LETTERS

Complete Khoisan and Bantu genomes from southern Africa

Stephan C. Schuster[1]*, Webb Miller[1]*, Aakrosh Ratan[1], Lynn P. Tomsho[1], Belinda Giardine[1], Lindsay R. Kasson[1], Robert S. Harris[1], Desiree C. Petersen[2], Fangqing Zhao[1], Ji Qi[1], Can Alkan[3], Jeffrey M. Kidd[3], Yazhou Sun[1], Daniela I. Drautz[1], Pascal Bouffard[4], Donna M. Muzny[5], Jeffrey G. Reid[5], Lynne V. Nazareth[5], Qingyu Wang[1], Richard Burhans[1], Cathy Riemer[1], Nicola E. Wittekindt[1], Priya Moorjani[6], Elizabeth A. Tindall[2,7], Charles G. Danko[8], Wee Siang Teo[2,7], Anne M. Buboltz[1], Zhenhai Zhang[1], Qianyi Ma[1], Arno Oosthuysen[9], Abraham W. Steenkamp[10], Hermann Oostuisen[11], Philippus Venter[12], John Gajewski[1], Yu Zhang[1], B. Franklin Pugh[1], Kateryna D. Makova[1], Anton Nekrutenko[1], Elaine R. Mardis[13], Nick Patterson[14], Tom H. Pringle[15], Francesca Chiaromonte[1], James C. Mullikin[16], Evan E. Eichler[3], Ross C. Hardison[1], Richard A. Gibbs[5], Timothy T. Harkins[4] & Vanessa M. Hayes[2,7]*

The genetic structure of the indigenous hunter-gatherer peoples of southern Africa, the oldest known lineage of modern human, is important for understanding human diversity. Studies based on mitochondrial[1] and small sets of nuclear markers[2] have shown that these hunter-gatherers, known as Khoisan, San, or Bushmen, are genetically divergent from other humans[1,3]. However, until now, fully sequenced human genomes have been limited to recently diverged populations[4–8]. Here we present the complete genome sequences of an indigenous hunter-gatherer from the Kalahari Desert and a Bantu from southern Africa, as well as protein-coding regions from an additional three hunter-gatherers from disparate regions of the Kalahari. We characterize the extent of whole-genome and exome diversity among the five men, reporting 1.3 million novel DNA differences genome-wide, including 13,146 novel amino acid variants. In terms of nucleotide substitutions, the Bushmen seem to be, on average, more different from each other than, for example, a European and an Asian. Observed genomic differences between the hunter-gatherers and others may help to pinpoint genetic adaptations to an agricultural lifestyle. Adding the described variants to current databases will facilitate inclusion of southern Africans in medical research efforts, particularly when family and medical histories can be correlated with genome-wide data.

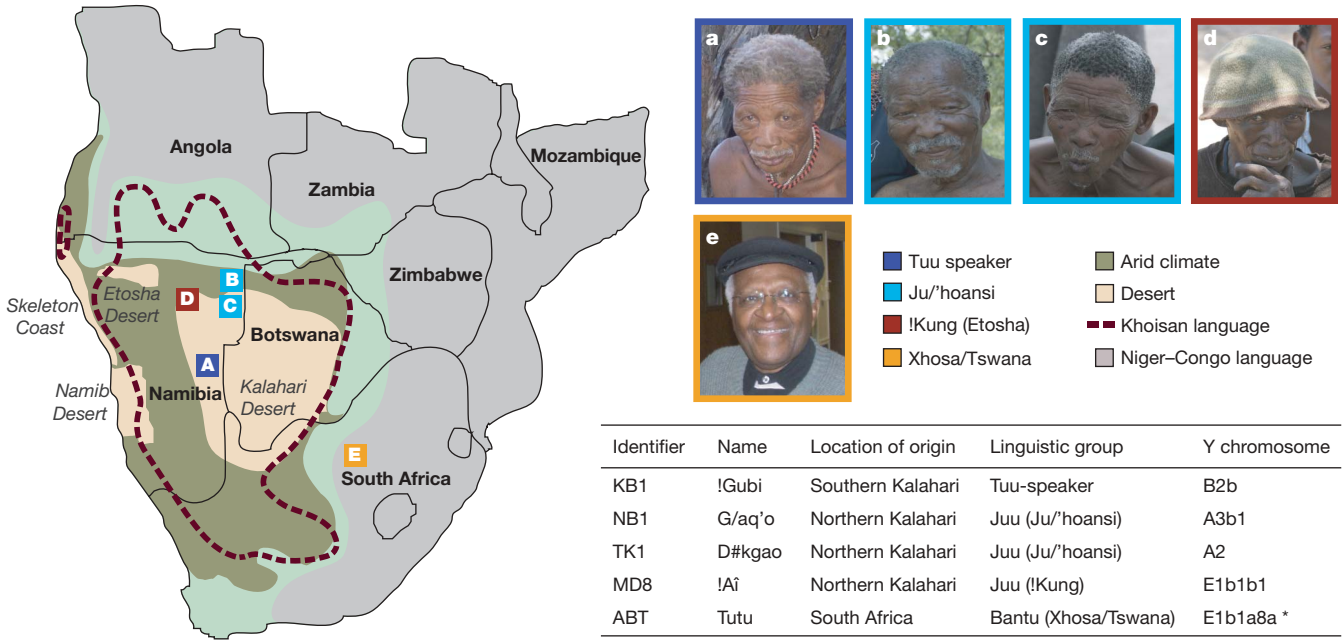
Four indigenous Namibian hunter-gatherers !Gubi, G/aq'o, D#kgao and !Aî (referred to here as KB1, NB1, TK1 and MD8, respectively), each the eldest member of his community, were chosen for genome sequencing based on their linguistic group, geographical location and Y-chromosome haplogroup representation (Fig. 1 and Supplementary Table 1). The Bantu individual is Archbishop Desmond Tutu (ABT), who represents Sotho-Tswana and Nguni speakers (from the broad Niger–Congo languages), the two largest southern African Bantu groups.

