R, et al. 2012. Molecular dissection of the basal clades in the human Y chromosome phylogenetic tree. PLoS One 7(11): e49170.
Scozzari R, et al. 2014. An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. Genome Res. 24(3):535–544.
Skoglund P, et al. 2017. Reconstructing prehistoric African population structure. Cell 171(1):59–71 e21.
Underhill PA, et al. 2000. Y chromosome sequence variation and the history of human populations. Nat Genet. 26(3):358–361.
Underhill PA, Kivisild T. 2007. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu Rev Genet. 41(1):539–564.
Van der Auwera GA, et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43:11.10.11–11.10.33.
Van Geystelen A, Decorte R, Larmuseau MH. 2013. AMY-tree: an algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications. BMC Genomics 14(1):101.
van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MH. 2014. Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. Hum Mutat. 35(2):187–191.
Wei W, et al. 2013. A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res. 23(2):388–395.
Y-Chromosome Consortium. 2002. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res. 12:339–348.

Associate editor: Partha T.E. Majumder