Downloaded from genome.cshlp.org on May 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations

Juan Rodriguez-Flores, Khalid Fakhro, Francisco Agosto-Perez, et al.

Genome Res. published online January 4, 2016
Access the most recent version at doi:10.1101/gr.191478.115

P<P Published online January 4, 2016 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Indigenous Arabs are Descendants of the Earliest Split from Ancient Eurasian Populations

Juan L. Rodriguez-Flores[1]*, Khalid Fakhro[2,3]*, Francisco Agosto-Perez[1,4],
Monica D. Ramstetter[4], Leonardo Arbiza[4], Thomas L. Vincent[1], Amal Robay[3], Joel A. Malek[3],
Karsten Suhre[5], Lotfi Chouchane[3], Ramin Badii[6], Ajayeb Al-Nabet Al-Marri[6],
Charbel Abi Khalil[3], Mahmoud Zirie[7], Amin Jayyousi[7], Jacqueline Salit[1], Alon Keinan[4],
Andrew G. Clark[4], Ronald G. Crystal[1]** and Jason G. Mezey[1,4]**

[1]Department of Genetic Medicine
Weill Cornell Medical College
New York, New York,

[2]Sidra Medical and Research Center
Doha, Qatar,

[3]Department of Genetic Medicine
Weill Cornell Medical College – Qatar
Doha, Qatar,

[4]Department of Biological Statistics and Computational Biology
Cornell University
Ithaca, NY,

[5]Bioinformatics Core
Weill Cornell Medical College – Qatar
Doha, Qatar

and

[6]Laboratory Medicine and Pathology
[7]Department of Medicine
Hamad Medical Corporation
Doha, Qatar

* JLRF and KF contributed equally to this study
** JGM and RGC contributed equally as senior investigators for this study

Running head: Uninterrupted ancestry in Arabia

Correspondence:
Department of Genetic Medicine
Weill Cornell Medical College
1300 York Avenue, Box 164
New York, New York 10065
Phone: (646) 962-4363
Fax: (646) 962-0220
E-mail: geneticmedicine@med.cornell.edu

Abstract

An open question in the history of human migration is the identity of the earliest Eurasian populations that have left contemporary descendants. The Arabian Peninsula was the initial site of the out of Africa migrations that occurred between 125,000 - 60,000 years ago, leading to the hypothesis that the first Eurasian populations were established on the Peninsula and that contemporary indigenous Arabs are direct descendants of these ancient peoples. To assess this hypothesis, we sequenced the entire genomes of 104 unrelated natives of the Arabian Peninsula at high coverage, including 56 of indigenous Arab ancestry. The indigenous Arab genomes defined a cluster distinct from other ancestral groups and these genomes showed clear hallmarks of an ancient out of Africa bottleneck. Similar to other Middle Eastern populations, the indigenous Arabs had higher levels of Neanderthal admixture compared to Africans but had lower levels than Europeans and Asians. These levels of Neanderthal admixture are consistent with an early divergence of Arab ancestors after the out of Africa bottleneck but before the major Neanderthal admixture events in Europe and other regions of Eurasia. When compared to worldwide populations sampled in the 1000 Genomes Project, while the indigenous Arabs had a signal of admixture with Europeans, they clustered in a basal, outgroup position to all 1000 Genomes non-Africans when considering pairwise similarity across the entire genome. These results place indigenous Arabs as the most distant relatives of all other contemporary non-Africans and identify these people as direct descendants of the first Eurasian populations established by the out of Africa migrations.

Introduction

All humans can trace their ancestry back to Africa (Cann 1987), where the ancestors of anatomically modern humans first diverged from primates (Patterson 2006), and then from archaic humans (Prufer 2014). Humans began leaving Africa through a number of coastal routes, where estimates suggest these “out of Africa” migrations reached the Arabian Peninsula as early as 125,000 years ago (Armitage 2011), and as late as 60,000 years ago (Henn 2012). After entering the Arabian Peninsula, human ancestors entered South Asia and spread to Australia (Rasmussen 2011), Europe, and eventually, the Americas. The individuals in these migrations were the most direct ancestors of ancient non-African peoples and they established the contemporary non-African populations recognized today (Cavalli-Sforza and Feldman 2003).

The relationship between contemporary Arab populations and these ancient human migrations is an open question (Shriner 2014; Lazaridis 2014). Given that the Arabian Peninsula was an initial site of egress from Africa, one hypothesis is that the original out of Africa migrations established ancient populations on the Peninsula that were direct ancestors of contemporary Arab populations (Lazaridis 2014). These people would therefore be direct descendants of the earliest split in the lineages that established Eurasian and other contemporary non-African populations (Rasmussen 2011; Armitage 2011; Henn 2012; Shriner 2014; Lazaridis 2014). If this hypothesis is correct, we would expect that there are contemporary, indigenous Arabs who are the most distant relatives of other Eurasians. To assess this hypothesis we carried out deep coverage genome sequencing of 104 unrelated natives of the Arabian Peninsula who are citizens of the nation of Qatar (Supplemental Figure 1), including 56 of indigenous Bedouin ancestry who are the best representatives of autochthonous Arabs, and compared these genomes to contemporary genomes of Africa, Asia, Europe, and the Americas (The 1000 Genomes Project Consortium 2012; Lazaridis 2014).

Results

Population Structure of the Arabian Peninsula

Previous analyses of the populations of the Arabian Peninsula (Hunter-Zinck 2010; Alsmadi 2013) have found 3 distinct clusters that reflect primary ancestry: Q1 (Bedouin), Q2 (Persian-South Asian), and Q3 (African) (Omberg 2012). By assessment of medical records and ancestry-informative SNP genotyping (Supplemental Figure 2), a sample of 108 purportedly unrelated individuals was selected for sequencing, including 60 Q1 (Bedouin), 20 Q2 (Persian-South Asian), and 20 Q3 (African), as well as 8 Q0 (Subpopulation Unassigned) that could not be cleanly placed in one of these three groups (Supplemental Table I). Each of these genomes was sequenced to a median depth of 37* (minimum 30*) by Illumina technology, identifying a total of 23,784,210 SNPs (see Methods, Supplemental Table II).

To confirm that none of the 108 individuals were closely related, we used KING-robust (Manichaikul 2010) and PREST-plus (McPeek and Sun 2000) to estimate family relationships based on a set of 1,407,483 SNPs after pruning of the full set of 22,958,844 autosomal SNPs in Qatar (see Methods). Both analyses identified five pairs of related individuals greater than third-degree that were subsequently confirmed by investigative re-assessment of medical records (Supplemental Table III, Supplemental Figure 3). Three of the pairs form a trio, hence 2 individuals from the trio were removed, and one individual from each of the two remaining pairs was removed, such that the remaining 104 individuals analyzed further included 8 Q0 (Subpopulation Unassigned) and 96 Q1-Q3 Qatari: 56 Q1 (Bedouin), 20 Q2 (Persian-South Asian), and 20 Q3 (African).

An analysis of inbreeding for these remaining individuals showed the Q1 (Bedouin) to have a more positive inbreeding coefficient than most of the non-admixed 1000 Genomes (The 1000 Genomes Project Consortium 2012) populations (Supplemental Table IV, Supplemental

Figure 4), consistent with the known inbreeding of this group (Hunter-Zinck 2010; Omberg 2012), although we also found the Q1 (Bedouin) to be less inbred than many small and/or isolated populations worldwide represented in the Human Origins samples (Lazaridis 2014) (Supplemental Table V, Supplemental Table VI, Supplemental Figure 4). The Q2 (Persian-South Asian) had a positive, but slightly lower, inbreeding coefficient than the Q1 (Bedouin). In contrast, the Q3 (African) had a non-negative coefficient that reflects known admixture with African populations (Hunter-Zinck 2010; Omberg 2012).

We confirmed the primary ancestry classifications of the 104 Qataris by principal component analysis (Price 2006). We combined the 104 Qataris, the Human Origins populations (Lazaridis 2014), and 1000 Genomes populations (The 1000 Genomes Project Consortium 2012) (excluding individuals already in Human Origins), and performed principal component analysis on a set of 197,714 linkage disequilibrium pruned autosome SNPs (Figure 1A and Supplemental Figure 5A). We also confirmed these clusterings just with the 104 Qataris and 1000 Genomes samples based on the same set of autosomal SNPs (Supplemental Figure 5B). These analyses reproduced the population clustering observed previously (Hunter-Zinck 2010; Omberg 2012), with the Q1 (Bedouin) closest to Europeans, the Q2 (Persian-South Asian) between Q1 (Bedouin) and Asians, and the Q3 (African) closest to African populations. A plot of just the Middle Eastern populations on the principal components also showed clustering as expected, with the Q1 (Bedouins) clustering with previously sampled Bedouins and Arabs, Q2 (Persian-South Asians) with Iranians, and Q3 (African) outside of the Middle Eastern cluster (not shown) (Figure 1B).

Y Chromosome and Mitochondrial DNA Haplogroups

We next analyzed the Y Chromosome (ChrY) and Mitochondrial DNA (MtDNA) to assess the degree to which the Q1 (Bedouin), Q2 (Persian-South Asian), or Q3 (African) Qatari ancestry groups represent distinct subpopulations (Figure 2). The ChrY haplogroups showed al-most no overlap between the Q1 (Bedouin) Qataris and Q2 (Persian-South Asian) Qataris, where an Analysis of Molecular Variance (AMOVA) was highly significant ($p<0.018$; Supplemental Table VII). The Arab haplogroup J1 was the dominant haplogroup in the Q1 (Bedouin) Qataris, but this haplogroup was not represented at all among the Q2 (Persian-South Asian) Qataris (Figure 2A). This confirmed that these are genetically well defined subpopulations that are relatively isolated from one another (Omberg 2012). There was also a strong partitioning of the ChrY haplogroups when considering the Q3 (African) Qataris, both when considering Q1 (Bedouin) *vs* Q3 (African) (AMOVA $p<1x10^{-5}$) and Q2 (Persian-South Asian) *vs* Q3 (African) (AMOVA $p<0.028$). The Q3 (African) had largely African haplogroups, a result consistent with the known recent African admixture of this subpopulation (Omberg 2012).

The MtDNA haplogroups were less partitioned among the Qataris, although they still showed significant partitioning between each pair of subpopulations (AMOVA Q1 *vs* Q2 $p<0.035$, Q1 *vs* Q3 $p<1x10^{-5}$ Q2 *vs* Q3 $p<0.017$) and among all three considered simultaneously (AMOVA $p<1x10^{-5}$; Supplemental Table VII). The MtDNA haplogroups also included more worldwide geographic diversity overall, indicating a different male *vs* female pattern of intermarriage among these subpopulations (Sandridge 2010). Together the ChrY and MtDNA haplogroups indicate that the Q1 (Bedouin), Q2 (Persian South-Asian), and Q3 (African) ancestry groups represent genetic subpopulations that not only reflect known migration history (Hunter-Zinck 2010; Omberg 2012) but that also represent units defined by a patrilocal society with strong historical barriers to intermarriage (Esposito 2001; Cavalli-Sforza and Feldman 2003), where gene flow has been dominated by female movement (i.e., admixture occurring through females marrying into the relatively isolated subpopulations), as well as female influxes from other geographic areas.

X-linked and Autosomal Diversity

To further analyze the relative male and female contributions to the genetics of the Qatari Q1 (Bedouin), Q2 (Persian-South Asian), and Q3 (African) subpopulations, we analyzed genome-wide ratios of X-linked and autosomal (X/A) diversity and X/A diversity ratios for genome intervals >0.18 cM from genes (Supplemental Table VIII, Supplemental Figure 6). For both of these ratios, the Q1 (Bedouin) and Q2 (Persian-South Asian) were lower than for African populations but were higher than for Europeans and Asians. This points to a higher effective population size of females in the Q1 (Bedouin) and Q2 (Persian-South Asian), possibly a consequence of the out of Africa migrations, which were believed to be biased toward migration of males over females (Gottipati 2011; Arbiza 2014). The Q3 (African) Qataris had X/A diversity ratios that were higher, even when compared to African populations. This may be driven by a smaller male effective population size; a possible consequence of polygamous culture and the ancestry of the Q3 (African) subpopulation that was a result of the historical slave trade into the region from Africa (Omberg 2012).

