**EVOLUTIONARY GENETICS**

**Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago**

Veronika Siska,1* Eppie Ruth Jones,1,2 Sungwon Jeon,3 Youngjuine Bhak,3 Hak-Min Kim,3 Yun Sung Cho,3 Hyunho Kim,4 Kyusang Lee,5 Elizaveta Veselovskaya,6 Tatiana Balueva,6 Marcos Gallego-Llorente,1 Michael Hofreiter,7 Daniel G. Bradley,1 Anders Eriksson,1 Ron Pinhasi,8† Jong Bhak,3,4* Andreea Manica1†

Ancient genomes have revolutionized our understanding of Holocene prehistory and, particularly, the Neolithic transition in western Eurasia. In contrast, East Asia has so far received little attention, despite representing a core region at which the Neolithic transition took place independently ~3 millennia after its onset in the Near East. We report genome-wide data from two hunter-gatherers from Devil’s Gate, an early Neolithic cave site (dated to ~7.7 thousand years ago) located in East Asia, on the border between Russia and Korea. Both of these individuals are genetically more similar to geographically close modern populations from the Amur Basin, all speaking Tungusic languages, and, in particular, to the Ulchi. The similarity to nearby modern populations and the low levels of additional genetic material in the Ulchi imply a high level of genetic continuity in this region during the Holocene, a pattern that markedly contrasts with that reported for Europe.

**INTRODUCTION**

Ancient genomes from western Asia have revealed a degree of genetic continuity between preagricultural hunter-gatherers and early farmers 12 to 8 thousand years ago (ka) (1, 2). In contrast, studies on southeast and central Europe indicate a major population replacement of Mesolithic hunter-gatherers by Neolithic farmers of a Near Eastern origin during the period 8.5 to 7 ka. This is then followed by a progressive “resurgence” of local hunter-gatherer lineages in some regions during the Middle/Late Neolithic and Early Neolithic periods and a major contribution from the Asian Steppe later, ~5.5 ka, coinciding with the advent of the Bronze Age (3–5). Compared to western Eurasia, for which hundreds of partial ancient genomes have already been sequenced, East Asia has been largely neglected by ancient DNA studies to date, with the exception of the Siberian Arctic belt, which has received attention in the context of the colonization of the Americas (6, 7). However, East Asia represents an extremely interesting region as the shift to reliance on agriculture appears to have taken a different course from that in western Eurasia. In the latter region, pottery, farming, and animal husbandry were closely associated. In contrast, Early Neolithic societies in the Russian Far East, Japan, and Korea started to manufacture and use pottery and basketry 10.5 to 15 ka, but domesticated crops and livestock arrived several millennia later (8, 9). Because of the current lack of ancient genomes from East Asia, we do not know the extent to which this gradual Neolithic transition, which happened independently from the one taking place in western Eurasia, reflected actual migrations, as found in Europe, or the cultural diffusion associated with population continuity.

**RESULTS**

**Samples, sequencing, and authenticity**

To fill this gap in our knowledge about the Neolithic in East Asia, we sequenced to low coverage the genomes of five early Neolithic burials (DevilsGate1, 0.059-fold coverage; DevilsGate2, 0.023-fold coverage; DevilsGate3, DevilsGate4, and DevilsGate5, <0.001-fold coverage) from a single occupational phase at Devil’s Gate (Chertovy Vorota) Cave in the Primorye Region, Russian Far East, close to the border with China and North Korea (see the Supplementary Materials). This site dates back to 9.4 to 7.2 ka, with the human remains dating to ~7.7 ka, and it includes some of the world’s earliest evidence of ancient textiles (10). The people inhabiting Devil’s Gate were hunter-fisher-gatherers with no evidence of farming; the fibers of wild plants were the main raw material for textile production (10). We focus our analysis on the two samples with the highest sequencing coverage, DevilsGate1 and DevilsGate2, both of which were female. The mitochondrial genome of the individual with higher coverage (DevilsGate1) could be assigned to haplogroup D4; this haplogroup is found in present-day populations (11). The people inhabiting Devil’s Gate were hunter-fisher-gatherers with no evidence of farming; the fibers of wild plants were the main raw material for textile production.

