SUPPLEMENTARY INFORMATION

Archaeological and chronological context

The human hair tuft (registration no. Qt 86 85/261: 12) was excavated in 1986 at Qeqertasussuk (Qt), a Saqqaq Culture site, located in the southernmost part of Disko Bay, West Greenland (Fig. 1a). Details of the site can be found in [1,2]. Due to its permafrost conditions, the Qt site enables excellent organic preservation – specifically, together with the Qajaa site, located 80 kilometers to the north, Qt is the only early Palaeo-Eskimo site to show preservation of organic materials such as wood, skin, hair and baleen [3,4,1]. In contrast to Qajaa, the Qt site holds no traces of later cultures (Dorset- or Thule), and Qt is the only Saqqaq site where human remains have been recovered (four limb bones and four tufts of hair).

At the Qt site, the sample was excavated from Section C, a 1 meter wide trench through the culture layers (different dumping areas) at the northern part of the Qt site (Fig. S1). The tuft was found embedded in the permafrost adjacent to a reindeer cranium in Square 85/261, Layer 15, approximately 40 cm below the present surface (Fig. S2). The hair tuft was found by a team member of Danish origin (Egon Geisler). The tuft was excavated over a period of several hours from the thawing culture layer 15 using a trowel, subsequently dried in the open air, and then stored in a labeled plastic bag. As with all other finds from Qeqertasussuk, the hair tuft was sent to Denmark for further analyses. Here, the hair was analyzed at the Copenhagen University Hospital, where it was determined to be of possible human origin based on morphological studies. Subsequently, it was stored at the University of Copenhagen and, later, at the National Museum of Denmark in a plastic bag awaiting further analyses.

According to the archaeologist in charge of the excavation, who also curates the hair sample (author BG), it has not been touched by anyone likely to be of Asian ancestry, and although some ethnic Greenlanders were present during the excavation, they did not handle the sample. The sample has, on the contrary, been handled by a number of ethnically Danish scientists. Given the low level of European contamination in the raw sequence reads (a maximum possible estimate of ~0.8%, see main text), despite this unprotected handling by Danes, it is unlikely that our findings of clearly Asian DNA should result from sample based contamination. The results confirm previous results showing that ancient hair can be effectively decontaminated (see main text).

Morphological characteristics of Saqqaq hair

A small subsample of hairs from the Saqqaq hair tuft used for DNA sequencing was subjected to a morphological examination. The hairs, to the unaided eye were black in colour and coarse to the touch. A number of hairs were selected for a more detailed microscopic examination. Briefly, the hairs were mounted in permanent mounting medium (XAM) on glass microscope slides for detailed transmitted light microscopy, which was performed using a Leica compound microscope (100-400x magnification range). Scale cast patterns and cross-sections were produced in accordance with Brunner and Coman [5]. Additionally, fibres were mounted onto aluminium stubs, gold-coated and imaged under high vacuum conditions using an FEI Quanta 400 scanning electron microscope.

Transmitted light microscopy revealed hairs that are moderately pigmented, with a continuous opaque medulla present in most of them. The hairs, despite their age, were comparable to scalp hairs of modern hominids in general appearance, as illustrated in

Figure 1b-d. The shaft diameters of the widest Saqqaq hairs measured approximately 100 μm. This is atypical of Caucasian scalp hairs that have an average shaft diameter of approximately 70 - 80 μm. However, this is consistent with the shaft diameters of Mongoloid scalp hairs are typically in the region of approximately 90-100 μm (measured on longitudinal axis of the mounted hairs).

Fungal tunnelling was apparent in the majority of the Saqqaq hairs; this phenomenon is apparent in both forensic and archaeologically derived hairs, and appears as fine cracks in the shaft. Fungal tunnelling has been associated with hairs that have been buried in, or in contact with, soil containing keratinophilic fungi [6]. Figures S3 and S4 illustrate examples of the fungal tunnelling apparent in many of the hair shafts which bore several of these 'tunnels'. Surface debris adhered to the majority of the Saqqaq hair shafts. It is likely that the debris found on the surface of the hair shafts are fungal growths responsible for the tunnelling of the hair shafts. Some of the growths had protuberances emanating from them. The image in Figure S5 illustrates an example of the type of debris present on many of the hair shafts. There are three stages in the fungal attack on hair, the first is the lifting of the cuticle, the second is the erosion of the cortex and the third is the tunnelling of the hair by specialised fungal hyphae [7].

Scale cast patterns were produced from Saqqaq hairs and modern hair. The results, illustrated in Figures S6 and S7, clearly show that the Saqqaq hair, stripped of its cuticle yields an almost smooth scale cast pattern of the exposed cortex which is in stark contrast with the scale cast pattern derived from a healthy modern hair. Probable remnants of a tiny portion of the cuticle ware apparent at the edge of one of the

Saqqaq hairs, which appears to show the cuticular scales comparable with human scale cast pattern (Fig. S8). These striae do not appear to be the result of fungal tunnelling, as none were seen in close succession to each other. Furthermore, these striae are very typical of scale patterns exhibited by hairs.

In reflected light (i.e. to the naked eye) the Saqqaq hairs were black in colour; however, under transmitted light the Saqqaq hairs were red/orange in colour. This phenomenon of ancient hairs having a red hue, implying the donor had red hair, is well documented [7]. Hair contains a mixture of black-brown eumelanin (high amount in brown/black hair) and red-yellow phaeomelanin (low amount in black/brown hair). Phaeomelanin is much more stable than eumelanin, thus, the phaeomelanin in the hair is better preserved over time than the eumelanin thus giving the hair an eventual reddish coloration.

Cross-sections of modern Caucasian scalp hairs and the Saqqaq hairs were produced (Fig. 1d). The modern hair were found to be oblong/oval in shape with a crisp, thin white outline which is the cuticle covering the cortex (pigment granules are visible in the modern scalp hair cuticle due to the hairs being very heavily pigmented). In comparison the Saqqaq hairs revealed a ‘fuzzy’ outline due to the lack of cuticle and appear to be more circular in cross-section than the modern hair (Fig. 1d). This circular cross-section is typical for Mongoloid scalp hairs than Caucasian scalp hairs, which supports the wide shaft diameter of many of the Saqqaq hairs.

Radiocarbon dating

Radiocarbon dating was undertaken at the Oxford Radiocarbon Accelerator Unit (ORAU, RLAHA, University of Oxford (Table S1). The dating process was enhanced by the high level of preservation of the materials, a result of the cold preservation conditions.