We also analyzed the relative ratios of X-linked and autosomal (X/A) diversity in non-genic regions of the female Q1 (Bedouin), Q2 (Persian-South Asian), and Q3 (African) genomes compared to females in African populations of 1000 Genomes (Supplemental Table IX). The relative X/A ratios of both the Q1 (Bedouin) and Q2 (Persian-South Asian) to African populations were slightly higher than when comparing European to African populations (Gottipati 2011; Arbiza 2014). This could indicate a slightly less extreme set of bottleneck events encountered since the out of Africa migrations by the direct ancestors of the Q1 (Bedouin) and Q2 (Persian-South Asian) compared to the bottlenecks encountered by the direct ancestors of Europeans. The relative X/A diversity ratios of Q3 (African) to African populations were closer to one, consistent with the known African admixture of this subpopulation (Omberg 2012).

Pairwise Sequential Markov Coalescent Analysis

We next analyzed the full complement of autosomal polymorphisms for signals of ancient bottlenecks by applying the pairwise sequential Markov coalescent (PSMC) (Li and Durbin 2011) (Figure 3). This analysis showed that the Q1 (Bedouin) and Q2 (Persian-South Asian) had clear hallmarks of a bottleneck event, with effective population size hitting a trough in the range of 100,000 to 30,000 years ago with a minimum at approximately 60,000 years ago. This same pattern is observed for a European individual from the 1000 Genomes and is consistent with what has been observed in other non-African human genomes using the pairwise sequential Markov coalescent, as well as related methods (Gronau 2011; Schiffels and Durbin 2014; Fu 2014). These data, therefore, point to the ancestors of Q1 (Bedouin) and Q2 (Persian-South Asian) as having migrated out of Africa at the same time as the ancestors of other non-African populations (Henn 2012). While PSMC estimates in the more recent past tend to have larger confidence intervals (Li and Durbin 2011), the Q1 (Bedouin) do appear to have a lower population size than the Q2 (Persian South-Asian) in the region <30,000 years ago, consistent with high levels of inbreeding in the Q1 (Bedouin) (Hunter-Zinck 2010; Sandridge 2010; Mezzavilla 2015). For the Q3 (African), the median effective population size was more similar to an African individual from the 1000 Genomes in the range 100,000 to 30,000 years ago, consistent with Sub-Saharan African ancestry that is relatively recent (Omberg 2012).

Admixture Analysis

The signal of an ancient bottleneck in the Q1 (Bedouin) is not unexpected given previous analyses of genomic admixture that found <1% African ancestry in this subpopulation (Omberg 2012) and studies of worldwide population structure, which have inferred that the Q1 (Bedouin) genomes have the greatest proportion of Arab genetic ancestry, even when compared to Bedouins from outside Qatar and to Arabs in surrounding countries, including Yemen and Saudi Arabia

(Shriner 2014; Hodgson 2014). To confirm a similarly minute amount of African admixture for the Q1 (Bedouin) in our sample, we applied three methodologies: (1) An ADMIXTURE (Alexander 2009) analysis of the genome-wide ancestry proportions in the 104 Qataris, 1000 Genomes (The 1000 Genomes Project Consortium 2012), and Human Origins samples (Lazaridis 2014); (2) An ALDER (Loh 2013) analysis of the proportion and timing of African ancestry in these same populations, and (3) A SupportMix (Omberg 2012) analysis of the population assignments of local genomic segments of the 96 Q1 (Bedouin), Q2 (Persian-South Asian), or Q3 (African) Qatari genomes.

The ADMIXTURE analysis identified K=12 ancestral populations as having the lowest cross-validation error (Supplemental Figure 7A). At this level of resolution the Q1 (Bedouin) had a high average (84%) proportion of ancestry that was also present in the Human Origins Bedouin B population at a high average proportion (93%) (Supplemental Figure 7B and C), where this same ancestry was also shared with Saudis, and at lower levels among other Middle Eastern populations. This ancestry therefore appears to be the signal of an indigenous Arab ancestral population. The Bedouin A population also shared this ancestry but at a lower average proportion (45%) and appeared to be more admixed overall. The Q2 (Persian-South Asian) shared a large proportion (45% on average) of ancestry that dominates in Iranians (46% on average), consistent with a Persian ancestral population (Omberg 2012). The Q3 (African) shared the majority of ancestry with African populations as expected and were considerably admixed overall, again consistent with the known history of this subpopulation (Omberg 2012) (Supplemental Figure 7A).

The ALDER analysis determined the relative percent of African (Yoruba) ancestry in the Q1 (Bedouin) (2.6%±1.37) and Q2 (Persian-South Asian) (5.0%±1.41) at levels on par with estimates for other populations sampled in the region (Supplemental Figure 8, Supplemental Table X), including Human Origins Bedouin and Saudi. This confirmed that recent African admixture is limited to the Q3 (African) subpopulation (37.6%±0.9), where this estimate is on par with African American populations. An estimate of the timing of African admixture placed the number of generations for Q1 (Bedouin) (15.2) and Q2 (Persian-South Asian) (14.0) slightly higher than Q3 (African) (9.3), consistent with the Q1 (Bedouin) and Q2 (Persian-South Asian) reflecting more distant African admixture events and with the Q3 (African) reflecting the historical timing of the African slave trade in the region (Omberg 2012).

The SupportMix analysis used 6 of the 1000 Genomes populations (2 European, 2 Asian, and 2 African) (see Supplemental Methods for details) as ancestral proxy reference panels and produced a set of "best guess" admixture assignments based on highest similarity to these genomes. While these 1000 Genomes populations do not include appropriate local populations most closely related to the Qataris needed for assessment of the true admixture composition of the genomes, the ancestry track length distribution of haplotypes assigned to African populations (Yoruba or Luhuya) provides a qualitative indicator of whether the subpopulations experienced recent admixture with African populations. As expected, the track lengths of the Q1 (Bedouin) and Q2 (Persian-South Asian) assigned to African 1000 Genomes populations were far shorter than those for Q3 (African) (Supplemental Figure 9), again confirming that recent African admixture is limited to the Q3 (African) subpopulation.

Neanderthal Ancestry

We next analyzed Neanderthal admixture contributions to the ancestry of Q1 (Bedouin) compared to the Q2 (Persian-South Asian) and Q3 (African) Qataris, the 1000 Genomes populations, and the populations of the Human Origins samples using the F_4 ratio and Patterson's *D-statistic* (Patterson 2012) (Figure 4, Supplemental Figure 10, Supplemental Table XI). The results for both methods were highly correlated (Supplemental Figure 10A). The Q1 (Bedouin; F_4 ratio=0.026, *D-statistic*=0.000) had more Neanderthal admixture than all African populations, including Q3 (African; F_4 ratio range=-0.017 to 0.024, *D-statistic* range=-0.031 to -0.003). The Q1 (Bedouin) also had Neanderthal admixture at levels comparable to Q2 (Persian South-Asian; F_4 ratio=0.024, *D-statistic* = -0.003) and to other Middle Eastern populations, including other Bedouin populations (Human Origins Bedouin A F_4 ratio=0.022, *D-statistic*=-0.003 and Bedouin B F_4 ratio=0.024, *D-statistic*=-0.003) and Saudi (F_4 ratio=0.026, *D-statistic*=-0.001). Interestingly, the Q1 (Bedouin) did not tend to have higher Neanderthal admixture levels when considering populations outside of the Middle East, where the bulk of European populations had higher Neanderthal admixture (F_4 ratio range=0.018 to 0.041, *D-statistic* range=0.003 to 0.010). Yet, the percent Neandethal admixture with the Q1 (Bedouin) was higher than expected if it could be entirely explained by later admixture events between the Q1 (Bedouin) and Europeans (observed F_4 ratio=0.026 *vs* expected F_4 ratio=0.00247).

The higher Neanderthal ancestry in the Q1 (Bedouin) Qatari compared to African populations places the divergence of ancestral Arabs after the out of Africa bottleneck. Given the current evidence of the geographic range of Neanderthal populations stretching from Europe and the Mediterranean through Northern and Central Asia (Fu 2014; Hershkovitz 2015), the lower Neanderthal Ancestry in the Q1 (Bedouin) Qatari compared to populations within the ancestral Neanderthal range is also consistent with an early divergence of the ancestors of indigenous Arabs from other lineages that populated Asia and Europe. Yet, since the Neanderthal admixture in the Q1 (Bedouin) cannot be entirely explained by admixture with Europeans, this indicates there was some admixture between Neanderthals and ancestors of the Q1 (Bedouin) in the region of the Arabian Peninsula.

TreeMix Analysis

We also analyzed the autosomes of the combined 96 Q1 (Bedouin), Q2 (Perisan-South Asian) or Q3 (African) Qataris and non-admixed populations of the 1000 Genomes using the population split and mixture inference method TreeMix (Pickrell and Pritchard 2012) to assess the relative genetic similarity of populations based on high-density, genome-wide allele frequencies. The analysis returned an overall tree for the 1000 Genomes populations that mirrored those found previously (Shriner 2014) with the addition of the Q1 (Bedouin) and Q2 (Persian-South Asian) clustering on the branch that includes Europeans (Perez-Miranda 2006) and the Q3 (African) clustering with African populations (Figure 5). When migrations were allowed in the analysis, no migration events were observed between the Q1 (Bedouin) and African populations, even when allowing as many as 5 migration events (Supplemental Figure 11). These results are also consistent with what is known of the migration history of the Arabian Peninsula, including migration both to and from Europe during ancient and more recent eras of civilization, where this resulted in detectable admixture from European populations in both the Q1 (Bedouin) and Q2 (Persian-South Asian) (Omberg 2012).

Proportion of Shared Alleles Neighbor-joining Analysis

As the principal component analysis and the TreeMix population level clusterings depend on allele frequencies, the clustering of the Q1 (Bedouin) on a common branch with European populations could be driven by the haplotypes introduced by migrants, which would be expected to shift the allele frequencies of these populations towards each other. As such, these clusterings based on allele frequencies do not necessarily argue against significant and deep ancestry of the Q1 (Bedouin) on the Arabian Peninsula, as indicated by the levels of Neanderthal admixture in this subpopulation. Additionally, these population level clusterings are disproportionately influenced by common segregating alleles (Pickrell and Pritchard 2012), while rare alleles can be more informative about deeper shared ancestry (Mathieson and McVean 2014) as the identity by state of a rare variant can more accurately reflect identity by descent (Hochreiter 2013).

In contrast to population level clustering, a pairwise clustering of individual genomes based on shared variants provides a relative measure for comparing total shared ancestry between individuals. Also, when applied to a common set of genome-wide, high density markers that include the low-minor allele frequency alleles of the 1000 Genomes, such pairwise clustering also provides an appropriate weight to rare alleles. We therefore performed a proportion of shared alleles (Mountain and Cavalli-Sforza 1997) analysis on the combined samples in the 104 Qatari and 1000 Genomes samples, where pairwise proportion of shared alleles was calculated for the 11,711,386 autosomal, biallelic SNPs segregating in both the 104 Qatari and 1000 Genomes samples. A robust version of the neighbor-joining algorithm was used to perform a pairwise clustering of the samples (Criscuolo and Gascuel 2008) (Figure 6A-F), where bootstrap support values were calculated for the observed trees using 100 random samplings of the SNPs.