As the genomes of our study participants were expected to diverge more from the human reference genome than do the publicly accessible Yoruban, European and Asian genomes[4–8], we aimed to generate a genome sequence that would provide sufficient quality for both mapping against the human reference and *de novo* assembly. Therefore, the genome of KB1 was sequenced to 10.2-fold coverage using the Roche/454 GS FLX platform with Titanium chemistry, giving an average read length of 350 base pairs (bp). To address aspects of genome structure, additional long-insert libraries for KB1 were sequenced using the Roche/454 Titanium paired-end technology, with insert sizes up to 17 kilobases (kb) and 12.3-fold non-redundant clone coverage. The genome of NB1 was sequenced using the same platform to twofold coverage. The genome of ABT was sequenced to over 30-fold coverage using Applied Biosystems' short-read technology, SOLiD 3.0. In addition, all five of the study participants' genomes were sequenced to at least 16-fold coverage in protein-coding regions (exomes) that were enriched by Nimblegen sequence capture (2.1 M array) and subsequently sequenced on the Roche/454 Titanium platform (1.5–1.9 gigabases (Gb) of sequence per individual). Supplementary Table 2 reports the volume of data obtained, whereas Supplementary Table 3 gives exome statistics.

The sequence data were validated by a variety of techniques, including comparison of the whole-genome and exome sequences, whole-genome sequencing by another platform (Illumina, 23.2-fold for KB1 and 7.2-fold for ABT), high-density genotyping (Illumina 1 Million SNPs), comparison of read-depth information with comparative genomic hybridization data, as well as validation of selected variants using TaqMan allelic discrimination and/or Sanger sequencing. We estimate the false-positive rate of our final single nucleotide polymorphism (SNP) calls for KB1 as 0.0009, and the false-negative rate as 0.09 (see Supplementary Information for details).

We created a *de novo* assembly of the KB1 genome, using the Phusion assembler[9]. The assembled contigs total 2.79 Gb, with an N50 contig size of 5.5 kb. The total scaffold size, including estimated gaps, is 3.09 Gb, with an N50 scaffold size of 156 kb. The largest scaffold assembled spans 3.2 Mb. Frequently, the Roche GS FLX

[1]Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, 310 Wartik Lab, University Park, Pennsylvania 16802, USA. [2]Cancer Genetics Group, Children's Cancer Institute Australia for Medical Research, C25 Lowy Cancer Research Centre University of New South Wales, High Street, New South Wales 2031, Australia. [3]University of Washington, Department of Genome Sciences, and Howard Hughes Medical Institute, Foege S-413-C, Box 355065, Seattle, Washington 98195-5065, USA. [4]Roche Diagnostics Corporation, Indianapolis, Indiana 46250-0414, USA. [5]The Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. [6]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [7]University of New South Wales, Randwick, New South Wales 2031, Australia. [8]Department of Biological Statistics and Computational Biology, 101 Biotechnology Building, Cornell University, Ithaca, New York 14853, USA. [9]PO Box 1899, Tsumeb, Namibia. [10]PO Box 180, Arnos, Namibia. [11]PO Box 1077, Grootfontein, Namibia. [12]University of Limpopo, Turfloop Campus, P/Bag X1106, 0727 Sovenga, South Africa. [13]Washington University in St Louis, School of Medicine, The Genome Center, 4444 Forest Park Boulevard, St Louis, Missouri 63108, USA. [14]Broad Institute of MIT (Massachusetts Institute of Technology) and Harvard University, Cambridge Center, Cambridge, Massachusetts 02142, USA. [15]Sperling Foundation, Eugene, Oregon 97405, USA. [16]National Human Genome Research Institute, National Institutes of Health, 5625 Fishers Lane, Room 5N-01Q, MSC 9400, Rockville, Maryland 20892-9400, USA.
*These authors contributed equally to this work.