Samples of reindeer (*Rangifer tarandus*) bone were pre-treated for dating using the Oxford gelatinisation/ultrafiltration procedure [8], which is based on previous work by Brown et al. [9]. Coarsely ground bone powder (between 360-590 mg) was loaded into a pre-combusted 20 mL test tube and decalcified with 20ml 0.5M HCl: two times at RT for 1-2hrs, overnight (or until CO_2 ceased to be evolved). The collagen was rinsed three times with water and then with 20ml 0.1M NaOH for 30 mins at RT, rinsed again and then reacted with 20ml 0.5M HCl for 15 mins at RT. After a third water rinse, the crude collagen was gelatinised in pH3 solution at 75°C for 20 hours.

The gelatin solution was filtered using an 8mm polyethylene Eezi-filter™ and the insoluble residues discarded. The filtered gelatin was then pipetted into an ultra-filter (Vivaspin™ 15 30kD MWCO), centrifuged at 2500-3000 rpm until 0.5-1 mL of the >30 kD gelatin fraction remained and then freeze dried ready for combustion in a CHN analyser. The collagen yields ranged from 3.3 to over 11%, indicative of the good state of preservation of the bones, as expected.

22.8 mg of the human hair sample was rinsed in 80°C 1M HCl for one hour, 0.1 M NaOH at RT for one hour and finally with a repeat 1M HCl rinse. Between each reaction, the hair was rinsed to neutrality with distilled water. Finally, the pre-treated

hair was freeze-dried and weighed. The yield after pre-treatment suggests that the material is excellently preserved.

Sub-samples of the pre-treated hair or collagen were combusted and mass spectrometrically analysed using a Europa ANCA Roboprep interfaced to a Europa 20/20 MS operating under continuous-flow mode. The C:N atomic ratio for the hair sample was 3.7, wholly within the acceptable range for this sample type according to the work of others analysing the C:N ranges of hair keratin (3.0-3.8; [10]). The C:N ratios of the collagen samples were all within the range 2.9—3.5, indicating acceptable values. All other analytical data shown in Table S1 supports the observation that the bones were well preserved and the collagen extracted acceptable in its quality.

Graphite was prepared by reduction of CO_2 over an iron catalyst in an excess H_2 atmosphere at 560°C prior to AMS radiocarbon measurement [11]. The conventional radiocarbon ages BP are listed in Table S1.

The fragmented reindeer cranium (Qt 86 FC 85/261: 11), that was found in direct association with the human hair tuft (Qt 86 FC 85/261: 12) was AMS dated to 3628 ± 28 BP (OxA 18749). Layer 15, in which these specimens came from, has previously been conventionally dated using twig materials to 3680 – 3310 BP [1].

The human hair determination is from an individual whose diet included high levels of marine mammal protein (see below; Table S1), which is in concurrence with the analysis of faunal remains from the site, that show an extraordinarily heavy reliance

on marine mammals, especially harp and ringed seal (*Pagophilus groenlandicus* and *Pusa hispida*, respectively) [1]. This strongly implies the presence of a marine reservoir effect. This reservoir effect can be corrected by determining the local offset compared with the average world ocean value. Using modern collections of shellfish from the region, we calculated a local ΔR value of 140 ± 99 yr, (based on the ages recovered from modern shellfish from Disko Bay, Greenland, using the standard deviation as the square root of variance [12-16] (Marine Reservoir database, http://intcal.qub.ac.uk/marine/). This was used in the calibration of the result in Table S2. Note that this assumes that marine protein was the exclusive source of protein for the human. This is unlikely, but the stable isotope values do show that this was the overwhelming source of protein. The calibrated age ranges are shown in Figure 1f. We plot the calibrated human hair determination in two ways, one corrected for the reservoir effect, the other not.

Taken together, the dating of the hair tuft itself (corrected for the marine effect), supports an early Saqqaq context for the human remains and confirms earlier determinations. We estimate Layer 15 to span the period 4100 – 3900 cal. BP.

Stable light isotope analysis of the Saqqaq hair

Subsamples of the Saqqaq hair sample were prepared for stable light isotope analysis at the University of Oxford and the University of Bradford according to standard protocols (adapted from [10,17]). The hair fibres were soaked in a 2:1 (v/v) mixture of methanol and chloroform (for two hours at Oxford and overnight at Bradford), followed by sonication and rinsing three times in deionised water. Following the cleaning, the samples were dried completely and weighed into tin foils. They were

analysed for carbon and nitrogen isotopes at the Research Laboratory for Archaeology and the History of Art (RLAHA, University of Oxford) using a Carlo Erba 1108 elemental analyzer coupled to a SerCon Geo 20/20 mass spectrometer in continuous-flow mode, and, at the University of Bradford using a ThermoFinnigan Delta plus XP mass spectrometer. Data are reported using conventional delta notation taken relative to internationally recognised standards.

Stable light isotopes can be used as a proxy for diet, with carbon differentiating between the major photosynthetic pathways in plants as well as terrestrial versus marine food intake and nitrogen indicative of the extent of animal protein consumption. Results from Oxford and Bradford are plotted in Figure 1e, which also includes the following hair samples processed in Oxford for comparison purposes: Qt 87 FB 20/20 (another Saqqaq sample from similar context); and, 6 Thule individuals (Uummannaq, 16th-17th century). All data points are presented as averages of 2 or 3 replicate runs, except 1 Thule sample, which was processed once due to small sample size. The expected and measured carbon and nitrogen isotopic measurements of standards run with the samples are noted in Table S3. The atomic C/N ratios of the samples fell within the acceptable range for keratin of 3.0-3.8 [10].

The results unequivocally support the reliance of the Saqqaq on high trophic level marine food resources. This can be contrasted with published data for modern hunting and fishing Inuit Greenlanders from roughly the same geographic region (plotted as 'Modern Uummannaq Omnivores'), who supplement their diet high in local marine food (seal, whale, halibut) with foods imported from Denmark (bread, cereals, dairy products), as well as with modern Danish omnivores and vegans [18]. The data can

also be compared with other published carbon and nitrogen isotope data for Dorset remains, Thule-era burials from northwest Hudson Bay and proto-historic burials from Southampton Island [19], which together show varied exploitation of marine and terrestrial resources including fish, shellfish and avifauna, ringed seal (*Pusa hispida*), walrus (*Odobenus rosmarus*), bowhead whale (*Balaena mysticetus*), reindeer (*Rangifer tarandus*), musk ox (*Ovibos moschatus*), arctic fox (*Alopex lagopus*), hare (*Lepus*) and wolverine (*Gulo gulo*).

Sample preparation

To rid the hair of macroscopic contaminants such as small branches and soil, the hair was manually cleaned in certified DNA-free molecular biology grade water. This was followed by a wash in 1% hypochlorite (final concentration) solution, to remove DNA contaminants [20]. Lastly the hair was again washed several times in demineralised water to avoid damage from residual hypochlorite.