The neighbor-joining analysis revealed that 50 of the 56 Q1 (Bedouin), along with 3 Q2 (Persian-South Asian), 1 Q3 (African) and 2 Q0 (Subpopulation Unassigned) Qataris, clustered outside African lineages and were also the most extreme outgroup that are basal to all non-African populations lacking recent African admixture (Figure 6D). Strong bootstrap support was observed for this cluster (70 out of 100 iterations), and for its presence as an outgroup to the Eurasian cluster (68 out of 100 iterations), comparable to the support for the Japanese cluster (60 out of 100 iterations) and for the East Asians as an outgroup to Europeans and Americans (81 out of 100 iterations). The Q1 (Bedouin) therefore fit the criteria of having ancient migration from Africa and being most distantly related to all other non-Africans in total ancestry.

A total of 11 Q2 (Persian-South Asian), 3 Q1 (Bedouin), 1 Q3 (African), and 1 Q0 (Subpopulation Unassigned) defined an Asian outgroup more closely related to Asians than the main Q1 (Bedouin) outgroup (Figure 6C), likely driven by the ancestry of the the Q2 (Persian-South Asian) subpopulation traceable to Persia and South Asia (Omberg 2012) and indicating these individuals are most distantly related to other Asians present in this cluster. A total of 12 Q3 (African), 3 Q1 (Bedouin), 3 Q2 (Persian-South Asian), and 4 Q0 (Subpopulation Unassigned) cluster as long individual branches or small clusters between the major Q1 (Bedouin) cluster and the admixed individuals of African ancestry from Southwest US (ASW), potentially representing individuals with a higher proportion of African admixture. As expected from the analyses of population genetic similarity and prior neighbor-joining analysis of admixed populations (Kopelman 2013), the Q3 (African) and African-Americans do not form large clusters, but rather appear as multiple individual branches close to the indigenous African populations, most similar to their African admixture source (Figure 6E and 6F). A set of 3 Q2 (Persian South-Asian) clustered as an outgroup to the Tuscan Southern European (TSI) branch (Figure 6B), which is not unexpected given admixture with European populations (Omberg 2012; Pickrell 2014).

Discussion

The hypothesis that the first Eurasian populations were established on the Arabian Peninsula and that contemporary indigenous Arabs are direct descendants of this ancient population is supported by two major conclusions derived from the combined evidence of this study. First, the analysis results for X/A diversity, the pairwise sequential Markov coalescent, genome-wide admixture, timing of African admixture, local admixture deconvolution, Neanderthal admixture, and application of TreeMix, support the inference that the Q1 (Bedouin) can trace the bulk of their ancestry back to the out of Africa migration events. Second, the combination of lower levels of Neanderthal admixture in the Q1 (Bedouin) than European / Asian populations and the outgroup position of the Q1 (Bedouin) compared to non-Africans in the pairwaise similarity clustering of high-density variants measured genome-wide, place the Q1 (Bedouin) as being the most distant relatives of other contemporary non-Africans. Given that the Q1 (Bedouin) have the greatest proportion of Arab genetic ancestry measured in contemporary populations (Shriner 2014; Hodgson 2014) and are among the best genetic representatives of the autochthonous population on the Arabian Peninsula, these two conclusions therefore point to the Bedouins being direct descendants of the earliest split after the out of Africa migration events that established a basal Eurasian population (Lazaridis 2014). This is also consistent with the majority of Q1 (Bedouin) being able to trace a significant portion of their autosomal ancestry through lineages that never left the peninsula after the out of Africa migration events since such deep ancestry would not be expected if the entire Arabian Peninsula population had been re-established from Africa or a non-African population at a later point.

Given the complex history of migration patterns to and from European populations, and the complicated patterns of isolation and intra- and inter-marriage of the indigenous Bedouin populations (Hunter-Zinck 2010; Sandridge 2010), it is not surprising that among the Q1 (Bedouin) are individuals who retain an autosomal signal of being the most distant relatives of non-Africans, while population-level clustering based on migration-shifted allele frequencies places the Q1 (Bedouin) closer to Europeans. The basal position of the Q1 (Bedouin) also has interesting implications for theories about the frequency, timing, and path of major migration waves that established populations in Asia and Europe (Shi 2008; Shriner 2014; Lazaridis 2014). A few isolated Asian populations were previously suspected to be descendants of a separate out of Africa migration wave based on Y Chromosome data (Hammer 1998; Shi 2008). Yet, distinct out of Africa migration events or separate migration waves emanating from the Arabian Peninsula into Europe and West Asia would be expected to place Bedouins / Europeans and Asians on separate branches of a pairwise clustering tree, distinct from our finding that places the Q1 (Bedouin) as direct descendants of the earliest lineage that split from the ancient non-African population.

A demographic scenario consistent with the evidence presented here is that the population ancestral to the Q1 (Bedouin) migrated out of Africa, and a subset of this population remained in the peninsula until the present day, while a second subset of this population migrated onwards and colonized Eurasia. This migration scenario implies the signal of the same bottleneck would be present in all non-African populations, which has been observed thus far in coalescent analysis of contemporary non-African populations (Gronau 2011; Schiffels and Durbin 2014; Fu 2014) and for an anatomically modern humans who lived 45,000 years ago (Fu 2014). This is also consistent with the recent discovery of another anatomically modern human who lived 55,000 years ago just northeast of the Arabian Peninsula that had morphological features similar to European peoples (Hershkovitz 2015), where this individual could have been a descendant of the basal Eurasian population that remained on the peninsula. Under this migration scenario, while other waves of migration may have occurred, the descendants of these alternative waves either left no descendants or were integrated into the dominant populations.

Beyond the importance for disentangling human migration history, an early split of Eurasian lineages in the Arabian Peninsula has implications for the study of disease genetics for indigenous people in the region. For example, for a disease such as type 2 diabetes that has a prevalence of over 18% in the Qatari population, associated genetic variants would not *a priori* be expected to be the same as those discovered in Europeans, when considering indigenous Arabs able to trace a significant portion of their ancestry back to ancient lineages on the Arabian Peninsula. More generally, this suggests that for any genome-wide association study (GWAS) or rare variant association study (RVAS) of diabetes or other complex diseases in Qatar, inference of deep ancestry in the Arabian Peninsula, using rare variation sampled by genome or exome sequencing, is critical for identifying new disease risk genes. Given the dearth of next generation sequencing studies conducted in Middle Eastern and Arab populations, these results indicate that a considerable number of variants that make important contributions to disease risk in these populations are yet to be discovered.

This study is the first analysis of Arabian Peninsula migration making use of deeply sequenced genomes from a sample of unrelated inhabitants of the peninsula. While there have been many analyses of Y Chromosomes and Mitochondrial DNA sampled from Arab individuals (Abu-Amero 2007; Rowold 2007; Abu-Amero 2008; Abu-Amero 2009), and there have been previous surveys of genetic variation of people within the peninsula and immediately surrounding regions conducted with genotyping arrays (Hunter-Zinck 2010; Behar 2010; Alsmadi 2013; Shriner 2014; Markus 2014) and deep exome sequencing (Rodriguez-Flores 2012; Rodriguez-Flores 2014; Alsmadi 2014), and by individual high-coverage genomes (Alsmadi 2014; John 2015), the sample of rare and common genetic variation throughout the genome in our sample provides a far more complete picture of how both ancient and recent migration events have contributed to the genetics of the modern peoples of the Arabian Peninsula. For understanding how human migration history has determined the structure of modern genomes, our identification of a cluster of Q1 (Bedouin) as the most distant ancestors of non-Africans is of considerable interest, particularly given the suspected route of migration out of Africa and into the surrounding continents. The possibility that the Q1 (Bedouin) are descendants of the first Eurasians provides an additional piece of the puzzle concerning ancient migration routes and the establishment of ancient non-African populations.

Methods

Ethics Statement

Human subjects were recruited and written informed consent was obtained at Hamad Medical Corporation (HMC) and HMC Primary Health Care Centers Doha, Qatar under protocols approved by the Institutional Review Boards of Hamad Medical Corporation and Weill Cornell Medical College in Qatar.

Inclusion Criteria

Qatar is a peninsula nation on the eastern edge of the Arabian Peninsula (Supplemental

Figure 1). The population of Qatar includes over 2 million inhabitants, comprised of approximately 300,000 nationals with roots in Qatar predating the discovery of oil and gas and establishment of an independent nation in 1970 and the over 1.7 million immigrants who mostly arrived in the past decade (Qatar Statistics Authority 2013). As selection criteria, we required that subjects be third generation Qataris where all ancestors were Qatari citizens born in Qatar, as assessed by questionnaires. Recent immigrants or residents of Qatar who traced their recent ancestry to other geographic regions were excluded.

Natives of the Arabian Peninsula can be divided into at least 3 genetic subpopulations that reflect the historical migration patterns in the region: Q1 (Bedouin), Q2 (Persian-South Asian) and Q3 (African) (Hunter-Zinck 2010; Omberg 2012; Rodriguez-Flores 2012). A panel of 48 SNPs was genotyped by TaqMan (Life Technologies, Carlsbad, CA) sufficient for classification into one of the 3 subpopulations based on >70% ancestry in one cluster in a STRUCTURE analysis with k=3 used to identify individuals that could unambiguously be placed in one of these three groups (Pritchard 2000; Rodriguez-Flores 2012) (Supplemental Figure 2). Our primary focus was the Q1 (Bedouin) genetic subpopulation because of its deepest ancestry in Arabia (Ferdinand 1993), so we selected 60 Q1 (Bedouin) individuals to include in the sample. We additionally selected 20 Q2 (Persian-South Asian) and 20 Q3 (African) to use as controls in the analysis, and an additional 8 Q0 (Subpopulation Unassigned) individuals that could not be confidently placed in one of these subpopulations, defined as not having >70% ancestry in any of the three groups as determined by STRUCTURE. The total sample therefore included 108 individuals with an even distribution of males and females (see Supplemental Methods, Supplemental Table I).

Illumina Deep Sequencing of the Genomes

In order to characterize the spectrum of genetic variation, each of the 108 Qatari genomes were sequenced to a median depth of 37* (minimum 30*) though the Illumina Genome Network (see Supplemental Methods for details).

Relatedness among Qataris

Given the high rate of consanguineous marriage previously reported in the Qatari population (Hunter-Zinck 2010; Mezzavilla 2015), we sought to quantify the relatedness between individuals in our sample, and to exclude closely related individuals that could potentially confound population genetics analysis methods that assume the input sample is unrelated. In order to conduct the relatedness analysis, autosomal SNPs in 108 Qatari genomes (described above) were filtered using PLINK 1.9 (Chang 2015) and relatedness between the 108 Qatari genomes was assessed using kinship coefficients estimated by KING-robust (Manichaikul 2010) and PREST-plus (McPeek and Sun 2000) (see Supplemental Methods). Both methods found the same five first-degree and second-degree relationships, where these relationships were then confirmed by investigative re-assessment of medical records. One individual from each of 5 pair of relatives was then excluded from the study. Three of the pairs of relatives formed a trio, hence two individuals were excluded from the trio, and one individual was excluded from each of the other two pairs, resulting in exclusion of 4 relatives in total.

Integration with 1000 Genomes Project Phase 1

An integrated SNP call set was produced for ancestry analysis for a total of 1200 genomes, combining the 108 Qatari genomes with the 1092 genomes from 1000 Genomes Project Phase 1 (1000 Genomes) (The 1000 Genomes Project Consortium 2012) (see Supplemental Methods). The integrated call set included 11,711,411 autosomal biallelic SNPs. The transition:transversion ratio of this final set was 2.2, close to values previously observed in 1000 Genomes (The 1000 Genomes Project Consortium 2012). Based on the concordance and quality measures, the calls generated from our pipeline were considered to be high quality and these were used for all further aspects of this study. After exclusion of 4 related Qataris (Supplemental Table III), the final integrated call set included 11,711,386 autosomal biallelic SNPs in 1,196 genomes.