Identifier	Name	Location of origin	Linguistic group	Y chromosome
KB1	!Gubi	Southern Kalahari	Tuu-speaker	B2b
NB1	G/aq'o	Northern Kalahari	Juu (Ju/'hoansi)	A3b1
TK1	D#kgao	Northern Kalahari	Juu (Ju/'hoansi)	A2
MD8	!Aî	Northern Kalahari	Juu (!Kung)	E1b1b1
ABT	Tutu	South Africa	Bantu (Xhosa/Tswana)	E1b1a8a *

Figure 1 | Map of southern Africa. The figure shows ethnic grouping and localities of study participants, KB1, NB1, TK1, MD8 and ABT (**a**–**e**, respectively), areas of arid and desert climates and the geographic distribution of the Khoisan and Niger–Congo languages[30]. The Khoisan languages are characterized by clicks, denoting additional consonants. The ! is a palatal click; / is a dental click; and # is an alveolar click[26]. Note that the ABT Y chromosome haplogroup was determined using both genotyping and sequencing data generated by this study.

sequence data resulted in contigs and scaffolds that do not map against the human reference genome. Many of these scaffolds corresponded to gaps in the current human reference assembly, including gaps over 200,000 bp in length (see Supplementary Information).

Single-nucleotide differences from the human reference genome assembly (NCBI Build 36, also known as hg18) were identified for the five southern African genomes and compared with those from eight available personal genomes[4–8]. In what follows, the term 'SNP' means a single-nucleotide difference from the human reference assembly, not including insertions/deletions of a base, and without restrictions on allele frequency in a population. SNPs were called using the software Newbler (for Roche/454), Corona Lite (for SOLiD) and MAQ[10] (for Illumina).

Consistent with the view that southern Africans are among the most divergent human populations, we identified more SNPs in KB1, and to a lesser extent in ABT, than have been reported in other individual human genomes (Fig. 2 and Table 1), although a portion of the variation in SNP numbers may stem from differences in technology and levels of coverage. The number of SNPs that are novel (that is, not previously seen in other individuals) is far higher for KB1 and ABT than for other individual whole genomes (Table 1). KB1 and ABT each have approximately 1 million SNPs that are not shared with each other or with the published Yoruban, Asian or European complete genomes[4–8] (Fig. 2). In the 117 megabases (Mb) of sequenced exome-containing intervals, the average rate of nucleotide differences between a pair of the Bushmen was 1.2 per kilobase, compared to an average of 1.0 per kilobase differing between a European and Asian individual. The higher SNP rate in Bushmen is reflected by the offset of the red and black lines in Fig. 3b. The autosomal diversity of the study participants is mirrored by the diversity of the mitochondrial genomes. Whereas Europeans on average show approximately 20 differences from the Cambridge reference sequence (CRS)[11], our southern African participants show up to 100 mitochondrial SNPs relative to the CRS (Supplementary Tables 4 and 5 and Supplementary Figs 1 and 2). More importantly, despite all mitochondrial sequences belonging to the same haplogroup L0, up to 84 differences are observed between pairs of participants' mitochondrial genomes (Supplementary Table 4).

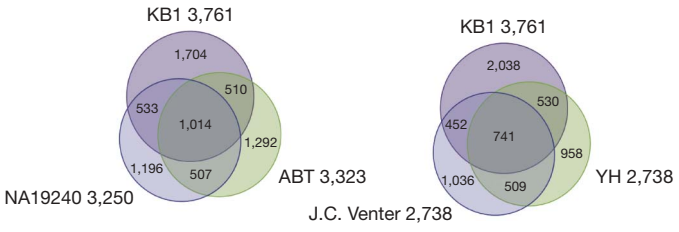


Figure 2 | Three-way relationships among SNPs. SNPs from KB1 are compared with those of the Yoruban NA19240 and ABT (left panel), and with an American of European descent (J. C. Venter) and a Chinese individual (YH) (right panel). Numbers are given in thousands. Variant positions that appear in all eight previous genomes were ignored, leading to a slightly smaller number of total SNPs (for example, 3,761,019 differences from the reference assembly for KB1, compared to 4,053,781 if they are included) and fewer SNPs in each three-way intersection. Similar relationships are found when other individuals from the geographical groups are examined.

To determine whether the novel SNPs represent ancestral alleles or arose since Bushmen separated from other populations, we examined the homologous nucleotide in the chimpanzee genome. SNPs that match the chimpanzee genome indicate that the difference is ancestral, whereas differences from chimpanzee indicate a derived allele. Of the 743,714 novel SNPs in KB1, the human reference genome matches with the chimpanzee genome for 87% of these, whereas the KB1 genome matches chimpanzee for only 6%. For the remaining 7%, the chimpanzee nucleotide could not be determined (6%) or differed from both the Bushman and the reference (1%). These fractions are essentially unchanged if we account for the estimated 3,600 false-positive SNP calls (that is, 0.0009 of 4 million), which can be assumed to appear as novel variants. Thus, very few of the novel differences in KB1's genome are ancestral nucleotides retained in the Bushmen; instead, the vast majority are changes that accumulated since the Bushmen lineage diverged from other human populations.