Three similarly sized cuts of hair (1.5-2g) were digested overnight at 55°C with agitation in 15ml of buffer containing, 10mM Tris pH 8, 10mM NaCl, 5mM $CaCl_2$, 2.5mM EDTA pH 8, 10% proteinase K solution, 40mM DTT and 2% SDS. An additional 5ml of buffer was added and the samples were incubated another 24hrs. The digested samples were phenol/phenol/chloroform extracted, and the aqueous phases were concentrated to 500μl using a 10kDa centricon column (Millipore, Billerica, MA). This was followed with 3 purifications on Qiaquick columns (Qiagen, Hilden, Germany), with the following modifications to the protocol, an additional wash with 500μl Salton buffer A (MP Biomedicals, Illkirch, France), 2 washes with Qiagen PE buffer, a 3 minute centrifugation at 13000 g to dry the filter, and finally a 5

minute incubation with 50µl Qiagen EB buffer. A second round of extractions were carried out with a few minor differences. Namely a single 50ml digestion was performed, and after the phenol/phenol/chloroform extraction concentration, a 30kDa filter was used to concentrate the digest to 2ml, which in turn was purified across 8 Qiaquick spin columns. Final elutions were in 30µl EB buffer.

Extracted DNA was built into Illumina index libraries, following manufactures protocol, but without DNA fragmentation. Volumes of DNA sample:adaptor were adjusted for short input DNA to approximate the suggested 1:10 ratio. Libraries were run on a 2% agarose gel, and gel purified using Qiagen gel extraction kit, following manufactures guidelines. Amplification of purified libraries were done using Phusion polymerase (Finnzymes, Espoo, Finland) with a final mixture of, 1X Phusion HF buffer, 0.2mM dNTP, 0.5µM Multiplexing PCR primer 1.0 (5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT), 0.01µM Multiplexing PCR primer 2.0 (5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT), 0.5µM PCR primer Index X (5'- CAAGCAGAAGACGGCATACGAGATN_6GTGACTGGAGTTC) (where X is one of 12 different indexes, and N_6 is the corresponding tag), 3% DMSO, 0.02U/µl Phusion polymerase, 3-5µl of template and water to 50µl final volume. Primers are part of Illumina's Multiplexing Sample Prep Oligo Kit. Cycling conditions were as stated in the Illumina protocol for a total of 18 cycles. PCR products were finally gel purified before sequencing.

Sequencing

The indexed DNA libraries were loaded on an Illumina Cluster Station to generate sequencing template clusters on flowcells by bridge amplification PCR. The workflow followed the manufacturer's instruction, briefly; template hybridisation, isothermal amplification, linearisation, blocking, denaturisation and hybridisation of the sequencing primers. The flowcells were loaded on to an Illumina Genome Analyzer II for sequencing. The read was sequenced first followed by 6bp-index. The images were transformed into intensity files that were then processed to produce DNA read sequences using the Illumina base-calling pipeline (Illumina Pipeline v1.4).

Estimates of DNA damage and quantitative PCR

Damage-driven deamination of cytosine yields uracil, while damage-driven deamination of 5'-methylcytosine yields thymine. Phusion polymerase is unable to replicate through uracil, thus is expected to stop whenever uracil is encountered in a sequence. In contrast, the Phusion will amplify through the thymine derivation of 5'-methylcytosine, although given both that 5'methylcytosine is relatively rare in the human genome, and that a damage event is required at this position, it can be hypothesised that this will contribute very little to the overall damage signal. The ability of Phusion to amplify damaged DNA (deaminated cytosine) was estimated by PCR amplifying samples with known high damage rates using both Phusion and HiFi polymerase. Samples were extracts of 2 reindeer bones (*Rangifer Tarandus*) dated 13,150±130 and 16,190±190 ^{14}C yr BP, and 1 woolly rhinoceros bone (*Coelodonta antiquitatis*) (not dated). The manufacturer's recommendations were followed for each enzyme, generic16S mtDNA mammalian fusion primers were used (Forward:

5'- GCCTCCCTCGCGCCATCAGN$_8$CGGTTGGGGTGACCTCGGA, reverse: 5'-GCCTTGCCAGCCCGCTCAGN$_8$GCTGTTATCCCTAGGGTAACT, N$_8$ represent different 8bp tags that were used for multiplexing purposes), and finally products were sequenced on a GS FLX (Roche, Germany). HiFi setup was as follows: 1X HiFi PCR buffer, 2mM $MgSO_4$, 0.4mM dNTP, 0.4μM primer (each), 1μl extract. Phusion setup was as follows: 1X HF buffer, 0.4mM dNTP, 0.4μM primer (each), 1 μl extract. PCR program were as follows, HiFi cycling parameters were: 94°C 4', (94°C 30", 54°C 30", 68°C 30") x 50 cycles, 72°C 7', Phusion parameters were: 98°C 30", (98°C 15", 54°C 30", 72°C 30") x 50 cycles, 72°C 7'. PCR products were gel purified and sequenced on a GS FLX following the manufacturer's protocol for amplicon sequencing; run size was LR25. A total of 63190 sequences with match to tag and primers were generated. 6 tags were used representing the 3 individuals with each enzyme. After trimming tag and primer, sequences were aligned using Muscle v3.7 [21], and for each position base composition was analysed to generate a 4x4 matrix with base calls and errors/damages (Table S4). When grouping into complementary pairs of errors, the data can be summarised in a graph showing the error rate for each individual (Fig. S9). When calibrated with this, the damage based errors drop from 1.6-6.1% to 0.2-0.5% ($p<2.2e-16$ for all, Fisher's exact test).

To estimate the copy number of DNA templates present in the extract both with, and without, cytosine deamination-based damage, we performed a number of quantitative PCRs (qPCRs). A single forward, and 4 reverse, primers, with products of lengths between 85-163bp (33-110bp excl. primers) were chosen (Forward: L16162 5'-CATAAATACTTGACCACCTGTAGTACATA, Reverse: H16196 5'-GATTGCTGTACTTGCTTGTAAGC, H16223 5'-

GCAGTTGATGTGTGATAGTTGAG, H16250 5'-GGTGAGGGGTGGCTTTGG, H16273 5'-GGGTGGGTAGGTTTGTTGGTATCC). Standards were generated for each primer pair from a purified PCR product (same setup as below) as follows. Firstly, the concentration of the PCR product was measured using a Nanodrop (Nanodrop Products, DE). Subsequently the copy number in the standard was estimated using, $c/M*N_A$, where c is concentration, M is molecular weight of PCR product and N_A is Avogadro's constant. qPCRs were performed with both Phusion and HiFi using the four primer pairs. HiFi setup was as follows: 1X HiFi PCR buffer, 2mM $MgSO_4$, 1µl Rox/SYBR mix, 0.2mM dNTP, 0.4µM primer (each), 0.02U/µl HiFi, 1µl template. Phusion setup was as follows: 1X HF buffer, 1µl Rox/SYBR mix, 0.2mM dNTP, 0.4µM primer (each), 0.02U/µl Phusion, 1µl template. qPCRs were run on MX3005 (Stratagene, La Jolla, CA) with the following cycling conditions, HiFi: 95C 10', (95°C 30", 60°C 30", 68°C 30") x 40 cycles; Phusion: 98°C 5', (98°C 15", 60°C 30", 72°C 30") x 40 cycles. The number of starting templates at each length was approximated for each of the two enzymes (Fig. S10). The results show approximately 1.8million copies/µl DNA extract of 85-bp mtDNA. Damage was estimated to be no more than 1%, by plotting a very simple model of damage on top of the HiFi amplified sample. The model estimated damage as H*(1-D)^L, where H is HiFi measurements, D is damage rate (here 1%) and L is length of PCR product exclusive primers.