© 2017 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. Distributed under a Creative Commons Attribution License 4.0 (CC BY).
Relation to modern populations
We compared the individuals from Devil’s Gate to a large panel of modern-day Eurasians and to published ancient genomes (Fig. 1A) (4, 5, 14–17). On the basis of PCA (18) and an unsupervised clustering approach, ADMIXTURE (19), both individuals fall within the range of modern variability found in populations from the Amur Basin, the geographic region where Devil’s Gate is located (Fig. 1), and which is today inhabited by speakers from a single language family (Tungusic). This result contrasts with observations in western Eurasia, where, because of a number of major intervening migration waves, hunter-gatherers of a similar age fall outside modern genetic variation (3, 20). We further confirmed the affinity between Devil’s Gate and modern-day Amur Basin populations by using outgroup f3 statistics in the form f3(African; DevilsGate, X), which measures the amount of shared genetic drift between a Devil’s Gate individual and X, a modern or ancient population, since they diverged from an African outgroup. Modern populations that live in the same geographic region as Devil’s Gate have the highest genetic affinity to our ancient genomes (Fig. 2), with a progressive decline in affinity with increasing geographic distance (r2 = 0.756, Fst = 301, P < 0.001; Fig. 3), in agreement with neutral drift leading to a simple isolation-by-distance pattern. The Ulchi, traditionally fishermen who live geographically very close to Devil’s Gate and are the only Tungusic-speaking population from the Amur Basin sampled in Russia (all other Tungusic speakers in our panel are from China), are genetically the most similar population in our panel. Other populations that show high affinity to Devil’s Gate are the Oroqen and the Hezhen—both of whom, like the Ulchi, are Tungusic speakers from the Amur Basin—as well as modern Koreans and Japanese. Given their geographic distance from Devil’s Gate (Fig. 3), Amerindian populations are unusually genetically close to samples from this site, in agreement with their previously reported relationship to Siberian and other north Asian populations (7).

Relation to ancient genomes from Asia
No previously published ancient genome shows marked genetic affinity to Devil’s Gate: The top 50 populations in our outgroup f3 statistic were all modern, an expected result given that all other ancient genomes are either geographically or temporally very distant from Devil’s Gate. Among these ancient genomes, the closest to Devil’s Gate are those from Steppe populations dating from the Bronze Age onward and Mesolithic hunter-gatherers from Europe, but these genomes are no closer to Devil’s Gate genomes than to genomes of modern populations from the same regions (for example, Tuvinian, Kalmyk, Russian, or Finnish). The two ancient genomes geographically closest to Devil’s Gate, Ust’-Ishim (~45 ka) and Mal’ta (MA1, 24 ka), also do not show high genetic affinity, probably because they both date to a much earlier time period. Of the two, MA1 is genetically closer to Devil’s Gate, but it is equally as distant from Devil’s Gate as it is from all other East Asians (Figs. S14 to S16). A similar pattern is found for Ust’-Ishim, which is equally as distant to all Asians, including Devil’s Gate; this is consistent with its basal position in a genealogical tree (Figs. S17 to S19).

Continuity between Devil’s Gate and the Ulchi
Because Devil’s Gate falls within the range of modern human genetic variability in the Amur Basin in a number of analyses and shows a high genetic affinity to the Ulchi, we investigated the extent of genetic continuity in this region. To look for signals of additional genetic material in the Ulchi, we modeled them as a mixture of Devil’s Gate and other modern populations using admixture f3 statistics. Despite a large panel of possible modern sources, the Ulchi are best represented by Devil’s Gate alone without any further contribution (no admixture f3 gave a significant negative result; tables S3 and S4). Because admixture f3 can be affected by demographic events such as bottlenecks, we also tested whether Devil’s Gate formed a clade with the Ulchi using a D statistic in the form D(African outgroup, X; Ulchi, Devil’s Gate). A number of primarily modern populations worldwide gave significantly nonzero results (|z| > 2), which, together with the additional components for the Ulchi in the ADMIXTURE analysis, suggests that the continuity is not absolute. However, it should be noted that the higher error rates in the Devil’s Gate sequence resulting from DNA degradation and low coverage can also decrease the inferred level of continuity. To compare the level of continuity between the Ulchi and the inhabitants of Devil’s Gate to that between modern Europeans and European hunter-gatherers, we compared their ancestry proportions as inferred by ADMIXTURE. We found that the proportion of Devil’s Gate–related ancestry in the Ulchi was significantly higher than the local hunter-gatherer–related ancestry in any European population (P < 0.01 from 100 bootstrap replicates for the five European populations with the highest mean hunter-gatherer–related component).