Integration with Human Origins Dataset

The 1000 Genomes Phase 1 is an excellent resource for rare variant discovery; however it is limited in terms of the breadth of global populations sampled. Unfortunately, at the time of writing, no global resource of sequenced genomes exists, hence the next best alternative for comparison of the Qataris to populations around the world is the “Human Origins Fully Public Dataset” (referred to here as “Human Origins” or abbreviated as “HO”), which includes genotype data for 1917 indivduals from Africa, West Eurasia (including Middle East), South Asia, East Asia, Central Asia/Siberia, and America. In particular, the West Eurasian, African, and South Asian datasets include populations sampled in countries close to Qatar where detection of shared ancestry is of interest in this study. The dataset also includes data from archaic genomes, such as Altai Neanderthal, Denisova, and chimpanzee, which are of interest in this study for quantification of Neanderthal ancestry. The Human Origins dataset includes a number of samples also present in the 1000 Genomes (Supplemental Table IV), and for these samples the Human Origins overlap data is kept.

In order to conduct population genetic analysis on a combined dataset of the 104 Qatari genomes (QG, n=104), 1000 Genomes Phase 1 (1000G-HO, n=1028 after exclusion of duplicates), and Human Origins Fully Public Dataset (HO, n=1862 after exclusion of archaic genomes, ancient genomes, and other genomes not relevant to this study) (Supplemental Table V), a set of sites overlapping between the integrated Qatari genomes plus 1000 Genomes minus Human Origins, and the Human Origins dataset were identified. Out of 600,841 SNPs in the Human Origins dataset and 11,711,386 SNPs in the Qatari genomes plus 1000 Genomes dataset, 388,805

SNPs overlapped. Further filtering was conducted on the datased, pruning SNPs based on linkage disequilibrium using PLINK (Purcell 2007) [‘--indep-pairwise 200 25 0.4’, matching parameters used previously (Lazaridis 2014)]. After linkage disequilibrium-pruning, the final dataset for analysis included 197,714 SNPs segregating in the three datasets (QG, 1000G-HO, and HO).

Inbreeding Coefficient

In order to place the high reported consanguinity in Qatar in a global context, the inbreeding coefficient was calculated using PLINK 1.9 (Chang 2015) for Q1 (Bedouin), Q2 (Persian-South Asian), and Q3 (African) Qataris, 1000 Genomes minus Human Origins overlap, and Human Origins populations (see Supplemental Methods).

Principal Component Analysis

A principal component analysis (PCA) (Price 2006) was carried out for the combined 104 Qatari genome, 1000 Genomes minus Human Origins overlap, and Human Origins samples using the 197,714 SNPs in the integrated dataset (filtering criteria described above). Using the results of this large-scale analysis, visual assessment of clustering and population overlap was used to confirm expected relationships between the analyzed populations. Four distinct views of a single PCA run were constructed: one comparing the Qatari genomes to 1000 Genomes populations (Supplemental Figure 5A), one comparing Qataris to 1000 Genomes and Human Origins Samples including two visualizations of the full dataset (Figure 1A color-coded by regional meta-populations and Supplemental Figure 5B color-coded by detailed population), and one comparing Qataris to Middle Eastern populations from the Human Origins dataset (Figure 1B). For the latter, in order to compare Qataris to Middle Eastern populations with potential for recent shared Bedouin ancestry with Qataris sampled by the Human Origins dataset, populations from the Middle East previously labeled in Lazardis et al (Lazaridis 2014) as “West Eurasia” were relabed as “Middle East”, including Bedouin A, Bedouin B, Druze, Egyptian Comas, Egyptian Metspalu, Iranian, Jordanian, Lebanese, Palestinian, Saudi, Syrian, Turkish, Turkish Adana, Turkish Aydin, Turkish Balikesir, Turkish Istanbul, Turkish Kayseri, Turkish Trabzon, and Yemen.

Y and Mitochondria Haplogroup Assignment

In order to determine the prevalence of known ChrY and MtDNA haplogroups in Qatar, SNP genotypes were generated simultaneously for the 108 Qatari genomes using an updated version of GATK (v3.1.1) that supports haploid chromosome calling (n=53 ChrY, n=108 MtDNA). For one of the genomes, the sample was originally thought to be male but is most likely female due to low call rates on ChrY. This sample was excluded from ChrY analysis and X/A diversity analysis, but was included in autosomal and MtDNA analysis. Mean coverage of mapped reads was 11* in ChrY and 3,892* in MtDNA. After exclusion of related and Subpopulation Unassigned (admixed) Qataris, the remaining samples included 47 ChrY and 96 MtDNA.

Haplogroup assignments for the ChrY and MtDNA were made using previously characterized variants. For ChrY, these assignments were made using YFitter (Jostins 2014) using variants limited to known SNPs catalogued by the International Society of Genetic Genealogy (Jobling and Tyler-Smith 2003) within a 10 Mb interval of the Y Chromosome that is known to be amenable to analysis based on short read sequencing (Skaletsky 2003; Poznik 2013). For mtDNA, these assignments were made using HaploGrep (Kloss-Brandstatter 2011) using the set of known haplogroup-specific variants in the PhyloTree (van Oven and Kayser 2009) database.

In order to quantify the differences between MtDNA and ChrY in terms of diversity of haplogroups identified, the proportion of variance among and within populations was quantified for ChrY and MtDNA using the AMOVA function in Arlequin (Excoffier 1992; Excoffier and Lischer 2010) (see Supplemental Methods). The analysis was repeated 8 times, including separate analysis of ChrY and MtDNA, for three-way comparison of the populations, as well as all possible 2-way comparisons (Q1/Q2, Q1/Q3, Q2/Q3). The proportion of variance among and within populations was tabulated, as well as the estimated F_{st} and p-value for both.

Comparison of X Chromosome to Autosomal Diversity

The ratio of X-linked to autosomal nucleotide diversity (X/A) for different populations was computed following the approach in Gottipati et al (Gottipati 2011) and Arbiza et al (Arbiza 2014) (see Supplemental Methods).

Coalescent Analysis

To infer the extent and timing of bottlenecks, the pairwise sequential Markov coalescent (PSMC) (Li and Durbin 2011) was applied to the 96 Q1 (Bedouin), Q2 (Persian-South Asian) or Q3 (African) Qatari genomes. A plot of effective population size *vs* years in the past was generated for each of the genome using instructions from the PSMC manual (Li and Durbin 2011) (see Supplemental Methods). For comparison, the same PSMC pipeline was run on BAM files of Illumina deep sequencing reads mapped to the GRCh37 human reference genome for an individual of European ancestry (NA12878, Utah resident with Northern and Western European ancestry, CEU) and an individual of African ancestry (NA19239, Yoruba in Ibadan, Nigeria, YRI) sequenced as part of the 1000 Genomes Pilot (The 1000 Genomes Project Consortium 2010). The resulting PSMC plots for these two individuals were shifted slightly, such that they align with Qatari PSMC plots at distant (>200 thousand years ago) timescales (Fu 2014).

Genome-Wide Admixture Analysis

In order to learn more about the ancestry of the sampled Qataris, a genome-wide admixture analysis was conducted on the combined dataset of 104 Qatari genomes, 1000 Genomes minus Human Origins overlap, and Human Origins using ADMIXTURE (Alexander 2009) (see Supplemental Methods). The cross-validation error was calculated for a range of expected number of ancestral populations (K), and the K with the lowest cross-validation error was used to quantify ancestry, in this case K=12.

African Admixture Proportion and Timing

In order to estimate the proportion and timing of African admixture in Qatari populations, the genomes of Qataris and world populations were analyzed using ALDER 1.2 (Loh 2013) (see Supplemental Methods).

Local Admixture Analysis

An admixture deconvolution analysis was performed on the 96 Q1 (Bedouin), Q2 (Persian-South Asian), or Q3 (African) Qatari genomes using the 11,711,386 autosomal SNPs segregating in both 1000 Genomes and Qatari genomes using SupportMix (Omberg 2012) (Supplemental Figure 9; see Supplemental Methods).

Neanderthal Ancestry

In order to compare the proportion of Neanderthal admixture in Q1 (Bedouin) Qataris with that of other populations in 1000 Genomes (The 1000 Genomes Project Consortium 2012) and Human Origins (Lazaridis 2014), the F_4 ratio (Patterson 2012) and Patterson's *D-statistic* (Patterson 2012) were estimated using the qpF4ratio and qpDstat programs, respectively, from the the AdmixTools 3.0 package (Patterson 2012) (see Supplemental Methods).

We additionally considered the expected F_4 ratio for the Q1 (Bedouin) under the scenario of no admixture between Neanderthal and direct ancestors of Q1 (Bedouin), such that observed Neanderthal ancestry in Q1 (Bedouin) would be entirely due to European admixture. From the estimated components of the ADMIXTURE analysis with K=12, the Southern European ancestry in the Q1 (Bedouin) is 8.2% on average, and the Northern European ancestry in Q1 (Bedouin) is 1.3% on average, totaling 9.5% of the genome. If the Q1 (Bedouin) had never mixed with Neanderthal prior to introduction of European admixture, assuming no selection against introgressed genomic intervals, we would therefore expect an F_4 ratio in Q1 (Bedouin) to be on the order of 1/10 of those observed in European populations.

TreeMix Analysis

We performed a TreeMix analysis (Pickrell and Pritchard 2012) of the 96 Q1 (Bedouin), Q2 (Persian-South Asian), or Q3 (African) Qatari genomes and 1000 Genomes excluding admixed populations (Puerto Rican, Mexican, Colombian, and African Ancestry in Southwest US; see Supplemental Methods).

Neighbor-joining Tree Clustering

In order to determine if any of the Qatari genomes were the most distant ancestors of all non-African populations, neighbor-joining trees were constructed for the 104 Qatari genomes and 1000 Genomes using the 11,711,386 autosomal SNPs segregating in both datasets. For each pair of genomes, the proportion of shared alleles (PSA) (Mountain and Cavalli-Sforza 1997) was calculated using the “--distance -square -1-ibs” function in PLINK 1.9 (Purcell 2007; Chang 2015), which outputs a 1196*1196 matrix of distances (1 minus IBS distance or PSA). A neighbor-joining (NJ) tree was constructed using a recently updated version of the original NJ (Saitou and Nei 1987) algorithm called NJS (Criscuolo and Gascuel 2008) that is better at handling missing values, as implemented in the APE package in R (Paradis 2004). Overall, this approach is computationally tractable for millions of markers genotyped in thousands of genomes and produces similar topologies to maximum-likelihood clustering methods but requires only a fraction of the compute time, where the trade-off is a sacrifice in the accuracy of branch lengths (Tateno 1994). The algorithm takes the distance matrix as input and outputs a tree. In order to confirm the robustness to sample ordering, the order of samples in the matrix was shuffled and re-clustered 100 times, where all re-clusterings recovered the same tree. In order to produce bootstrap support values for the tree, 100 re-clusterings of the tree were generated based on random sampling of SNPs. For each bootstrap iteration, 11,711,386 random (with replacement) SNPs were selected using a Python (Python Software Foundation 2015) script, and then the PSA distance matrix and NJ tree were recalculated using these SNPs. Bootstrap support was calculated using the Python package SumTrees (Sukumaran and Holder 2010).

For visualization, the tree was rooted at the most recent common ancestor (MRCA) node of the largest cluster of 1000 Genomes Yoruba (YRI) genomes in the tree. A color version of the tree was produced using TreeGraph 2 (Stover and Muller 2010) by manually coloring the branches leading to each node. A single color is assigned to each population, with populations from the same continent having similar colors: Europeans in shades of purple, Asians in shades of brown, Americans in shades of green, Africans in shades of orange, Q1 (Bedouin) in red, Q2 (Persian-South Asian) in blue, Q3 (Sub-Saharan African) in black, and Q0 (Subpopulation Unassigned) in grey. When a cluster of nodes includes different populations, the terminal branches were given population-specific colors, while the shared higher-order branches for the cluster were given the color of the population in majority. For example, if 10 Q1 (Bedouin) and 1 Q0 (Subpopulation Unassigned) were in a cluster, the branches above where the nodes come together were colored red.