Y-chromosome SNP verification

Primers were designed to target specific diagnostic Y-chromosome markers, in part to verify the shotgun sequencing and in part to gain information on uncovered regions. Selected markers (marker names as defined in [22]) were MEH2 (MEH2_F: 5'-

AAATTTTGAGTAAGCCATCACCCCA, MEH2_R: 5'-CCACATGTAATTGCAAAAACTGCATTG), M346 (M346_F: 5'-GGGAAAGGCAGCCAAGAGGACA, M346_R: 5'-TCCACTCACTCTGCCTACCTGA), M3 (M3_F: 5'-GGGCATCTTTCATTTTAGGTACCAGC, M3_R: 5'-TCTGCTGCCAGGGCTTTCAA), M216 (M216_F: 5'-CTCAACCAGTTTTTATGAAGCTAG, M216_R: 5'-CTGAATTCTGACACTGCTAGTTAT), and M242 (M242_F: 5'-TAGTATCTTGAACTTATATATG, M242_R: 5'-CACGTTAAGACCAATGC). Amplification was performed using HiFi. HiFi setup: 1X HiFi PCR buffer, 2mM $MgSO_4$, 1μl Rox/SYBR mix, 0.2mM dNTP, 0.4μM primer (each), 0.02U/μl HiFi, 1μl template. PCR program were as follows, HiFi: 94°C 4', (94°C 40", 52°C 35", 68°C 40") x 48 cycles, 72°C 10'. PCR products were purified using Qiaquick columns (Qiagen, Hilden, Germany), and cloned using the TOPO TA cloning kit (Invitrogen, Carlsbad, CA) following the manufacturer's guidelines. Cloned product were amplified and sequenced on an ABI3130xl (Applied Biosystems, Foster City, CA), following the manufacturer's protocol.

Genome assembly

As described in the main text, short index sequences of length 6 nucleotides were used to distinguish DNA from the PCR amplification from contaminants. Only reads with the correct index were used in the downstream analyses. We used a stringent criterion of 100% match in the index sequences to lower the risk of contamination. 93.17% of all reads had the correct index and were used in the mapping.

The correctly indexed reads were mapped to the human genome chromosomes 1-22, X, Y, and mtDNA (hg18, NCBI build 36.1, http://hgdownload.cse.ucsc.edu/goldenPath/hg18/ chromosomes/). Due to the fragmentation of the ancient DNA, most reads contain part of the 3' primer/adapter sequence. It is difficult to correctly identify a short primer segment in the end of a read where the sequencing quality is generally lower. Therefore we used a program (*Sesam*) that was developed to perform the mapping such that the most likely split between a genomic match and the primer was found. The quality scores are ascii "phred scores" and translated to error probabilities $p=10^{(i-d)/10}$, where i is the ascii value and d is the base of the conversion (33 or 64). After translating the quality scores to probabilities, the read sequence is turned into a position specific scoring matrix with a score of $\log2((1-p)/q(a))$ for the nucleotide a, which is called and $\log2(p/3q(b))$ for another nucleotide b, where $q(a)$ is the frequency of base a in the human reference genome. The search is done using an enhanced suffix array index [23] of the genome. In this search only sequences scoring more than 50% of the maximal score for the sequence are reported, and to speed up the search, it is also required that the sum of scores of the first 20 nucleotides is such that we only allow a single mismatch with the lowest quality score. The probability of each match is approximated by $p_i = 2^{s(i)}/\sum_j 2^{s(j)}$, where $s(j)$ is the score of match j of the read. Only if this probability is larger than 0.9, is the match reported as a unique match (this is more stringent than most other methods that rely only on the number of mismatches). Reads with more than 100 matches above the threshold are also ignored.

Using Sesam, we uniquely mapped 49.20% of all the correctly indexed reads. The average length of the mapped reads after primer removal is 55.27 nucleotides, and the

GC-content is 0.53. The mapped reads were sorted by chromosome and position for further analysis. To address the problem of clonal expansion from the PCR, reads of the same length from the same PCR library that map to the same position and strand in a chromosome were considered clones. Instead of removing them completely, a collapsed representative sequence was generated from the 'clones', that took into account the nucleotides and quality scores at each position. The collapsed clone was made by creating a Lx4 scoring matrix, where L is the read length of the clones. For each position 1 to L, the sum of the probabilities (in log-space) of observing each of the 4 nucleotides was calculated from the read nucleotides and quality scores seen in all the clones at that position. The collapsed clonal representative was defined by choosing the most likely nucleotide in the matrix at each position, and the corresponding qualities were calculated from the summed log-transformed probabilities. Because of this, the new error rates can be much smaller than 10^{-4}, which is the minimum for the Illumina quality scores. We extend the quality range to 10^{-5}. The reason for not including even lower error rates is because errors might occur early in preparative process (e.g. PCR amplification), and these would get too much weight in the SNP calling and bias the result.

Of the mapped reads, 51.4% were retained unaltered, and the remaining clonal reads were collapsed into approximately 172 million so-called collapsed reads (corresponding to 17.23% of the total number of mapped reads after clonal removal). After this step, we effectively mapped 28.47% of all out initial reads uniquely to the human genome. Since the nucleotides in the collapsed reads are corrected based on all the clones in a set, we performed primer removal on the collapsed sequences to avoid including high quality primer sequences in the genotyping, which would significantly

bias the SNP calls. Even if Sesam could not locate the transition from genomic sequence to primer, the enhanced quality read helped us solve that problem. This was taken care of by performing ungapped alignment of the 3' end of the reads to the 5' end of the reverse-complemented primer sequences. The best alignment (if any) was used to truncate the sequence. To accommodate sequencing errors we allowed for a number of mismatches of up to 1/3 of the alignment length.