These results suggest a relatively high degree of continuity in this region; the Ulchi are likely descendants of Devil’s Gate (or a population genetically very close to it), but the geographic and genetic connectivity among populations in the region means that this modern population also shows increased association with related modern populations. Compared to Europe, these results suggest a higher level of genetic continuity in northern East Asia over the last ~7.7 thousand years (ky), without any major population turnover since the early Neolithic.

Southern and northern genetic material in the Japanese and the Koreans
The close genetic affinity between Devil’s Gate and modern Japanese and Koreans, who live further south, is also of interest. It has been argued, based on both archaeological (21) and genetic analyses (22–25), that modern Japanese have a dual origin, descending from an admixture event between hunter-gatherers of the Jomon culture (16 to 3 ka) and migrants of the Yayoi culture (3 to 1.7 ka), who brought wet rice agriculture from the Yangtze estuary in southern China through Korea. The few ancient mtDNA samples available from Jomon sites on the northern Hokkaido island show an enrichment of particular haplotypes (N9b and M7a, with D1, D4, and G1 also detected) present in modern Japanese populations, particularly the Ainu and Ryukyuans, as well as southern Siberians (for example, Udegey and Ulchi) (26, 27). The mtDNA haplogroups of our samples from Devil’s Gate (D4 and M) are also present in Jomon samples, although they are not the most common ones (N9b and M7a). Recently, nuclear genetic data from two Jomon samples also confirmed the dual origin hypothesis and implied that the Jomon diverged before the diversification of present-day East Asians (28).

We investigated whether it was possible to recover the Northern and Southern genetic components by modeling modern Japanese as a mixture of all possible pairs of sources, including both modern Asian populations and Devil’s Gate, using admixture f3 statistics. The clearest signal was given by a combination of Devil’s Gate and modern-day populations from Taiwan, southern China, and Vietnam (Fig. 4), which could represent hunter-gatherer and agriculturalist components, respectively. However, it is important to note that these scores were just barely significant (−3 < z < −2) and that some modern pairs also gave
Fig. 1. Regional reference panel, PCA, and ADMIXTURE analysis. (A) Map of Asia showing the location of Devil’s Gate (black triangle) and of modern populations forming the regional panel of our analysis. (B) Plot of the first two principal components as defined by our regional panel of modern populations from East Asia and central Asia shown on (A), with the two samples from Devil’s Gate (black triangles) projected upon them (18). (C) ADMIXTURE analysis (19) performed on Devil’s Gate and our regional panel, for K = 5 (lowest cross-validation error) and K = 8 (appearance of Devil’s Gate–specific cluster).
**Fig. 2. Outgroup $f_3$ statistics.** Outgroup $f_3$ measuring shared drift between Devil’s Gate (black triangle shows sampling location) and modern populations with respect to an African outgroup (Khomani). (A) Map of the whole world. (B) Fifteen populations with the highest shared drift with Devil’s Gate, color-coded by regions as in Fig. 1. Error bars represent 1 SE.
negative scores, even if not reaching our significance threshold (z scores as low as −1.9; see the Supplementary Materials). The origin of Koreans has received less attention. Also, because of their location on the mainland, Koreans have likely experienced a greater degree of contact with neighboring populations throughout history. However, their genomes show similar characteristics to those of the Japanese on genome-wide SNP data (29) and have also been shown to harbor both northern and southern Asian mtDNA (30) and Y chromosomal haplogroups (30, 31). Unfortunately, our low coverage and small sample size from Devil’s Gate prevented a reliable estimate of admixture coefficients or use of linkage disequilibrium–based methods to investigate whether the components originated from secondary contact (admixture) or continuous differentiation and to date any admixture event that did occur.