Data Access

The read data generated for this study in BAM format, as well as SNP genotypes in VCF format, are available at the NCBI Sequence Read Archive under BioProject accession number PRJNA288297 (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA288297). Allele frequencies for known and novel genomic SNPs were submitted to NCBI dbSNP and can be searched for using batch ID 1062298 and handle WEILL_CORNELL_DGM (http://www.ncbi.nlm.nih.gov/SNP/snp_viewBatch.cgi?sbid=1062298). PLINK and VCF files of genotypes for variants analyzed in this study, both before and after integration with 1000 Genomes and Human Origins, are available on our website http://geneticmedicine.weill.cornell.edu/genome.html.

.

Acknowledgments. We thank the three reviewers of this manuscript for helpful comments and suggestions, the 1000 Genomes Project for helpful advice on analysis methods; MR Staudt, Y Strulovici-Barel, A Al Shakaki, OM Chidiac, R Mathew, and the WCMC-Q Genomics Core for help with the study; Mezey lab students and N Mohamed for help in preparing this manuscript. These studies were supported, in part, by the Qatar Foundation and Weill Cornell Medical College in Qatar. JLRF was supported, in part, by NIH T32 HL09428.

Conflict of interest: The authors have no conflict to declare.

References

Abu-Amero, K.K., Gonzalez, A.M., Larruga, J.M., Bosley, T.M., and Cabrera, V.M. 2007. Eurasian and African mitochondrial DNA influences in the Saudi Arabian population. *BMC.Evol.Biol.* **7**:32

Abu-Amero, K.K., Hellani, A., Gonzalez, A.M., Larruga, J.M., Cabrera, V.M., and Underhill, P.A. 2009. Saudi Arabian Y-Chromosome diversity and its relationship with nearby regions. *BMC.Genet.* **10**:59

Abu-Amero, K.K., Larruga, J.M., Cabrera, V.M., and Gonzalez, A.M. 2008. Mitochondrial DNA structure in the Arabian Peninsula. *BMC.Evol.Biol.* **8**:45

Alexander, D.H., Novembre, J., and Lange, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**:1655-1664

Alsmadi, O., John, S.E., Thareja, G., Hebbar, P., Antony, D., Behbehani, K., and Thanaraj, T.A. 2014. Genome at juncture of early human migration: a systematic analysis of two whole genomes and thirteen exomes from Kuwaiti population subgroup of inferred Saudi Arabian tribe ancestry. *PLoS.One.* **9**:e99069

Alsmadi, O., Thareja, G., Alkayal, F., Rajagopalan, R., John, S.E., Hebbar, P., Behbehani, K., and Thanaraj, T.A. 2013. Genetic substructure of Kuwaiti population reveals migration history. *PLoS.One.* **8**:e74913

Arbiza, L., Gottipati, S., Siepel, A., and Keinan, A. 2014. Contrasting X-linked and autosomal diversity across 14 human populations. *Am J Hum Genet* **94**:827-844

Armitage, S.J., Jasim, S.A., Marks, A.E., Parker, A.G., Usik, V.I., and Uerpmann, H.P. 2011. The southern route "out of Africa": evidence for an early expansion of modern humans into Arabia. *Science* **331**:453-456

Behar, D.M., Yunusbayev, B., Metspalu, M., Metspalu, E., Rosset, S., Parik, J., Rootsi, S., Chaubey, G., Kutuev, I., Yudkovsky, G., et al. 2010. The genome-wide structure of the Jewish people. *Nature* **466**:238-242

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**:53-59

Cann, R.L., Stoneking, M., and Wilson, A.C. 1987. Mitochondrial DNA and human evolution. *Nature* **325**:31-36

Cavalli-Sforza, L.L. and Feldman, M.W. 2003. The application of molecular genetic approaches to the study of human evolution. *Nat.Genet.* **33 Suppl**:266-275

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* **4**:7

Criscuolo, A. and Gascuel, O. 2008. Fast NJ-like algorithms to deal with incomplete distance matrices. *BMC.Bioinformatics.* **9**:166

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del, A.G., Rivas, M.A., Hanna, M., et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat.Genet.* **43**:491-498

Esposito, J.L. 2001. *Women in Muslim Family Law*, 2nd Edition edition. Syracuse Universtiy Press, Syracuse, NY.

Excoffier, L. and Lischer, H.E. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol.Resour.* **10**:564-567

Excoffier, L., Smouse, P.E., and Quattro, J.M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**:479-491

Ferdinand, K., Nicolaisen, I., and Carlsberg Foundation's Nomad Research Project. 1993. *Bedouins of Qatar*. Thames and Hudson, New York.

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L., ximu-Petri, A., Prufer, K., de, F.C., et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**:445-449

Gottipati, S., Arbiza, L., Siepel, A., Clark, A.G., and Keinan, A. 2011. Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nat Genet* **43**:741-743

Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G., and Siepel, A. 2011. Bayesian inference of ancient human demography from individual genome sequences. *Nat.Genet.* **43**:1031-1034

Hammer, M.F., Karafet, T., Rasanayagam, A., Wood, E.T., Altheide, T.K., Jenkins, T., Griffiths, R.C., Templeton, A.R., and Zegura, S.L. 1998. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol.Biol.Evol.* **15**:427-441

Henn, B.M., Cavalli-Sforza, L.L., and Feldman, M.W. 2012. The great human expansion. *Proc.Natl.Acad.Sci.U.S.A* **109**:17758-17764

Hershkovitz, I., Marder, O., Ayalon, A., Bar-Matthews, M., Yasur, G., Boaretto, E., Caracuta, V., Alex, B., Frumkin, A., Goder-Goldberger, M., et al. 2015. Levantine cranium from Manot Cave (Israel) foreshadows the first European modern humans. *Nature*

Hochreiter, S. 2013. HapFABIA: identification of very short segments of identity by descent characterized by rare variants in large sequencing data. *Nucleic Acids Res.* **41**:e202

Hodgson, J.A., Mulligan, C.J., Al-Meeri, A., and Raaum, R.L. 2014. Early back-to-Africa migration into the Horn of Africa. *PLoS.Genet.* **10**:e1004393

Hunter-Zinck, H., Musharoff, S., Salit, J., Al-Ali, K.A., Chouchane, L., Gohar, A., Matthews, R., Butler, M.W., Fuller, J., Hackett, N.R., et al. 2010. Population genetic structure of the people of Qatar. *Am.J.Hum.Genet.* **87**:17-25

Jobling, M.A. and Tyler-Smith, C. 2003. The human Y chromosome: an evolutionary marker comes of age. *Nat.Rev.Genet.* **4**:598-612

John, S.E., Thareja, G., Hebbar, P., Behbehani, K., Thanaraj, T.A., and Alsmadi, O. 2015. Kuwaiti population subgroup of nomadic Bedouin ancestry – Whole genome sequence and analysis. *Genom Data* **3**:116-127

Jostins, L., Xu, Y., McCarthy, S., Ayub, Q., Durbin, R., Barrett, J., and Tyler-Smith, C. 2014. YFitter: Maximum likelihood assignment of Y chromosome haplogroups from low-coverage sequence data. *arXiv:1407.7988*

Keinan, A., Mullikin, J.C., Patterson, N., and Reich, D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* **39**:1251-1255

Kloss-Brandstatter, A., Pacher, D., Schonherr, S., Weissensteiner, H., Binna, R., Specht, G., and Kronenberg, F. 2011. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum.Mutat.* **32**:25-32

Kopelman, N.M., Stone, L., Gascuel, O., and Rosenberg, N.A. 2013. The behavior of admixed populations in neighbor-joining inference of population trees. *Pac.Symp.Biocomput.*:273-284

Lahiri, S.N. 2003. *Resampling methods for dependent data*. Springer, New York.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**:409-413

Li, H. and Durbin, R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**:493-496

Liu, R.Y. and Singh, K. 1992. Moving blocks jackknife and bootstrap capture weak dependence. In Exploring the limits of bootstrap (eds. R. LePage, L. Billard), pp. 225-248. John Wiley, New York.

Loh, P.R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., and Berger, B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**:1233-1254

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics.* **26**:2867-2873

Markus, B., Alshafee, I., and Birk, O.S. 2014. Deciphering the fine-structure of tribal admixture in the Bedouin population using genomic data. *Heredity (Edinb.)* **112**:182-189

Mathieson, I. and McVean, G. 2014. Demography and the age of rare variants. *PLoS.Genet.* **10**:e1004528

McPeek, M.S. and Sun, L. 2000. Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am J Hum Genet* **66**:1076-1094

Mezzavilla, M., Vozzi, D., Badii, R., Alkowari, M.K., Abdulhadi, K., Girotto, G., and Gasparini, P. 2015. Increased rate of deleterious variants in long runs of homozygosity of an inbred population from Qatar. *Hum Hered.* **79**:14-19

Mountain, J.L. and Cavalli-Sforza, L.L. 1997. Multilocus genotypes, a tree of individuals, and human evolutionary history. *Am J Hum.Genet.* **61**:705-718

Omberg, L., Salit, J., Hackett, N., Fuller, J., Matthew, R., Chouchane, L., Rodriguez-Flores, J.L., Bustamante, C., Crystal, R.G., and Mezey, J.G. 2012. Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. *BMC.Genet.* **13**:49

Paradis, E., Claude, J., and Strimmer, K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics.* **20**:289-290

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. 2012. Ancient admixture in human history. *Genetics* **192**:1065-1093

Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S., and Reich, D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**:1103-1108

Perez-Miranda, A.M., fonso-Sanchez, M.A., Pena, J.A., and Herrera, R.J. 2006. Qatari DNA variation at a crossroad of human migrations. *Hum.Hered.* **61**:67-79

Pickrell, J.K., Patterson, N., Loh, P.R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., and Reich, D. 2014. Ancient west Eurasian ancestry in southern and eastern Africa. *Proc Natl Acad Sci U S A* **111**:2632-2637

Pickrell, J.K. and Pritchard, J.K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS.Genet.* **8**:e1002967

Poznik, G.D., Henn, B.M., Yee, M.C., Sliwerska, E., Euskirchen, G.M., Lin, A.A., Snyder, M., Quintana-Murci, L., Kidd, J.M., Underhill, P.A., et al. 2013. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**:562-565

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat.Genet.* **38**:904-909

Pritchard, J.K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**:945-959

Prufer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de, F.C., et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**:43-49

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am.J.Hum.Genet.* **81**:559-575

Python Software Foundation. www.python.org [last accessed 10/10/15]

Qatar Statistics Authority; Table 4.3 Population (15+) by educational attainment, nationality, age, sex and marital status. http://www.qsa.gov.qa/QatarCensus/Pdf/Population%20above%2015%20by%20educational%20attainment,%20nationality,%20age,%20sex%20and%20marital%20status.pdf [last accessed 10/10/15]

Rambaut, A. http://tree.bio.ed.ac.uk/software/figtree/ [last accessed 10/10/15]

Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K.E., Rasmussen, S., Albrechtsen, A., Skotte, L., Lindgreen, S., Metspalu, M., Jombart, T., et al. 2011. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334**:94-98

Reich, D., Thangaraj, K., Patterson, N., Price, A.L., and Singh, L. 2009. Reconstructing Indian population history. *Nature* **461**:489-494

Rodriguez-Flores, J.L., Fakhro, K., Hackett, N.R., Salit, J., Fuller, J., gosto-Perez, F., Gharbiah, M., Malek, J.A., Zirie, M., Jayyousi, A., et al. 2014. Exome sequencing identifies potential risk variants for mendelian disorders at high prevalence in qatar. *Hum.Mutat.* **35**:105-116

Rodriguez-Flores, J.L., Fuller, J., Hackett, N.R., Salit, J., Malek, J.A., Al-Dous, E., Chouchane, L., Zirie, M., Jayoussi, A., Mahmoud, M.A., et al. 2012. Exome sequencing of only seven Qataris identifies potentially deleterious variants in the Qatari population. *PLoS.One.* **7**:e47614