After the clonal analysis, a coverage map was calculated for each chromosome. This step generates a file with a line for each position covered in the chromosome. Each line contains information on the nucleotide in the reference genome, the sequence depth and a list of the pairs of nucleotides and corresponding qualities that cover the position. To avoid false SNP calls due to insertion and deletion events, which would accumulate in the 3’ end of the read, we ignore the last 7 nucleotides in the low quality, 3’ end of each read in this step. The 7 nucleotide trimming is based on an analysis of error frequencies that showed an increase in the overall error rate in the last 7 positions of the reads. We cover 79.31% of the complete human genome with an average depth of 16x across the whole genome (if limiting ourselves to only the positions covered, we have on average depth of 20x) (Table S5). We calculated the percentage of sequences of length 50 or 60 that occur more than once in the complete genome. For length 50, 15.2% of the genome is repetitive and for length 60, 13.3% is repetitive, so we estimate the theoretically highest coverage to be 85-87% with an average read length of 55.

Genotyping

The limited coverage and the special sources of errors from ancient DNA sequencing pose special challenges to genotyping. We therefore developed a probabilistic genotyping method, called *SNPest* (**S**ingle **N**ucleotide **P**olymorphism **est**imation), which takes both quality-scores and alternative sources of read errors explicitly into account. This allows sensitive genotyping, while avoiding systematic wrong calls due to read errors. The method is based on a generative probabilistic model of the sampling and sequencing of read nucleotides from the genotype at a given position in the ancient diploid genome. The model defines a probability distribution over observed and unobserved random variables. The observed random variables correspond to the light intensities from the sequencing reactions (I) and to the human reference nucleotide (H). The unobserved variables correspond to the genotype at the given position in the true ancient genome (G), the nucleotides in the ancient DNA fragments corresponding to our reads (A), and the nucleotides present in the mapped reads (R). The combined probability distribution for a single position can be written as:

$$P(H,G,A,R,I) = P(H)P(G \mid H)\prod_{i=1}^{n} P(A_i \mid G)P(R_i \mid A_i)P(I_i \mid R_i)$$

Where n denotes the number of observed reads at a given position and the index variable i, used with A, G, and I, denotes a specific read.

As can be seen, the complete probability factorises into a number of conditional probability terms, of which each has a specific interpretation in the sampling of reads from the original genotype at a given position. P(H) defines a uniform prior over the

human reference bases. P(G|H) defines the conditional distribution of the ten possible genotypes (AA, AC, ... , TT), given the reference nucleotide. This distribution was derived from the genotype counts of the Yanhuang genome [24]. $P(A_i|G)$ defines a conditional distribution over nucleotides in the original (ancient) DNA fragment that gave rise to the i'th present day read. $P(R_i|A_i)$ defines a conditional distribution over nucleotides in the i'th present day read given the original ancient nucleotide. This term models any errors in the present day reads due to PCR amplification, failure to remove primers, wrongly mapped reads, etc. This error rate was estimated to 0.33% based on all positions larger than 20, since Sesam treats position 1-20 specially, from all reads that map to chromosome 1. Finally, $P(I_i|R_i)$ defines the probability of observing the observed read intensities, given the read nucleotide. Although we do not know the form of $P(I_i|R_i)$, we note that it is proportional to $P(R_i|I_i)$, which is given in the form of the quality scores calculated in the Illumina pipeline (v1.4). These proportionality factors cancel out in the calculation of the posterior probability given below.

Based on a marginalisation of the complete probability distribution, where we sum out all the unobserved nuisance parameters (A & R), we calculate the posterior probability of the genotypes P(G|I, H). The genotype with the highest posterior probability is our prediction at the given position.

For the two sex chromosomes and the mtDNA, a haploid genotyping model was developed. The only difference to the above defined diploid model is that the unobserved genotype of the ancient genome (G) is now defined over the four single nucleotides, instead of the ten possible dinucleotide genotypes of the diploid model.

This affects P(G|H), which is now estimated from only the X chromosome of the Yanhuang genome, and $P(A_i|G)$, which is now simply specified by the identity matrix.

Applying SNPest to all chromosomes we find a total of 2,209,739 differences from the reference. Comparing these to dbSNP (v130) we see an overlap of 82.4% with positions annotated as single nucleotide polymorphisms. When comparing to all types of polymorphisms in dbSNP, we see an overlap of 86.2%. Overlap to other known genomes is shown in Figure S11, with the biggest overlap to the Korean and Chinese genomes (39.0% and 38.2% respectively).

A high confidence set of genotype calls (Table S5) is defined as follows: 1) Since repeat regions will have a higher rate of mapping errors, we exclude these. (The repeat annotations were defined by RepeatMasker and downloaded from the UCSC Genome Browser, hg18 [25]. 2) We also exclude regions with depth lower than 10x or higher than 50x (to avoid poorly covered regions and un-annotated repeats, respectively), 3) Only genotype calls with a posterior probability $P \geq 0.9999$ are included. 4) For SNPs (defined relative to the reference genome), we require the distance to the nearest neighbouring SNP to be at least 5 nucleotides. This is done to avoid false SNPs due to indels, which tend to cluster together. The high-confidence SNP set contains 353,151 positions with an overlap to single nucleotide polymorphisms in dbSNP of 91.0% and an overlap with all polymorphisms in dbSNP of 93.2%.

Due to the high copy number of mitochondrial genomes in the cell, we handle this slightly different than the nuclear chromosomes. Genotyping is carried out using the

same model as used for the sex chromosomes, only we do not exclude positions with more than 50x depth (average depth on chrM is 3802X).

For positions with more than 200x depth, we randomly sample 200 reads at that position for the genotyping. We tested that this has minimal influence on the genotype, while the running time per chromosome was changed from days to hours (data not shown). By genotyping chromosome 15 and 21 using either approach, differences between the two approaches were seen in only 11 of 69,407,765 positions (0.000016%) or 58 of 29,842,561 positions (0.000194%) respectively.

Analyses of heterozygote calls made on chromosome X using the diploid genotyping model

Since the Saqqaq genome is from a male individual, only single copies of both sex chromosomes are present. Genotyping of these chromosomes is therefore done with an especially designed haploid-genotyping model. However, since several factors, such as read and mapping errors as well as contamination from modern humans (discussed below), can give rise to sites with several nucleotides present, it is of interest to identify these sites and analyse their cause. We do this by genotyping the sex chromosomes with the diploid genotyping model, and analysing sites that are called as heterozygous. Since chromosome Y is known to be notoriously difficult to assemble and correctly genotype [26], we focus this analysis on chromosome X only.

The most common cause for observing several nucleotides at a given site in a haploid chromosome is read errors. At low quality-scores, these are very frequent. However, our genotyping model (*SNPest*) takes this into account, and sites called as

heterozygous with a high posterior probability are not well-explained by read errors. Errors introduced in the polymerase chain reaction (PCR) are related to these, but will be rarer, albeit they can be assigned high quality scores.