The large number of novel SNPs raises concerns regarding the ability of current genotyping arrays to capture effectively the true extent of genetic diversity and haplotype structure represented in southern Africa. Assessing percentage heterozygosity for 1,105,569 autosomal SNPs using current-content Illumina arrays, we were surprised to find lower heterozygosity in KB1 compared to a region-matched European control (Supplementary Data and Supplementary Fig. 3a, b), because it is well known that genetic diversity is highest in Africa. However, analysis of whole-genome sequencing data for KB1 and ABT revealed

Table 1 | Number of SNPs in the genome and the sequenced exome-containing regions

Individual	Genomic SNPs	Novel SNPs	Coding-exon SNPs
KB1	4,053,781	743,714	22,119
NB1	1,181,663	181,427	19,593
MD8	125,848	25,485	17,739
TK1	136,985	30,963	19,226
ABT SOLiD	3,624,334	412,754	17,342
ABT exome	121,383	20,294	18,994
NA18507	2,639,169	115,843	16,431
NA19240	3,586,490	216,968	17,268
J. Watson	2,060,544	98,926	11,868
J. C. Venter	3,074,574	160,370	15,079
NA12891	2,968,312	35,575	13,375
NA12892	2,972,120	36,120	13,317
Chinese	3,074,061	84,786	15,759
Korean (SJK)	3,439,097	130,566	16,637

Novel means that an SNP is not in dbSNP126, other personal genomes (except that it can be shared between our participants), PhenCode, ENCODE resequencing, the Environmental Genome Project, or the 1000 Genomes data set. NA18507 and NA19240 are Yoruban; NA12891 and NA12892 are of Caucasian origin.

high percentages of heterozygous SNPs (59% and 60%, respectively), as expected. This discrepancy underscores the inadequacy of current SNP arrays for analysing southern African populations.

The local density of SNPs identified in KB1 varies considerably across the genome (Supplementary Fig. 4), and this variation in density is also seen in other individual genomes (data not shown). Some of the hotspots are common to all individuals examined, whereas others show striking local differences among individuals, such as the statistically significant ($P < 10^{-5}$; see Supplementary Information) KB1 hotspot shown in Fig. 3a. This region corresponds to the 17q21.3 inversion[12], which contains several genes, including those encoding CRHR1 (a corticotropin-releasing hormone receptor) and MAPT (microtubule-associated protein tau). Analysis of diagnostic sequence variants as well as direct typing of a 238-bp indel[13] (Supplementary Fig. 5) confirm that KB1 is heterozygous for the 17q21.3 H2 haplotype, a surprising finding because the H2 allele is found at low frequencies in non-European populations[12]. Read depth and array-CGH indicate that the H2 allele carried by KB1 does not contain the 75-kb duplication present on all analysed European H2 alleles[14–16] (Supplementary Fig. 6a, b). The KB1 H2 haplotype may represent the ancestral sequence and structure of the H2 haplotype that was present in African populations before its increased frequency in European and Middle Eastern populations[12].

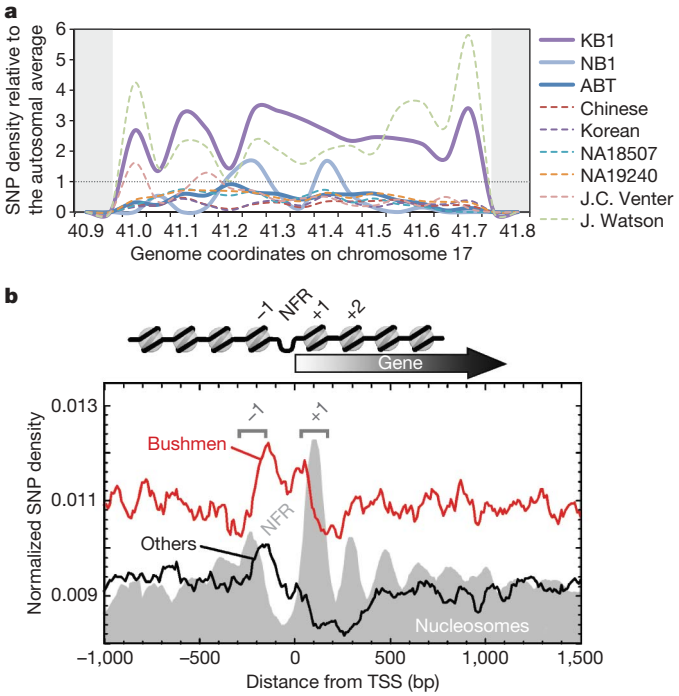


Figure 3 | Variation in SNP densities. a, An SNP hotspot for KB1 and J. Watson on chromosome 17; both individuals are heterozygous for the 17q21.3 H2 haplotype. On either side are repetitive regions where SNPs cannot be called (grey). Local SNP rates are divided by the individual's autosome-wide rate, so the expected rates are 1.0 (horizontal dotted line). KB1 has a nearly 2.5-fold enrichment of SNPs for 650,000 bases. **b**, Distribution of SNPs from Bushmen genomes (red line) and non-Bushmen genomes (black line), compared with nucleosome positions (filled grey plot), indicating the nucleosome-free region (NFR) and the −1 and +1 nucleosomes. TSS, transcription start site.