**Phenotypes of interest**

The low coverage of our sample does not allow for direct observation of most SNPs linked to phenotypic traits of interest, but imputation based on modern-day populations can provide some information. We focused on the genome with highest coverage, DevilsGate1, using the same imputation approach that has previously been used to estimate genotype probabilities (GPs) for ancient European samples (5, 16, 17). DevilsGate1 likely had brown eyes (rs12913832 on HERC2; GP, 0.905) and, where it could be determined, had pigmentation-associated variants that are common in East Asia (see section S11) (32). She appears to have at least one copy of the derived mutation on the EDAR gene, encoding the Ectodysplasin A receptor (rs3827760; GP, 0.865), which gives increased odds of straight, thick hair (33), as well as shovel-shaped incisors (34). She almost certainly lacked the most common Eurasian mutation for lactose tolerance (rs4988235, LCT gene; GP > 0.999) (35) and was unlikely to have suffered from alcohol flush (rs671, ALDH2 gene; GP, 0.847) (36). Thus, at least with regard to those phenotypic traits for which the genetic basis is known, there also seems to have been some degree of phenotypic continuity in this region for the last 7.7 ky.

**DISCUSSION**

By analyzing genome-wide data from two early Neolithic East Asians from Devil’s Gate, in the Russian Far East, we could demonstrate a high level of genetic continuity in the region over at least the last 7700 years. The cold climatic conditions in this area, where modern populations still rely on a number of hunter-gatherer–fisher practices, likely provide an explanation for the apparent continuity and lack of major genetic turnover by exogenous farming populations, as has been documented in the case of southeast and central Europe. Thus, it seems plausible that the local hunter-gatherers progressively added food-producing practices to their original lifestyle. However, it is interesting to note that in Europe, even at very high latitudes, where similar subsistence practices were still important until very recent times, the Neolithic expansion left a significant genetic signature, albeit attenuated in modern populations, compared to the southern part of the continent. Our ancient genomes thus provide evidence for a qualitatively different population history during the Neolithic transition in East Asia compared to western Eurasia, suggesting stronger genetic continuity in the former region. These results encourage further study of the East Asian Neolithic, which would greatly benefit from genetic data from early agriculturalists (ideally, from areas near the origin of wet rice cultivation in southern East Asia), as well as higher-coverage hunter-gatherer samples from different regions to quantify population structure before intensive agriculture.
as described by Gamba et al. (17). Libraries were sequenced on either an Illumina MiSeq or HiSeq platform (for further details and sequencing statistics, see Supplementary Materials and Methods).

**Data processing and alignment.**

For single-end sequencing data, adapter sequences were trimmed from the ends of reads using cutadapt (39), allowing an overlap of only 1 base pair (bp) between the adapter and the read. For paired-end data, adapters were trimmed using leeHom (40). leeHom was run using the –ancientdna option, and paired-end reads that overlapped were merged. For paired-end reads that could not be overlapped, only data from read 1 were used in downstream analyses. Reads were aligned using the Burrows-Wheeler alignment tool (BWA) (41), with the seed region disabled, to the GRCh37 build of the human genome, with the mitochondrial sequence replaced by the revised Cambridge reference sequence (National Center for Biotechnology Information accession number NC_012920.1). Reads from different sequencing experiments were merged using Picard MergeSamFiles (http://picard.sourceforge.net/), and clonal reads were removed using SAMtools (42). A minimum read length of 30 bp was imposed, and for the higher-coverage (above 0.01X) samples, DevilsGate1 and DevilsGate2, indels were realigned using RealignerTargetCreator and IndelRealigner from the Genome Analysis Toolkit (GATK) (43). SAMtools (42) was used to filter out reads with a mapping quality of less than 30, and reads were rescaled using mapDamage 2.0 (44) to reduce the qualities of likely damaged bases, therefore lessening the effects of ancient DNA damage-associated errors on analysis (44). Average genomic depth of coverage was calculated using the genomcov function of bedtools (45).