A second big source of nucleotide variation at a site is mapping errors. These are especially common in regions that align well (e.g. are homologous) to many other regions around the genome. This is the case for repeats, such as SINEs and LINEs, which will inevitably have a higher than average rate of mapping errors. In addition, the genome is also rich in other types of repetitive structures, such as recently duplicated regions, paralogous genes, and pseudo-genes, which will also have elevated rates of mapping errors. A further challenge in this respect, is that 7.3% of the human genome is still not assembled and therefore represented as gaps in the reference. Reads from repetitive regions within assembly gaps that have homologs elsewhere are likely to be mis-aligned. This effect can be especially large for repeat families that are mostly found in assembly gaps with only a smaller fraction outside.

A third source of nucleotide variation is alignment errors due to indels or structural difference between the reference genome and the genome being sequenced. A read from a region with a short indel, may map to the correct location, but since most genomic mapping methods, *Sesam* included, do not model gaps, part of the read will be wrongly aligned and give rise to a run of discrepancies. Similarly, structural variation, such as copy-number variation, can give rise to such alignment errors.

The diploid model uses the reference genome as part of its input data. This has the effect that heterozygote calls, involving the reference genome nucleotide, are much

more probable than homozygote calls away from the reference. Especially at low coverage, this will create a bias toward heterozygote calls, which can be wrongly interpreted as evidence for nucleotide variation at a given site. Since the sex chromosomes are present in single copies, their overall coverage is about half that of the nuclear chromosomes and this effect will be more pronounced.

Finally, contamination from modern humans may also contribute to sequence variation at a given site. Specifically, this may happen at polymorphic sites where the Saqqaq genome and the modern human differ. In this respect, chromosome X is more informative than chromosome Y, since contamination from both men and women will contribute chromosome X sequence and hence nucleotide variation at individual polymorphic sites. In contrast to the above described mapping errors, contamination will not be localised or biased toward certain regions, but should be uniformly represented along the genome.

Even in the absence of contamination, it is inevitable to have some high-confidence heterozygote calls when genotyping a haploid chromosome with a diploid-genotyping model. This is exemplified by the sequencing of the Yanhuang genome (also male), which resulted in 8% heterozygous calls on chromosome X. Generally, contamination is not considered a problem in modern human genomes.

When genotyping chromosome X of the Saqqaq genome with the diploid model, we found a total of 1783 high confidence SNPs. Of these, 76 are heterozogyte calls. Below we analyse the most likely cause for these and give some illustrative examples.

We find that 22 (29%) of the 76 heterozygote SNPs are located within 10 bases of indels and structural variation known to segregate in the human population (as defined in dbSNP version 130). Since these indel-and-structural-variation-proximal regions only compose 4.2% of chromosome X, we conclude that they are most likely the cause for these SNP calls.

We then analysed the remaining SNPs with respect to the repetitive structure of the genome. 18 (33%) of the 54 remaining high confidence heterozygote SNPs (24% of all) were found in regions with extensive sequence homology to at least ten other regions in the genome, as defined by the self-chain track of the UCSC Genome Browser [25]. Overall, only 2.2% of chromosome X shows this level of homology to other regions. Again, we conclude that these heterozygote calls are most likely explained by the repetitive nature of these regions.

Of the remaining 36 heterozogyte SNPs, 15 (42%; 20% of all) were found within 50 bases of a neighboring SNP call (with respect to the full set of SNP calls). In comparison, only 99 (5.8%) of the 1707 high confidence homozygote calls were within 50 bases of a neighboring SNP. This shows that a large fraction of the remaining heterozygote calls are in regions with an unusual high SNP density, suggestive of additional un-annotated structural variation or repetitive regions with high incidence of mapping errors.

Of the remaining 21 heterozogyte SNPs, 13 (62%; 17% of all) have some homologous regions around the genome (albeit less than ten). For comparison, in all of chromosome X, outside of repeats, only 9.0% show homology to other regions

(again based on the self-chain track of the UCSC Genome Browser). Some of these are in genes that are part of large gene families, with several homologous regions around the genome. Others have numerous homologous regions within chromosome X and appear to part of locally duplicated regions. Finally, some appear to be homologous to only a single other region in the genome. One representative example of this is a CT heterozygote SNP called at position 71,850,461, where the reference genome has a C (see Figure S12). This region is almost identical to a region on chromosome 1, probably due to a recent pseudo-gene insertion, where the corresponding position is a T. Wrongly mapped reads from this region thus explains the observed nucleotide variation at the site.

Finally, six of the eight heterozygote SNPs with no annotated sequence homology to other regions of the genome, are located in short gaps (1-4 nucleotides) between annotated repeats of the same type or regions annotated to be of low complexity. (Note that SNPs overlapping repeats are not included in the high confidence set and that annotated repeats are also not included in the self-chains used to identify repetitive regions above). With little doubt, these heterozygote calls are thus also due to the repetitive structure of the genome. This leaves two heterozygote calls with no clear indication of cause. They could potentially be explained by contamination. However, they could also be explained by the presence of un-annotated repeats, homologous regions located in assembly gaps, structural variation, etc., in line with the other heterozygote calls categorised above.

Since the majority (74 of 76) of the high confidence heterozygote SNP calls can be referred to alignment and mapping errors, we conclude that their presence does not indicate contamination of the Saqqaq genome with modern day human DNA.

The full set of SNP calls include calls based on low coverage and calls made in repeats. For both of these reasons we expect to see a high rate of (low-confidence) heterozygote SNP calls in this set. There are 48,057 SNPs in the full set and 18,102 (38%) are heterozygotes, confirming this expectation. The majority of these heterozygote SNPs are located in repeats (96%), which is much higher than the fraction of homozygote calls located in repeats (62%), again showing that mapping errors are more common in repeats and that the repetitive structure of the genome is the main source for the heterozygote SNP calls in the haploid X chromosome.

Genotyping the mitochondrial genome with the diploid model

We also genotype the mitochondrial genome using the diploid genotyping model. In total 41 SNPs are called of which one is a heterozygote. However, it is located only four nucleotides from another (homozygote) SNP call and can be attributed to an indel in the Saqqaq mitochondrial genome relative to the reference. It therefore does not pass the quality filters for the high confidence set, which contain 38 homozygote SNP calls (one call overlaps a repeat and the last call also has a distance of four nucleotides to the nearest neighbour). Again, we find no evidence of contamination.