We also observed a genome-wide trend for elevated SNP levels in promoter regions (Fig. 3b). Promoter regulatory elements tend to be enriched near nucleosome borders, which are where we observed peak SNP levels, particularly in the composite Bushmen genomes. It is possible that increased SNP frequency in these genomic regions could drive phenotypic changes in humans.

We identified 27,641 distinct amino acid substitutions among our five participants, compared to the human reference sequence, many occurring in more than one individual. Of these, 10,929 appear in one or more of the previously sequenced personal genomes considered here, an additional 3,566 are found in public databases (see Supplementary Information) and the remaining 13,146 are novel and distributed among 7,720 distinct genes. The following discussion of putative phenotypes for the genotypes found in Bushmen is intended to illustrate how the presence of observed SNPs and their previous association with phenotypes can lead to testable hypotheses. These are only candidates for the suggested functions, and experimental tests must be conducted to investigate them further.

Of the 14,495 (that is, 10,929 + 3,566) previously identified amino acid SNPs, 621 were found in databases providing disease associations or other phenotypic information. Some of these are easily related to the Bushmen lifestyle, such as lack of the European-derived lactase persistence allele (a functional promoter variant in the *LCT* gene) and of the *SLC24A5* allele associated with light-coloured skin. In other instances, agreement with the human reference sequence is informative, such as the lack of the African-specific Duffy null (*DARC*) malaria-resistance allele[17]. The lack of malaria-resistance alleles in the Bushmen populations might have significant consequences on an already dwindling population of well-adapted foragers, when forced into a farming lifestyle that brings increased pathogen loads[17]. Therefore, these genetic markers may allow for the tracing of the rate of human adaptation in changing environments[18] (see Supplementary Information).

Although a number of SNPs observed in the Bushmen have been related to phenotypes in other ethnic groups in the literature and online databases, one should remain sceptical about the validity of untested associations. In the Supplementary Information, we illustrate this point with dbSNP entry rs1051339 for the *LIPA* gene, which is annotated in one public database as associated with 'Wolman's syndrome', a devastating failure in lipid metabolism (Supplementary Fig. 7).

We observed SNPs reported to be associated with enhanced physiology (Supplementary Table 6). KB1, MD8, TK1 and ABT are homozygous for an allele of *VDR* associated with higher bone mineral density; KB1 is homozygous for an allele of *UGT1A3* associated with increased metabolism of endo- and xenobiotics; KB1, NB1 and ABT are homozygous for an allele of *ACTN3* associated with increased sprint and power performance; KB1 is heterozygous for an allele of *CLCNKB* encoding a chloride channel that has a greater ability to reabsorb chloride ions from the renal glomerulus—a property that would probably be advantageous in the desert. Other interesting SNPs include one that retains the function of the *CYP2G* gene (Supplementary Fig. 8a, b), and two at positions in the taste receptor gene *TAS2R38* conferring the ability to taste a bitter compound (phenylthiocarbamide), which may reflect a need in hunter-gatherers to avoid toxic plants (see Supplementary Information for detailed discussion).

The 13,146 novel amino acid SNPs reported here will be a rich resource for future work, providing many new candidate functional sites that have not been included in whole-genome association studies

so far. Approximately 25% of these SNPs are predicted to have functional implications by a suite of computational methods (see Supplementary Information). The Gene Ontology categories that are prominently represented in the 6,623 genes with one or more novel Bushmen SNP (that is, excluding from the 7,720 genes with novel SNPs those unique to ABT) include many functions that are known to evolve quickly in humans, such as immune response, reproduction and sensory perception (Supplementary Table 7). See the Supplementary Information for detailed descriptions of computational analyses of genes related to lipid metabolism and sensory perception.

As all of our study participants are of old age (~80 years) and seemingly in good health, the novel coding variants described in this study can be correlated to health status and phenotypes over the entire human lifespan. The Bushmen participants have reached their advanced age despite living under harsh conditions due to periodical famine and untreated illnesses. As some of the Bushmen coding alleles have been associated in the published literature with disease, our results may help to reassess those earlier reports, as well as help to identify potential population-specific pharmacogenetic incompatibilities of certain drugs that are globally prescribed.

Segmental duplications were detected in 17,601 distinct autosomal genes in the KB1 genome and copy numbers estimated following procedures described earlier[19] (Supplementary Fig. 6a, b). Copy numbers estimated from read depth are more reliable for longer segments, so we specifically targeted regions larger than 20 kb. In total, we detected 886 intervals (each >20 kb) of autosomal segmental duplication (93.5 Mb), which includes 100 intervals (3.9 Mb) that are not predicted to be duplicated in sample NA18507 (a HapMap sample from Yoruba, Nigeria)[19]. Using array-CGH, 58 of these intervals (2.6 Mb) had increased copy numbers in KB1 relative to NA18507, the only other published African genome. The set of validated duplications includes a 140-kb interval on chromosome 10 spanning the *CYP2E1* gene, which encodes a cytochrome P450 protein that is induced by ethanol and metabolizes many toxicological substrates[20] (Supplementary Fig. 6a).