**Authenticity of results and contamination estimates.**

Patterns of molecular damage and the length distribution of reads were assessed using all reads for DevilsGate3, DevilsGate4, and DevilsGate5. Because a portion of the reads from DevilsGate1 and DevilsGate2 was derived from 50-bp single-end sequencing, only reads sequenced with 150-bp paired-end sequencing were considered in the following analyses to avoid using truncated reads (library ID MOS5A.E1 for DevilsGate1 and MOS4A.E1 for DevilsGate2). MapDamage 2.0 (44) was used to assess patterns of molecular damage, which are typical of ancient DNA. We replicated all results using our data without applying MapDamage to avoid biases from dropping true mutations that look like damage on our low-coverage data. For further details, see Supplementary Materials and Methods.

We assessed the rate of mitochondrial contamination for our highest-coverage samples, DevilsGate1 and DevilsGate2. This was calculated by evaluating the percentage of nonconsensus bases at haplogroup-defining positions (haplogroup D4 for DevilsGate1 and M for DevilsGate2) using bases with quality ≥20. We also used schmutzi, a tool that uses a Bayesian maximum a posteriori algorithm (12), to estimate the mitochondrial contamination for DevilsGate1. Last, we replicated the analyses that yielded the most robust results (outgroup f3 scores and PCA) using only reads showing evidence of postmortem damage (PMD score of at least 3) (13). For further details, see Supplementary Materials and Methods.

**Statistical analysis**

**Mitochondrial haplogroup determination and molecular sex determination.**

Mitochondrial consensus sequences were generated for DevilsGate1 and DevilsGate2 using Analysis of Next Generation Sequencing Data (ANGSD) (46). Called positions were required to have a depth of coverage ≥3, and only bases with quality ≥20 were considered. The resulting

**Materials and methods**

**Experimental design**

**Sample preparation and sequencing.**

Molecular analyses were carried out in dedicated ancient DNA facilities at Trinity College Dublin, Ireland. Samples were prepared, and DNA was extracted using a silica column-based protocol following the methods of Gamba et al. (17), which were based on the study of Yang et al. (37). DNA extracted from both the first and second lysis buffers (17) was used for library preparation, which was carried out using a modified version of the protocol of Meyer and Kircher (38),
FASTA files were uploaded to HAPLOFIND (47) for haplogroup determination, with coverage calculated using GATK DepthOfCoverage (43). Mutations defining the assigned haplogroup were also manually checked. Molecular sex was assigned using the script described in the study of Skoglund et al. (48). For further details, see Supplementary Materials and Methods.

**SNP calling and merging with reference panel.**

To compare our sample to modern and ancient human genetic variation, we called SNPs using the hg19 reference FASTA file at positions overlapping with the Human Origins (HO) reference panel (591,356 positions) (49) using SAMtools 1.2 (42). Bases were required to have a minimum mapping quality of 30 and base quality of 20; all triallelic SNPs were discarded. Because of our low coverage does not provide sufficient information to infer diploid genotypes, a base was chosen with probability proportional to its depth of coverage. This allele was duplicated to form a homozygous diploid genotype, which was used to represent the individual at that SNP position (48). This method of SNP calling (referred to as the proportional method from now on) will artificially increase the appearance of drift on the lineage leading to calling (referred to as the proportional method from now on) will artificially increase the appearance of drift on the lineage leading to ancestry. Inferences about artificial ancestry do not influence our conclusions.

The resulting SNP data for DevilsGate1 and DevilsGate2 were then merged with a reference panel containing modern genomes from the HO panel and selected ancient genomes [this data set was described by Jones et al. (5)] as well as an additional 45 Korean genomes from the Personal Genome Project Korea (http://opengenome.net/) using PLINK 1.07 (50). Additional sample information is available in extended data table S1, including sample IDs, populations, and groupings used throughout the article. Last, a transversion-only version of all the above data was created by converting all T’s to C and G’s to A’s. This alternative data set was used to confirm that potential biases originating from ancient DNA damage do not influence our conclusions.