Contamination estimated by private European allele frequency

To estimate European contamination in the Saqqaq data, we used the HGDP genotype data to define a set of alleles that were private to Europeans. The Europeans were

defined as 135 individ- uals from France, Italy, Italy (Bergamo), Orkney Islands, and Russian Caucasus.To represent a population closely related to the Saqqaq individual, we chose individuals from Siberia, Cambodia, China and Japan. Only 70 individuals from China were chosen in order to match the same number of individuals as in the European sample. All SNPs with missing data or missing annotation were removed. We denote a set of 28,403 sites as having private European alleles, because these alleles were observed only in the European sample. Of these sites, we cover 25,597 positions with a total of 392,470 read nucleotides from the Saqqaq individual. Of these, we observe the European allele in 1,286 cases (0.33%). The error rate for the reads was estimated based on the neighbouring positions to the private alleles. We extracted 5 positions on both sides (if covered) yielding a total of 2,700,147 read nucleotides. For each position, the fraction of nucleotides that deviate from the reference is calculated, giving a total error estimate of 0.35%. Since this error can be in three directions relative to the reference, the probability of seeing the particular private allele by chance is 0.12% on average.

In order to quantify the contamination rate in the Saqqaq individual we use the following model for the probability of observing a private allele *O* given an error rate *F*:

$$\begin{aligned} p(O|F) &= p(O|F,EU)p(EU) + p(O|F,AS)p(AS) \\ &= p(O|EU,true)p(EU)(1-F) + p(O|EU,error)p(EU)F \\ & \quad p(O|AS,true)p(AS)(1-F) + p(O|AS,error)p(AS)F, \end{aligned}$$

where *p(EU)* is the probability of an allele being European (contamination rate), *p(AS)* the probability of an allele being Asian (1-contamination rate), *true* meaning that no error on the read has occurred, and *error* meaning that an error occurred on the read (only errors between the two possible alleles are included).

Due to the SNP ascertainment, the allele frequencies in the population are hard to estimate without bias. Therefore, we estimate the expected number of private alleles using subsampling. In the procedure, we select the sites with private alleles after excluding a single individual and then obtain an estimate of the number of private alleles by the number of observed private alleles in the excluded individual. We perform this for every individual, and our final estimate is the average over all individuals. Thus an unbiased estimate of the fraction of private alleles based on the observed number of private alleles in the whole sample *Oj* for site *j* can be written as:

$$p(O|EU,true) = \frac{\sum_{j=1}^{M}\left(O_j\left(1 - I_{unique_j}\right)\right)}{NM - \sum_{j=1}^{M}\left(2I_{unique_j}\right)},$$

where I_{unique} is the indicator function which returns 1 if the private allele is unique to one individual, and zero otherwise.

Because we denote alleles as private if they are only observed in one population of our finite sample, some of the alleles are not really private in the full population. Therefore, we still expect so observe some alleles in one population that are denoted as private for the other population in our sample even without contamination. We also estimate this fraction, *p(O|AS,true)*, using subsampling where we denote the private alleles after excluding a single individual. We then estimate the fraction of sites where we expect to find a private allele from the observed number in the sample that was excluded. This gives us an estimate for each individual. Because some of the Asian individuals in our sample show a significant portion of admixture, the estimate will be inflated. Therefore we use the median of the estimate for the individuals instead of the mean because this will be more robust in the presence of outliers.

We estimate standard error using a blocked jackknife procedure. The alleles across the genome will be correlated due to linkage disequilibrium. Therefore we bin the

genome in 5Mb large regions and estimate the standard errors using bootstrap where for each bootstrap one of the blocks is removed from the analysis. We chose a bin size of 5Mb as the LD does not extend this far and because this is what is recommended in Reich et al. [27].

Using the error estimate of 0.12% in the model described above, we get an expected contamination of 0.85% ± 0.21% (Table S6). To get around possible issues with noise, we also introduced a quality cutoff on the reads where we disregard all nucleotides with an error rate of 0.33% or worse. This only changes the results slightly. Of the 344,819 read nucleotides above the cutoff, we observe the European allele in 888 cases. The error estimate is based on 2,329,850 read nucleotides yielding a total error rate of 0.0886%, or 0.0295% when correcting for the three possible errors. Using this approach, the estimated level of contamination is 1.1% +/- 0.21% comparable to the result above. It is also important to note that even a level of contamination around 1% at the sequence level results in a much lower level of contamination at the genotyping level, and therefore this potential source of error will not affect our SNP calls. This is also seen on the genotype level on the high confidence SNPs as seen in Table S7.

It should be noted that this contamination estimate ignores many factors that would influence the result: For instance, some of the individuals in our Asian sample show signs of having European admixture (see plot in main text), which will bias our estimate of the contamination. Also, using this Asian sample as a proxy for the ancestral Saqqaq population and ignoring errors in the SNP chip data will affect the results. The error rate is calculated from the total number of discrepancies from the

reference and does not take transitions/transversion into account. Since both SNPs and read errors are mainly due to transitions, we underestimate the background chance of observing the European allele by ignoring this in the error rate. We also treat all positions identically in the error estimate, although the error rate will be higher in repeats. The standard errors here are based on a fixed error rate and a fixed estimate of the fraction of private European alleles in the Saqqaq individual. Thus we do not take into account the uncertainty in these estimates, which makes the standard errors a little to narrow.

Y chromosome based demography

We found 4,024 positions at which Saqqaq sequence differed from the reference sequence, 243 of these in the high confidence subset. For majority of these positions there is no information available in Y-chromosome consortium SNP database and these are likely to include primarily private or low frequency mutations, and changes due to DNA damage and/or sequencing errors. Out of the 590 informative Y chromosome SNPs (as defined by the Y chromosome consortium), which have been positioned on the phylogenetic tree of Y chromosome haplogroups so far [22] there are 398 which come from non-repetitive regions. For 219 of these our quality criteria were met and the nucleotide status was ascertained for the Saqqaq sample (Table S8). In 211 positions the Saqqaq sample shares the genotype with the reference, which belongs to haplogroup R1b for most of the Y chromosome sequence. At 8 positions the genotype was different, including 6 SNPs that define haplogroup R. Combined with the derived status at 8 haplogroup P defining positions, plus the ancestral status that Saqqaq and the reference share at the remaining 203 loci, this evidence shows

that the Saqqaq sequence can be unambiguously assigned to a single source of Y chromosome haplogroup P (Fig. 3d).

Among the extant human populations, almost all haplogroup P chromosomes belong either to haplogroup R or Q, the latter being also one of the two most common Y chromosome haplogroups among Northeast Siberians [28], Aleut Islanders [29], North American Athapaskan speakers [30] and among modern Greenland Inuits [31]. According to the current YCC nomenclature the Northeast Siberian and North American haplogroup Q lineages fall further into Q1(a)* and Q1a3 subclades of this haplogroup [22]. However, none of the basal haplogroup Q defining sites were mapped while the Saqqaq sequence appeared to have the ancestral state at markers M346 and M3 that are derived in approximately one third of haplogroup Q1a3 lineages among modern Inuits. Scanning of the raw data outside the high confidence subset, however, revealed that the haplogroup Q defining position M242 had been ascertained as T, which is the derived state, in three reads. Given these data, the Saqqaq sequence most likely belongs to haplogroup Q, but the sequencing data were not sufficient to (a) confirm its derived status in this haplogroup, and (b) to specify more accurately the sub-clade in haplogroup Q to which the ancient DNA specimen would belong.