Next, we specifically estimated copy numbers for all autosomal RefSeq genes and designed a custom oligonucleotide array targeting genes where KB1 and NA18507 are predicted to differ by at least one copy. This validated 193 genes as differing in copy number between KB1 and NA18507 (53 where NA18507 has more copies and 140 where KB1 has more copies; Supplementary Table 8). For 26 of these genes, KB1 is estimated to have at least two copies more than in NA18507, Han Chinese YH, and European-descent J. Watson. This gene set includes salivary amylase (*AMY1A*, KB1 copy number estimate = 15; this may be consistent with a forager lifestyle[21]), the alpha defensins (*DEFA1*, KB1 copy number estimate = 12.5) and γ-glutamyltransferase 1 (*GGT1*, KB1 copy number estimate = 13.2).

Sequencing and extensive genotyping revealed genetic relationships among our participants and other human groups. Placement of complete mitochondrial genomes (Supplementary Table 9), including additional Tuu (KB2) and Juu (NB8) females on the maternal tree of ref. 1 (Supplementary Fig. 1a–c) positioned our participants within the clade L0 basal branch. Surprisingly, ABT was placed in clade L0d, a Bushmen-specific mitochondrial lineage. We identified 75 (of 1,220) Bushmen-informative SNPs on the Y chromosome (Supplementary Fig. 9). In contrast to the other Bushmen, MD8 showed a Bantu Y-chromosome lineage consistent with ABT. Clade A (Supplementary Table 10), B (Supplementary Table 11) and E (Supplementary Table 12) Y-marker analysis allowed for haplogroup validation and ABT's E1b1a8a classification (http://ycc.biosci.arizona.edu/)[22].

We performed principal component analysis (PCA) using the EIGENSTRAT software[23] on 174,272 autosome-wide SNPs common across the data sets (generated using 1M or 610K Illumina, or Affymetrix SNP6.0 arrays). Data on 10 Bushmen and 20 Xhosa[24] were projected with 20 Yoruba and 20 Europeans from available (HapMap and Coriell) data, and 5 Bushmen (SAN) from the Human Genome Diversity Panel (HGDP) data. Population-wide PCA defines the Bushmen as distinct from the Niger–Congo populations as from Europeans (Fig. 4a). Within-Africa analysis separates Bushmen from the divergent western and southern African populations (Fig. 4b), whereas ABT clearly falls within the Southern Bantu cluster. Variable relatedness of the Xhosa to Yoruba may suggest past admixture and/or historical diversity within this broadly defined population[24]. Within the Bushmen group, we predict that the Ju/'hoansi and HGDP San are essentially the same population. Divergence of KB1 and MD8 may be explained by recent Bantu admixture (assumed for MD8) or by unique sub-populations with a small percentage of ancient Bantu admixture. Although limited by sample size, a four population test[17] suggests weak and/or inconclusive admixture in KB1 and our Ju/'hoansi participants. A different test (see Supplementary Table 14) shows gene-flow between ancestors of KB1 and ABT, confirming the mitochondrial results, but without determining the direction of flow. In contrast to KB1, NB1 and TK1, gene flow between Bushmen and southern African Bantu could be confirmed through ABT's L0 type mitochondria and the Bantu-specific Y-chromosomal markers in MD8. Whether the migrations underlying these instances followed a general pattern of either patri- or matrilocality[25] will have to await a detailed population-structure analysis based on novel-content arrays that include the 1.3 million new genetic markers from this study.