Rowold, D.J., Luis, J.R., Terreros, M.C., and Herrera, R.J. 2007. Mitochondrial DNA geneflow indicates preferred usage of the Levant Corridor over the Horn of Africa passageway. *J Hum.Genet.* **52**:436-447

Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol.Biol.Evol.* **4**:406-425

Sandridge, A.L., Takeddin, J., Al-Kaabi, E., and Frances, Y. 2010. Consanguinity in Qatar: knowledge, attitude and practice in a population born between 1946 and 1991. *J Biosoc.Sci.* **42**:59-82

Schiffels, S. and Durbin, R. 2014. Inferring human population size and separation history from multiple genome sequences. *Nat.Genet.* **46**:919-925

Shi, H., Zhong, H., Peng, Y., Dong, Y.L., Qi, X.B., Zhang, F., Liu, L.F., Tan, S.J., Ma, R.Z., Xiao, C.J., et al. 2008. Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. *BMC.Biol.* **6**:45

Shriner, D., Tekola-Ayele, F., Adeyemo, A., and Rotimi, C.N. 2014. Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry. *Sci.Rep.* **4**:6055

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**:825-837

Stover, B.C. and Muller, K.F. 2010. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC.Bioinformatics.* **11**:7

Sukumaran, J. and Holder, M.T. 2010. DendroPy: a Python library for phylogenetic computing. *Bioinformatics.* **26**:1569-1571

Tateno, Y., Takezaki, N., and Nei, M. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol.Biol.Evol.* **11**:261-277

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**:1061-1073

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**:56-65

The Comprehensive R Archive Network. https://cran.r-project.org/ [last accessed 10/10/15]

van Oven, M. and Kayser, M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat.* **30**:E386-E394

Figure Legends

Figure 1. Principal component analysis (PCA) (Price 2006) of the 104 Qatari genomes (circle), 1000 Genomes (triangle), and Human Origins (square) study samples. Shown are individuals plotted on principal components PC1 and PC2, with genomes color-coded by study and population, with the Q1 (Bedouin) in red, Q2 (Persian-South Asian) in dodger blue, and Q3 (African) in black. **A.** Plot of all populations, defined by study and by population, where all populations from the same region and study are grouped and color-coded together (1000 Genomes: Africa, America, East Asia, and Europe; Human Origins: Africa, America, Central Asia / Siberia, East Asia, Middle East, Oceania, South Asia, and West Eurasia). **B.** Plot of Middle Eastern subpopulations from Human Origins that cluster near Q1 (Bedouin) and Q2 (Persian-South Asian).

Figure 2. Y chromosome (ChrY) and Mitochondrial DNA (MtDNA) haplogroup assignments. The ChrY and MtDNA haplogroups were determined for Q1 (Bedouin), Q2 (Persian-South Asian), and Q3 (African). **A.** Pie charts of the haplogroup frequencies for ChrY. **B.** Pie charts of the haplogroup frequencies for MtDNA.

Figure 3. Ancient bottlenecks in the 96 Q1 (Bedouin), Q2 (Persian-South Asian), or Q3 (African) Qatari genomes (56 Q1, 20 Q2, 20 Q3) determined by pairwise sequential Markov coalescent analysis (Li and Durbin 2011). Shown is the plot of the median effective population size (*y*-axis) across individuals in a subpopulation *vs* years in the past (log scale *x*-axis) for the samples in the three major Qatari subpopulations: Q1 (Bedouin) in red, Q2 (Persian-South Asian) in dodger blue, Q3 (African) in black. A single individual of European ancestry (NA12879, violet) and a single individual of African ancestry (NA19239, orange) from the 1000 Genomes Project deep-coverage pilot (The 1000 Genomes Project Consortium 2010) are shown for comparison.

Figure 4. Neanderthal ancestry in world populations. F_4 ratio estimation as implemented in AdmixTools 3.0 (Patterson 2012) was used to calculate the Neanderthal ancestry proportion for each population in the combined dataset of Qatari genomes, 1000 Genomes, and Human Origins. The F_4 ratio estimates α, the proportion of Neanderthal ancestry in a population. Shown are the results for populations of interest, including highest and lowest scoring populations from each region (1000 Genomes, Africa; 1000 Genomes, America; 1000 Genomes, East Asia, 1000 Genomes, Europe, Human Origins, Africa; Human Origins, America; Human Origins, Central Asia/Siberia; Human Origins, East Asia; Human Origins, Oceania; Human Origins, South Asia; Human Origins, West Eurasia), Middle Eastern populations (Human Origins), Q1 (Bedouin), Q2 (Persian-South Asian) and Q3 (African). Populations are color-coded by region, and a distinct color is used for each Qatari population. A full set of results is presented in Supplemental Figure 10 and Supplemental Table XI. The population codes are as in 1000 Genomes Project (The 1000 Genomes Project Consortium 2012).

Figure 5. TreeMix (Pickrell and Pritchard 2012) hierarchical clustering analysis of the Q1 (Bed ouin), Q2 (Persian-South Asian), and Q3 (African) and 1000 Genomes samples. Shown is a maximum-likelihood tree of population splits inferred without subsequent migration events, where branch lengths estimate divergence between populations (Europeans in shades of purple: CEU, FIN, GBR, IBS, TSI, East Asians in shades of brown: CHB, CHS, JPT, Africans in shades of orange: LWK, YRI, with the Q1 (Bedouin) in red, Q2 (Persian-South Asian) in dodger blue, and Q3 (African) in black). When allowing from 1-5 migration events in separate TreeMix analyses, none of the admixture loops connected the Q1 (Bedouin) with any African populations (Supplemental Figure 10), consistent with the Q1 (Bedouin) having no recent African admixture.

Figure 6. Neighbor-joining tree hierarchical clustering analysis of the combined Qatari genomes and 1000 Genomes Phase 1 samples based on pairwise proportion of shared alleles calculated across the entire autosome. **A.** The entire neighbor-joining tree with each of the branches leading to individuals in the 1000 Genomes samples color-coded by continent (Europeans in shades of purple: CEU, FIN, GBR, IBS, TSI, Asians in shades of brown: CHB, CHS, JPT, Africans in shades of orange: LWK, YRI, ASW, Americans in shades of green: CLM, MXL, PUR) and with the Q1 (Bedouin) in red, Q2 (Persian-South Asian) in dodger blue, Q3 (African) in black, and Q0 (Subpopulation Unassigned) in grey. **B.** Detail of the 3 (15%) Q2 (Persian-South Asian) that cluster with Europeans. **C.** Detail of the 11 (55%) Q2 (Persian-South Asian) individuals, with 3 (5%) Q1 (Bedouin), 1 (5%) Q3 (African), and 1 (13%) Q0 (Subpopulation Unassigned) that cluster as an outgroup to Asians. **D**. Detail of the 50 (89%) Q1 individuals, with 3 (15%) Q2 (Persian-South Asian), 1 (5%) Q3 (African), and 2 (25%) Q0 (Subpopulation Unassigned), that cluster outside the Africans and African Ancestry in Southwest US and that also cluster as an outgroup to all other non-African populations, indicating that they are the most distant ancestors of all non-Africans. **E.** Detail showing the 3 (15%) Q1 (Bedouin), 3 (15%) Q2 (Persian-South Asian),12 (60%) Q3 (African), and 4 (50%) Q0 (Subpopulation Unassigned) that do not form large clusters but are all located within the admixed cluster. **F.** Detail of the 1 (5%) Q3 (African) that clusters between Yoruba (YRI) and Luhya (LWK).

Figure 1

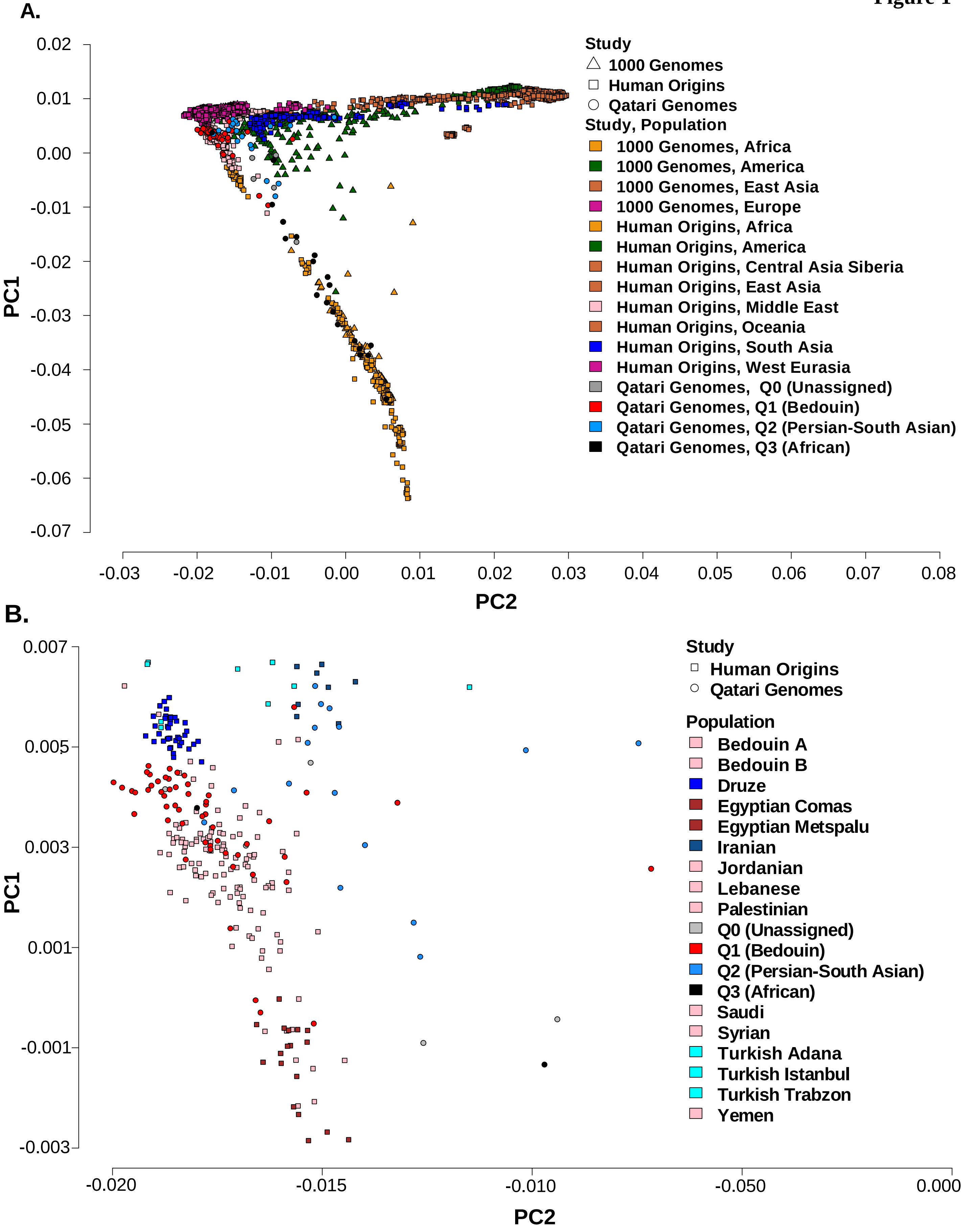
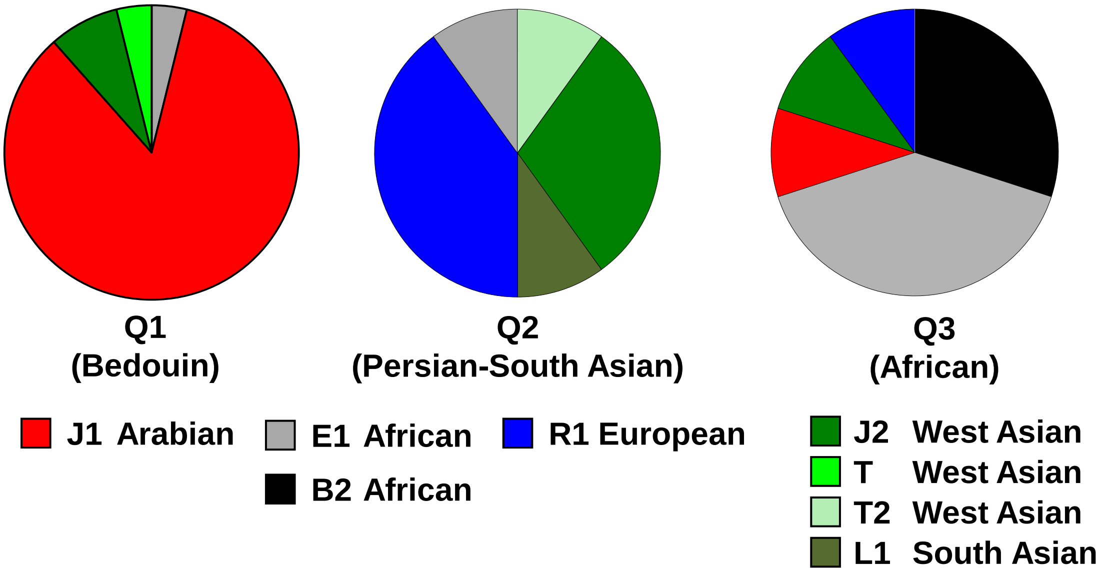