In the later analyses (outgroup and admixture $f_3$ statistics, PCA, and ADMIXTURE analysis), results using all mutations or only transversions were qualitatively similar, apart from increased noise in the transversion-only data due to the reduced information content. Thus, in the main text, we only report results using all mutations and the default calling method referred to as the proportional method (choosing a read uniformly at random from the reads covering any given position). We present results using the Khomani San as our African outgroup for outgroup $f_3$ and $D$ statistics, but other populations (the Yoruba, the Mbuti, and the Dinka) gave equivalent results. For further details, see Supplementary Materials and Methods.

**Population genetic analysis.**

PCA was performed with two different reference panels (Fig. 1 and extended data table S1), both subsets of the worldwide panel of contemporary and ancient individuals from the study of Jones et al. (5). The analysis was carried out using EIGENSOFT 6.0.1 smartpca (18), with the lsqproject and normalization options on, the outlier removal option off, and one SNP from each pair in linkage disequilibrium with $r^2 > 0.2$ removed. Ancient samples were projected onto the principal components defined by modern populations. For further details and results, see Supplementary Materials and Methods.

A clustering analysis was performed using ADMIXTURE version 1.23 (19). SNPs in linkage disequilibrium were thinned using PLINK 1.07 with parameters --indep-pairwise 200 25 0.5, resulting in a set of 334,359 SNPs for analysis (91,379 transversions). $K = 2$ to 20 clusters for the global panel and $K = 2$ to 10 clusters for the regional panel were explored using 10 independent runs, with fivefold cross-validation at each $K$ with different random seeds. The minimal cross-validation error was found at $K = 18$ for the global panel and $K = 5$ for the regional (East Asia and central Asia) panel, but the error already started plateauing around $K = 9$ for the global panel, suggesting little improvement. Furthermore, results for the regional panel were largely similar to those from the global panel for East Asian populations. To compare population-level frequencies of the inferred ancestry components, we simultaneously performed bootstrapping on SNPs and individuals for each population, with 100 bootstrap estimates, using inferred ancestral components from our ADMIXTURE runs on all SNPs and MapDamage-treated samples from Devil’s Gate. We projected each bootstrap replicate on the ancestral components by numerically maximizing the corresponding likelihood function, following the logic in the study of Allentoft et al. (16) and Sikora et al. (51). For further details and results, see Supplementary Materials and Methods and extended data figs. S1 to S8.

$D$ statistics (52) and $f_3$ statistics (49, 53) were used to formally assess the relationships between the samples using the qpDstat ($D$ statistics) and qp3PopTest ($f_3$ statistics) programs from the ADMIXTOOLS package (49). Significance was assessed by these programs using a block jackknife over 5-centimorgan chunks of the genome, and statistics were considered significant if their $z$ score was of magnitude greater than 2 or, for admixture $f_3$ scores, if they were smaller than $-2$. These correspond approximately to $P$ values of 0.046 and 0.023, respectively. For further details and results, see Supplementary Materials and Methods.

**Phenotypes of interest.**

We investigated phenotypes of interest in our highest-coverage sample, DevilsGate1, including some loci known to have been under selection in Eurasian populations. Because of the low quality of our samples, we used BEAGLE (54) to impute genotypes using a reference panel containing phased genomes from the 1000 Genomes Project (26 different populations). Following the study of Gamba et al. (17), GATK UnifiedGenotyper (43) was used to call genotype likelihoods at SNP sites in phase 3 of the 1000 Genomes Project. Equal likelihoods were set for positions with no spanning sequence data and positions where the observed genotype could be explained by deamination (17). We imputed at least 1 Mb upstream and downstream from the loci of interest using 10 iterations to estimate genotypes at markers for which we had no direct genotype. For further details and results, see Supplementary Materials and Methods and extended data table S9.

**SUPPLEMENTARY MATERIALS**

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/3/2/e1601877/DC1


Downloaded from http://advances.sciencemag.org on January 21, 2018
REFERENCES AND NOTES


74. B. L. Browning, S. R. Browning, Genotype imputation with millions of reference samples.
Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago

Sci Adv 3 (2), e1601877.
DOI: 10.1126/sciadv.1601877