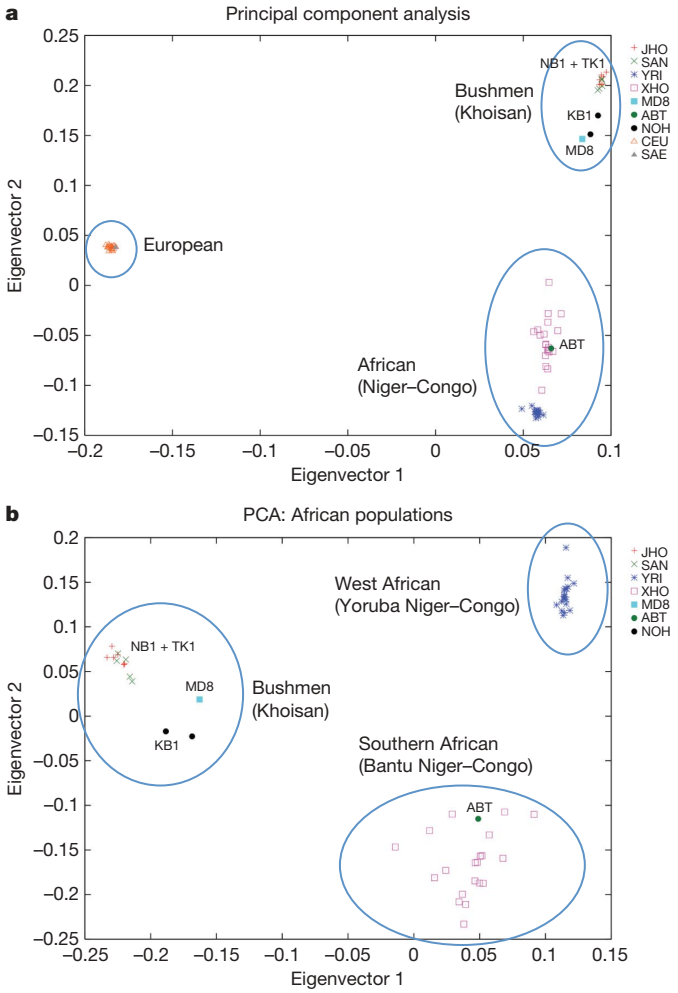


Figure 4 | Three-way population structure based on 174,272 autosomal SNPs using PCA. **a**, **b**, The PCA of Europeans, Africans (Niger–Congo) and Bushmen (**a**) and African populations only (**b**) distinguishes the Bushmen from Yorubans and Bantus. The fraction of the variance explained in **a** is 0.09 for eigenvector 1 and 0.04 for eigenvector 2, whereas in PCA **b** it is 0.06 and 0.02, respectively, with a Tracy–Widom *P* value $<10^{-12}$. ABT, sequenced Bantu; CEU, European HapMap; JHO, Juu speakers (including NB1 and TK1); MD8, sequenced !Kung; NOH, Tuu speakers (including KB1); SAE, South African European; SAN, HGDP San; XHO, South African Xhosa; YRI, Yoruba HapMap.

As the Bushmen hunter-gatherers have never adopted agricultural practices throughout their cultural history[26], the sequence variants found in their genomes may reflect an ancient adaptation to a foraging lifestyle. In the case of the Kalahari Bushmen, adaptation to life in arid climates must have occurred as well, as several phenotypic traits have been noted that are absent in other human groups, such as the ability to store water and lipid metabolites in body tissues[26]. These physiological and genetic differences may guide future studies into the much debated question of whether population replacement, rather than cultural exchange, has driven the expansion of agriculture in the southern regions of Africa[27], as was observed for late Stone Age populations in Europe[28,29].

METHODS SUMMARY

Using guidelines approved by the Institutional Review Board of Penn State University, USA, (IRB 28460 and IRB 28890), the University of Limpopo Ethics Committee, South Africa (Limpopo Provincial Government #011/2008), and the Human Research Ethics Committee of the University of New South Wales, Australia (HREC 08089 and HREC 08244), all participants consented either in writing (ABT) or via video-recorded verbal consent (Bushmen). The collection of human DNA in Namibia was conducted under a permit by the Ministry of Health and Social Services (MoHSS) of the Namibian Government. None of the Bushmen participants had any previously known genetic conditions. ABT is a survivor of poliomyelitis, tuberculosis and prostate cancer.

Several whole-genome shotgun DNA libraries for KB1 and NB1 were prepared and sequenced using methods previously described for the Roche/454 platform. Exome sequences for the five participants and whole-genome sequence for ABT were obtained as described in the paper. Mapping of the 454 sequencing reads to the human reference sequence (NCBI Build 36) was performed using a locally produced aligner called *lastz* (http://www.bx.psu.edu/miller_lab) and in-house scripts.

Access to our data. It is challenging to provide convenient access to the large and complex data sets resulting from the sequencing and analysis of a human genome. In addition to submitting data to standard repositories, we provide all data sets in an immediately useful form through the Galaxy bioinformatics platform (http://usegalaxy.org), a web application designed to integrate data with analysis tools. In addition to downloading, data sets can be transformed in a variety of ways and compared with existing annotations (see 'Data and analysis user's guide' at http://galaxycast.org). The positions of the SNPs for each Bushman and ABT can be viewed in a customized installation of the UCSC Genome Browser (http://main.genome-browser.bx.psu.edu/), along with supporting evidence (number of reads for each allele and hyperlinks to the actual reads) and computationally predicted phenotypic consequences for SNPs in coding regions.

Received 11 August 2009; accepted 6 January 2010.