A.
PC1
0.02
0.01
0.00
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06
-0.07
PC2
-0.03
-0.02
-0.01
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Study
△ 1000 Genomes
☐ Human Origins
○ Qatari Genomes
Study, Population
1000 Genomes, Africa
1000 Genomes, America
1000 Genomes, East Asia
1000 Genomes, Europe
Human Origins, Africa
Human Origins, America
Human Origins, Central Asia Siberia
Human Origins, East Asia
Human Origins, Middle East
Human Origins, Oceania
Human Origins, South Asia
Human Origins, West Eurasia
Qatari Genomes, Q0 (Unassigned)
Qatari Genomes, Q1 (Bedouin)
Qatari Genomes, Q2 (Persian-South Asian)
Qatari Genomes, Q3 (African)
B.
PC1
0.007
0.005
0.003
0.001
-0.001
-0.003
PC2
-0.020
-0.015
-0.010
-0.050
0.000
Study
☐ Human Origins
○ Qatari Genomes
Population
Bedouin A
Bedouin B
Druze
Egyptian Comas
Egyptian Metspalu
Iranian
Jordanian
Lebanese
Palestinian
Q0 (Unassigned)
Q1 (Bedouin)
Q2 (Persian-South Asian)
Q3 (African)
Saudi
Syrian
Turkish Adana
Turkish Istanbul
Turkish Trabzon
Yemen

Figure 2

A. Y chromosome haplogroups



B. Mitochondrial DNA haplogroups

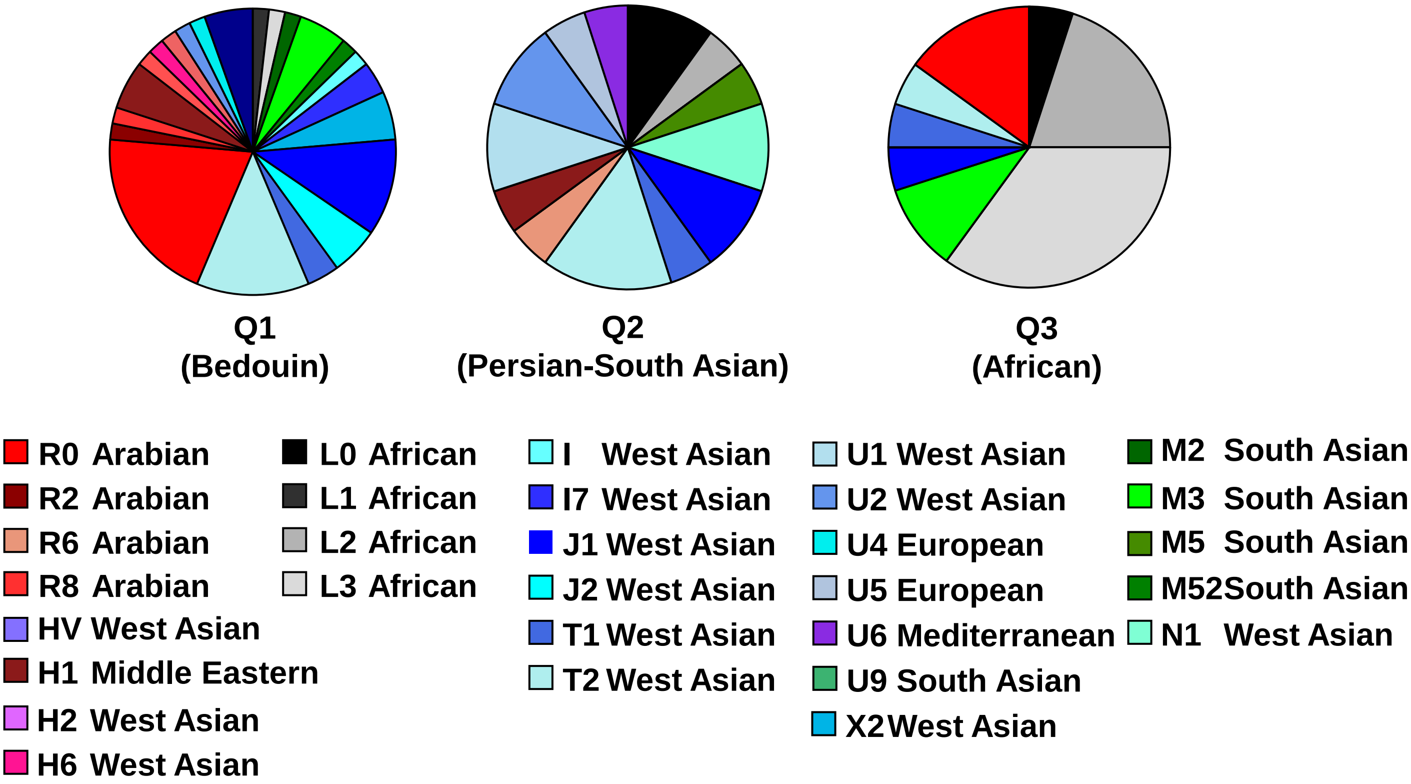
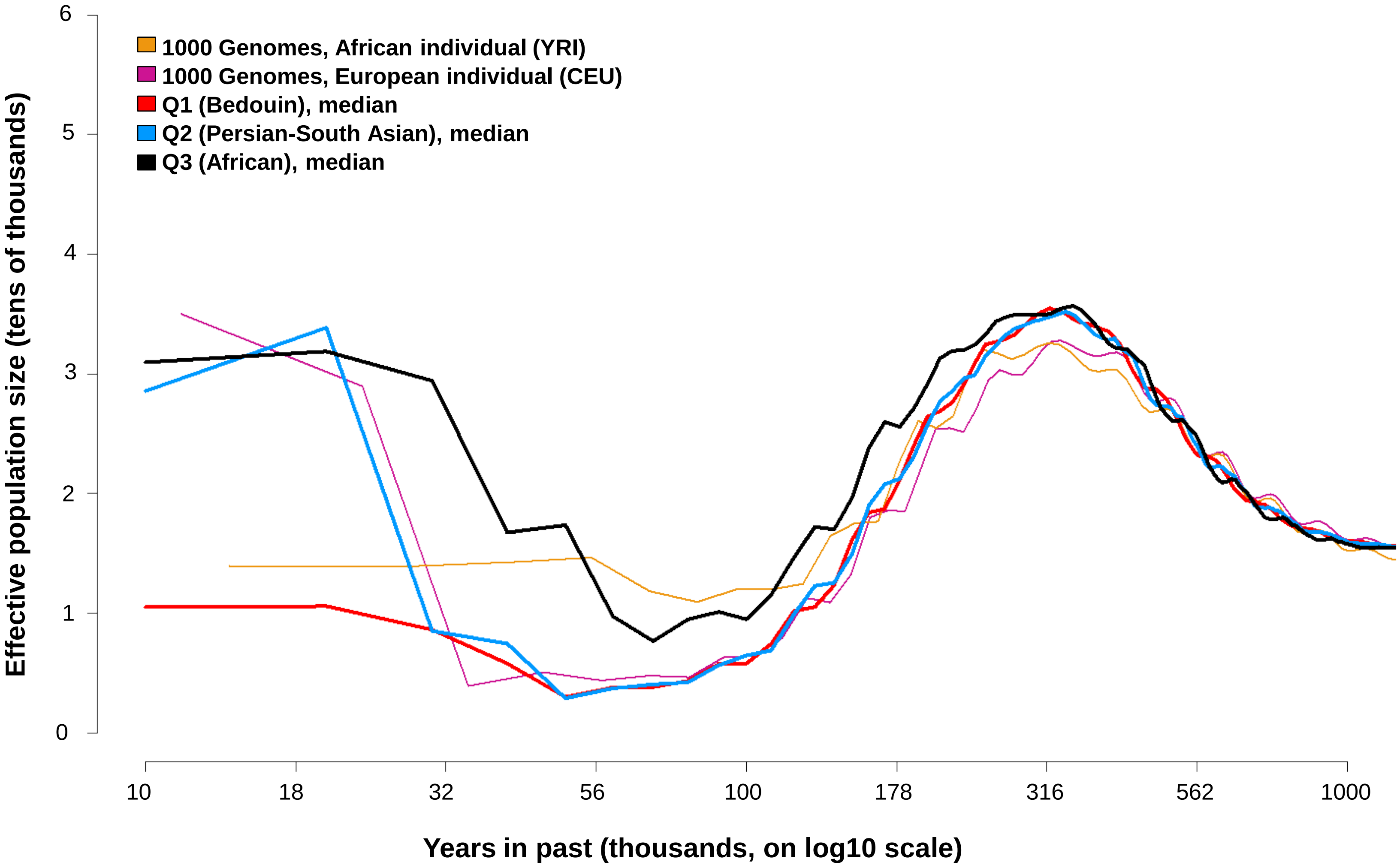
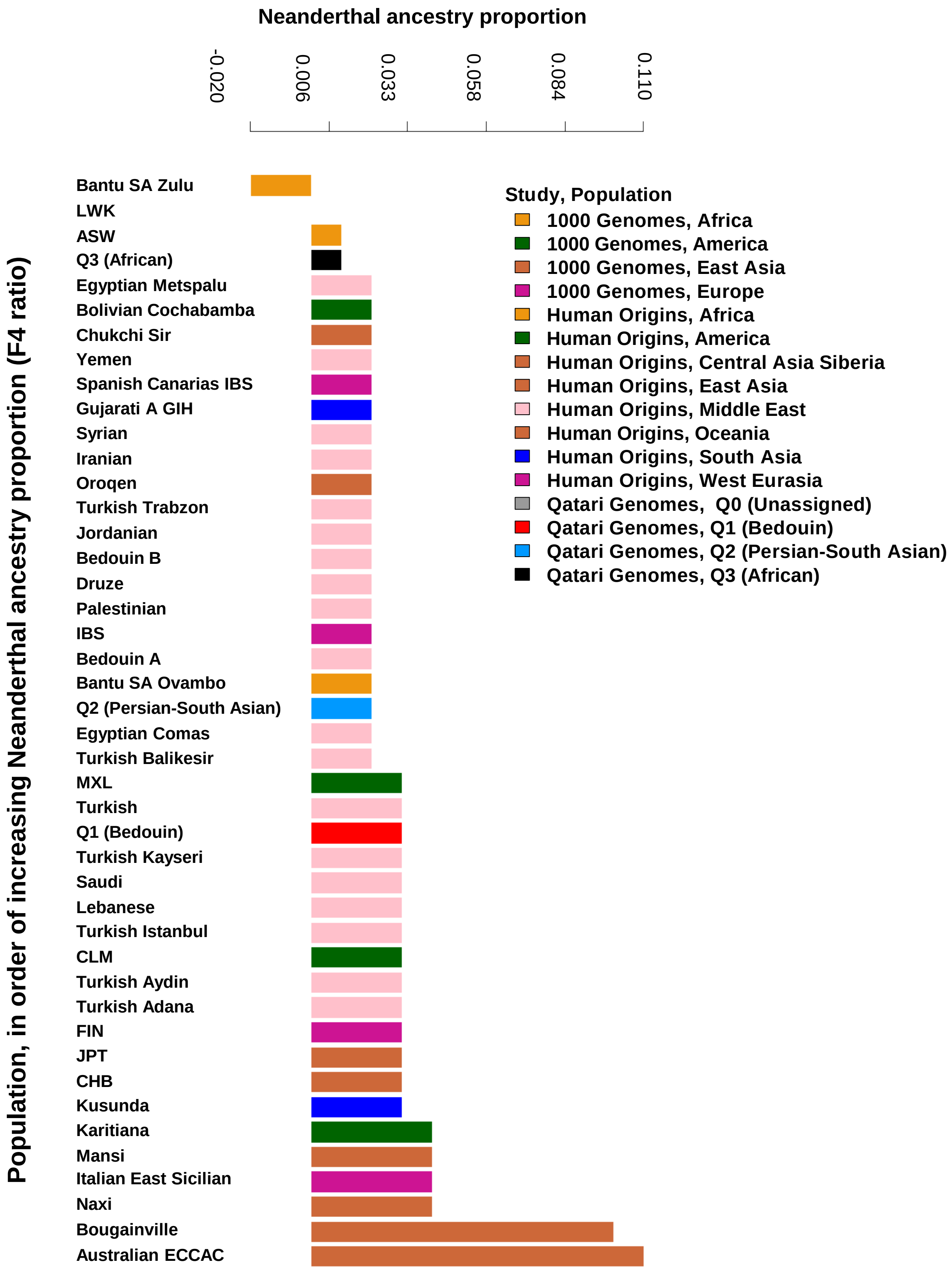