1. Gonder, M. K. *et al.* Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol. Biol. Evol.* **24,** 757–768 (2007).
2. Myers, S. *et al.* A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310,** 321–324 (2005).
3. Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324,** 1035–1044 (2009).
4. Ahn, S. M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* **19,** 1622–1629 (2009).
5. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
6. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
7. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456,** 60–65 (2008).
8. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
9. Mullikin, J. C. & Ning, Z. The phusion assembler. *Genome Res.* **13,** 81–90 (2003).
10. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18,** 1851–1858 (2008).
11. Andrews, R. M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genet.* **23,** 147 (1999).
12. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nature Genet.* **37,** 129–137 (2005).
13. Baker, M. *et al.* Association of an extended haplotype in the tau gene with progressive supranuclear palsy. *Hum. Mol. Genet.* **8,** 711–715 (1999).
14. Antonacci, F. *et al.* Characterization of six human disease-associated inversion polymorphisms. *Hum. Mol. Genet.* **18,** 2555–2566 (2009).
15. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genet.* **40,** 1166–1174 (2008).
16. Zody, M. C. *et al.* Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nature Genet.* **40,** 1076–1083 (2008).
17. Reich, D. *et al.* Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS Genet.* **5,** e1000360 (2009).
18. Coop, G. *et al.* The role of geography in human adaptation. *PLoS Genet.* **5,** e1000500 (2009).
19. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genet.* **41,** 1061–1067 (2009).
20. Kessova, I. & Cederbaum, A. I. CYP2E1: biochemistry, toxicology, regulation and function in ethanol-induced liver injury. *Curr. Mol. Med.* **3,** 509–518 (2003).
21. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nature Genet.* **39,** 1256–1260 (2007).
22. Karafet, T. M. *et al.* New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.* **18,** 830–838 (2008).
23. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
24. Patterson, N. *et al.* Genetic structure of a unique admixed population: implications for medical research. *Hum. Mol. Genet.* **19,** 411–419 (2009).
25. Oota, H. *et al.* Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nature Genet.* **29,** 20–21 (2001).
26. Le Roux, W. & White, A. *Voices of the San: Living in Southern Africa Today* (Kwela, 2004).
27. Berniell-Lee, G. *et al.* Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. *Mol. Biol. Evol.* **26,** 1581–1589 (2009).
28. Malmstrom, H. *et al.* Ancient DNA reveals lack of continuity between neolithic hunter-gatherers and contemporary Scandinavians. *Curr. Biol.* **19,** 1758–1762 (2009).
29. Bramanti, B. *et al.* Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science* **326,** 137–140 (2009).
30. Heine, B. & Nurse, D. in *African Languages: An Introduction* (Cambridge Univ. Press, 2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We particularly want to thank Archbishop Desmond Tutu, !Gubi, G/aq'o, D#kgao and !Aî, as well as their respective families and communities for their willingness to participate in this study. This work is supported by the Pennsylvania State University. The genome sequencing of the four Namibian individuals was supported by Roche Applied Sciences and the exome sequencing capture by Nimblegen (to S.C.S.). Sequencing of Archbishop Tutu's genome was supported by AppliedBiosystems (whole genome) (to R.A.G.) and Roche Applied Sciences (exome) (to S.C.S.). Whole-genome genotyping and NRY haplotyping was supported by the Cancer Institute of New South Wales (to V.M.H.). Travel was supported by Penn State University (to S.C.S.) and Hyperion Asset Management Australia (to V.M.H.). We thank K. Walters for assistance in researching geographic facts and help with Fig. 1, B. Schultz for his help with Supplementary Table 5 and R.-A. Hardie for her help with Supplementary Fig. 8b. T. Loughran assisted with medical sample collection. We thank the 1000 Genome Project for early access and use of their data. This work was also supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (J.C.M.). A.R. was supported by an NSF grant DEB-0733029 and K.D.M. was supported by a grant R01GM087472 from NIH. S.C.S. is supported by the Gordon and Betty Moore Foundation. V.M.H. is a Cancer Institute of New South Wales Fellow.

Author Contributions S.C.S. and V.M.H. managed the project. S.C.S. and V.M.H. collected and processed the blood samples during field trips in 2008 and 2009. S.C.S., V.M.H. and W.M. designed research. Sequencing data was generated by L.P.T., L.R.K., D.C.P., D.I.D., J.G., P.B., D.M.M., J.G.R., L.V.N., V.M.H. and S.C.S. Genotyping was performed by D.C.P., E.A.T., W.S.T. and V.M.H. Data were analysed by S.C.S., W.M., A.R., B.G., R.S.H., C.R., D.C.P., F.Z., Y.S., C.A., J.M.K., D.I.D., J.Q., R.B., Q.W., Q.M., Z.Z., N.E.W., A.M.B., P.M., C.G.D., R.S.H., K.D.M., A.N., E.R.M., N.P., T.H.P., Y.Z., F.C., J.C.M., R.C.H., B.F.P., E.E.E., R.A.G., T.T.H. and V.M.H.; A.O., A.W.S., H.O. and P.V. assisted with field work. S.C.S., W.M. and V.M.H. wrote the manuscript with input from the co-authors.

Author Information All sequence data have been deposited in the NCBI short read archive, with accession number SRA010356. The sequences and associated data are freely available from http://galaxy.psu.edu/bushman. SNP and indel information has been placed into the dbSNP database under handle BUSHMAN. The GEO id for the array data is GSE19048. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.C.S. (scs@bx.psu.edu) or V.M.H. (vhayes@ccia.unsw.edu.au).