Figure 3



■ 1000 Genomes, African individual (YRI)
■ 1000 Genomes, European individual (CEU)
■ Q1 (Bedouin), median
■ Q2 (Persian-South Asian), median
■ Q3 (African), median
Effective population size (tens of thousands)
0
1
2
3
4
5
6
10
18
32
56
100
178
316
562
1000
Years in past (thousands, on log10 scale)

Figure 4


Neanderthal ancestry proportion
-0.020
0.006
0.033
0.058
0.084
0.110
Population, in order of increasing Neanderthal ancestry proportion (F4 ratio)
Bantu SA Zulu
LWK
ASW
Q3 (African)
Egyptian Metspalu
Bolivian Cochabamba
Chukchi Sir
Yemen
Spanish Canarias IBS
Gujarati A GIH
Syrian
Iranian
Oroqen
Turkish Trabzon
Jordanian
Bedouin B
Druze
Palestinian
IBS
Bedouin A
Bantu SA Ovambo
Q2 (Persian-South Asian)
Egyptian Comas
Turkish Balikesir
MXL
Turkish
Q1 (Bedouin)
Turkish Kayseri
Saudi
Lebanese
Turkish Istanbul
CLM
Turkish Aydin
Turkish Adana
FIN
JPT
CHB
Kusunda
Karitiana
Mansi
Italian East Sicilian
Naxi
Bougainville
Australian ECCAC
Study, Population
1000 Genomes, Africa
1000 Genomes, America
1000 Genomes, East Asia
1000 Genomes, Europe
Human Origins, Africa
Human Origins, America
Human Origins, Central Asia Siberia
Human Origins, East Asia
Human Origins, Middle East
Human Origins, Oceania
Human Origins, South Asia
Human Origins, West Eurasia
Qatari Genomes, Q0 (Unassigned)
Qatari Genomes, Q1 (Bedouin)
Qatari Genomes, Q2 (Persian-South Asian)
Qatari Genomes, Q3 (African)

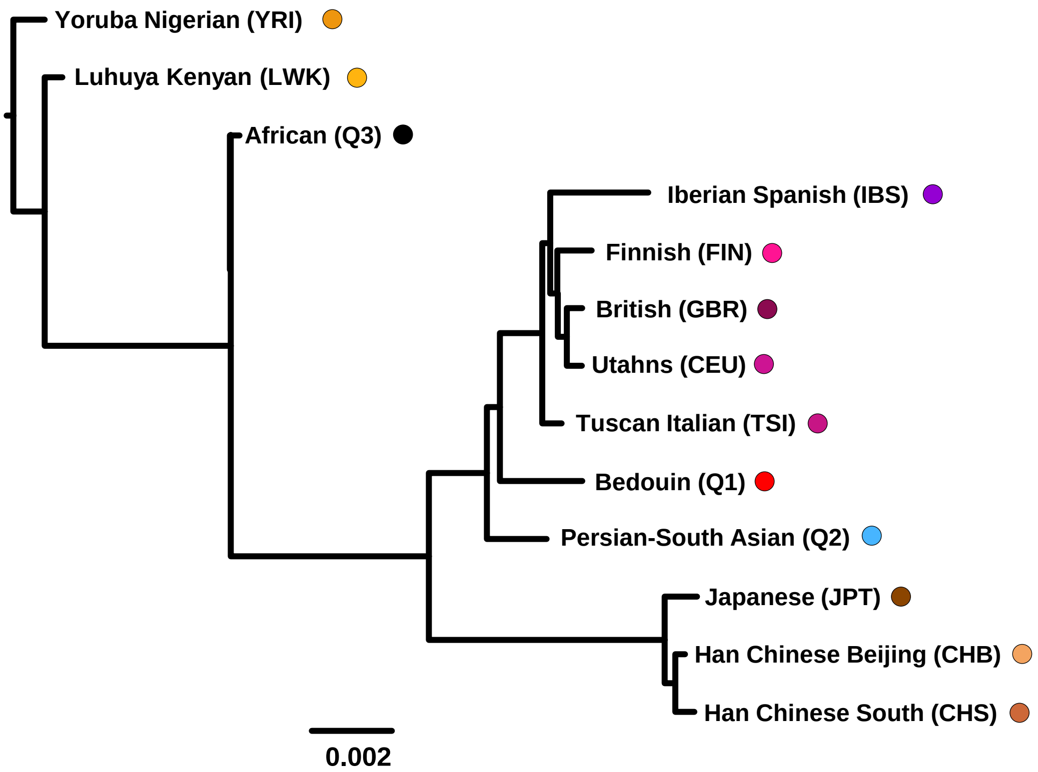
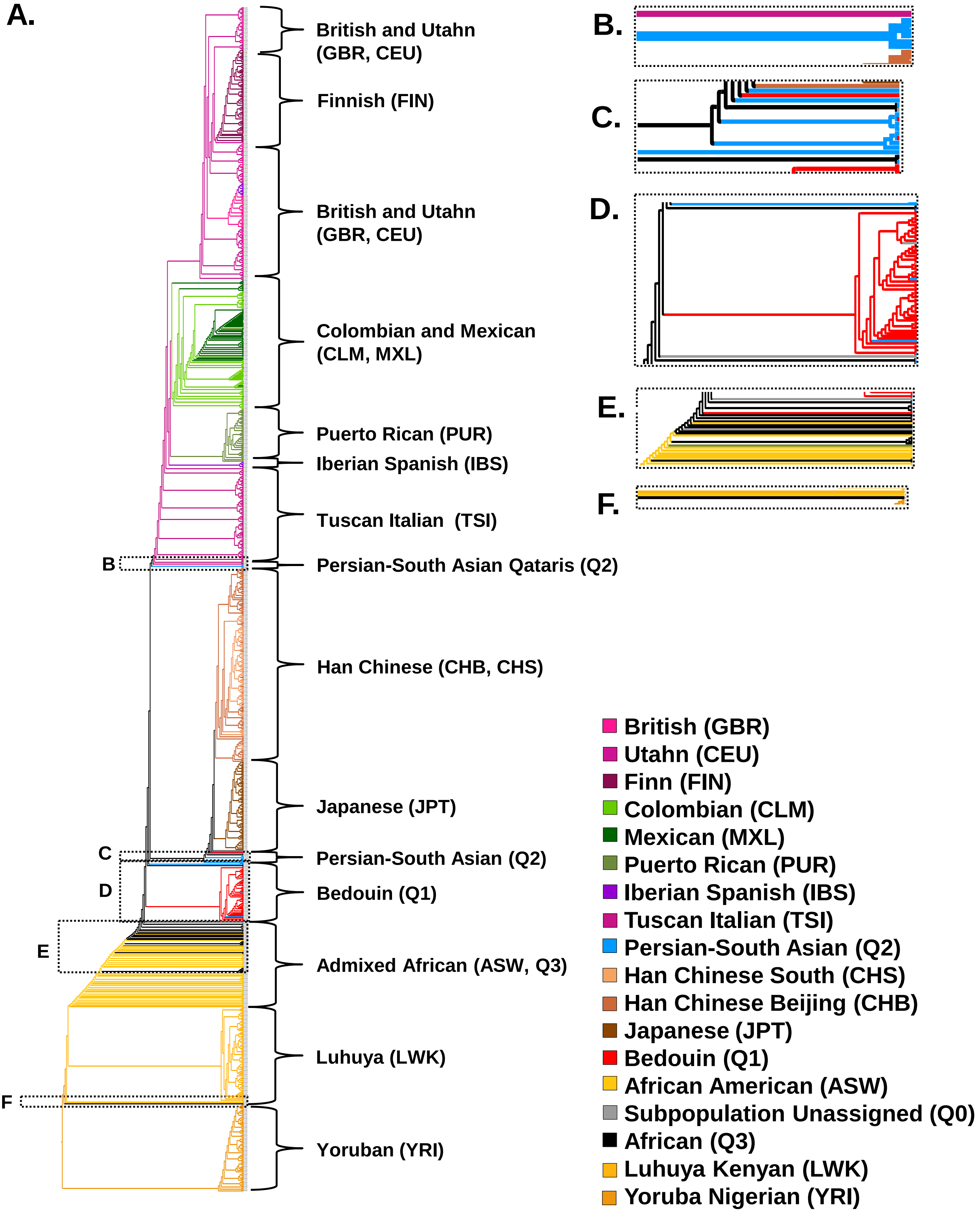

Yoruba Nigerian (YRI)
Luhuya Kenyan (LWK)
African (Q3)
Iberian Spanish (IBS)
Finnish (FIN)
British (GBR)
Utahns (CEU)
Tuscan Italian (TSI)
Bedouin (Q1)
Persian-South Asian (Q2)
Japanese (JPT)
Han Chinese Beijing (CHB)
Han Chinese South (CHS)
0.002

Figure 6


A.
British and Utahn (GBR, CEU)
Finnish (FIN)
British and Utahn (GBR, CEU)
Colombian and Mexican (CLM, MXL)
Puerto Rican (PUR)
Iberian Spanish (IBS)
Tuscan Italian (TSI)
B
Persian-South Asian Qataris (Q2)
Han Chinese (CHB, CHS)
Japanese (JPT)
C
Persian-South Asian (Q2)
D
Bedouin (Q1)
E
Admixed African (ASW, Q3)
Luhuya (LWK)
F
Yoruban (YRI)
B.
C.
D.
E.
F.
British (GBR)
Utahn (CEU)
Finn (FIN)
Colombian (CLM)
Mexican (MXL)
Puerto Rican (PUR)
Iberian Spanish (IBS)
Tuscan Italian (TSI)
Persian-South Asian (Q2)
Han Chinese South (CHS)
Han Chinese Beijing (CHB)
Japanese (JPT)
Bedouin (Q1)
African American (ASW)
Subpopulation Unassigned (Q0)
African (Q3)
Luhuya Kenyan (LWK)
Yoruba Nigerian (YRI)