To confirm and specify the phylogenetic assignment of the Saqqaq Y chromosome we further tested, through conventional methods involving PCR and cloning, five Y chromosome markers: MEH2, M346, M3, M216, and M242. Through these analyses the derived status at M242 was indeed confirmed, and we were able to further specify the phylogenetic affiliation of the Saqqaq Y chromosome in haplogroup Q1a by the

derived state at marker MEH2 (Fig. 3d). Overall, the combined use of Illumina and conventional PCR/cloning based sequencing showed that all Y chromosome reads from the Saqqaq sample could be explained as deriving from a single male individual of haplogroup Q1a affiliation. In case of two markers – P240 and P257 – the difference between the Saqqaq individual and the reference sequence appeared to be due to the chimaeric nature of the reference sequence rather than because of possible contamination or ascertainment problems with the Saqqaq sample.

The fact that the Saqqaq sample belongs to haplogroup Q1a is not surprising considering the wide geographic distribution of haplogroup Q lineages among extant populations in this region. Because modern Inuits, Aleutians, Native Americans and various Siberian populations all carry Q1a lineages at moderate to high frequencies it makes this high resolution Y chromosome evidence from the ancient Saqqaq sample – as it currently stands – non-informative for questions addressing the origins of palaeo and neoeskimos, and more particularly, about the continuity between the two. However, this ambiguity is mainly due to the lack of high resolution Y chromosome SNP data for extant populations. The newly revealed 4,024 markers potentially include a number of phylogenetically informative markers that, when future additional Y chromosome-wide sequence data will becomes available for extant populations, can reveal with higher accuracy the phylogenetic affiliation of the paternal ancestry of the Saqqaq individual.

PCA analysis

We included two datasets in the principal component analysis (PCA) analysis: The Stanford Human Genome Diversity Project SNP Genotyping Data

(http://hagsc.org/hgdp/files.html) and genotype data from 16 additional populations. Both datasets were genotyped on Illumina 650K arrays.

Using the high confidence dataset from the Saqqaq genome, we selected all sites that overlapped with the Illumina SNPs with the exception of two sites, which had alleles that were inconsistent with the annotation in dbSNP 126. After merging the Saqqaq data with the SNP chip data we removed all non-autosomal SNPs, all SNPS with a minor allele frequency of less than 0.05 and all SNPs with missing data. We then performed the PCA analysis using the method described in Patterson et al. [32] (Eigensoft), which we implemented in the statistical language R. We used different subsets of the populations in the analysis. The eigenvectors were ordered according to the rank of the eigenvalues i.e. the first principal component has the largest corresponding eigenvalue.

Estimation of inbreeding

Inbreeding, i.e. identity by descent (IBD) sharing between the two chromosome copies of an individual, is usually not an observable quantity. But it can be inferred from genetic data. For example if we have genotype information from an individual and the frequency of the genotypes in the population we can estimate the amount of IBD sharing within the individual by writing up the probability of the data as the following sum over the hidden IBD state, X, of the different markers

$$p(G \mid F) = \prod_{i=1}^{m} \sum_{x=0}^{1} p(G_i \mid X = x) p(x) \quad (1)$$

where m is the number of loci, $G=\{G_1,G_2,...,G_m\}$ are the observed genotypes with $G_i \in \{AA,Aa,aa\}$, X indicates whether or not the two chromosomes are shared IBD in locus i, and the inbreeding coefficient F is the probability of being IBD at a random

locus (*p(x)*). The probabilities of the genotypes given the IBD state are here the probabilities of the alleles i.e. $P(AA|X=0)=p_A^2$, $P(AA|X=1)= p_A$, $P(Aa|X=0)=2p_Ap_a$, $P(Aa|X=1)=0$, $P(aa|X=0)= p_a^2$ and $P(aa|X=1)=p_a$. The probabilities of the alleles, p_A and p_a, at each locus *i* are usually obtained from the frequency of the alleles in a sample. If the markers are located far from each other they will be conditionally independent the expression in equation (1) is a proper likelihood function, and a maximum likelihood estimate of *F* can be obtained. Given a dense data set, where the markers are not independent, the equation can be seen as a composite likelihood function and will still be a consistent estimator. However, significance testing based on likelihood ratios using standard asymptotic theory is no longer possible [33].

For the Saqqaq genome, as for many other re-sequenced genomes based on Next-Generation sequencing, it is more difficult to call heterozygous than homozygous genotypes. Thus there will be an under-calling of heterozygotes, which will bias the above calculation. However, by removing the sites that are heterozygous in the Saqqaq individual we can still estimate the inbreeding using a modified approach. In this new modified approach we have:

$$P(G_i/G_i \neq Aa,F) = \sum_{x=0}^{1} P(G_i \mid G_i \neq Aa, X = x)P(X = x \mid G_i \neq Aa,F)$$

and the composite likelihood function becomes

$$P(G/G \neq Aa,F) = \prod_{i=1}^{m}\sum_{x=0}^{1} P(G_i \mid G_i \neq Aa, X = x)P(X = x \mid G_i \neq Aa,F)$$

The probabilities for observing the genotypes conditionally on observing only homozygous genotypes and on the IBD state *X* can be seen in Table S9. And the probability of the IBD state in locus *i* is

$$P(X = x \mid G_i \neq Aa, F) = \frac{P(X = x, G_i \neq Aa \mid F)}{P(G_i \neq Aa \mid F)} = \frac{P(X = x \mid F)P(G_i \neq Aa \mid X = x)}{P(G_i \neq Aa \mid X = 0)(1 - F) + F}$$

where the probability of being homozygous, $P(G_i \neq Aa \mid X=0)$, is estimated from the allele frequencies assuming Hardy Weinberg equilibrium in the reference population and the probability of being inbred $P(X=1|F)=F$ and not being inbreed $P(X=0|F)=1-F$. Using this composite likelihood function F can be estimated using the maximum likelihood approach. It might be counterintuitive that there is any information is left about IBD when all heterozygous sites are removed. However, some information is indeed left as can be seen in table S9 the probabilities of observing the homozygote genotypes differ depending on whether or not the individual is inbred and this difference increases with the rarity of the observed alleles.

The intuition behind this is that if the individual is not inbred then the allele that the individual is homozygous for in a given locus has to be sampled twice by chance for the observation to occur whereas if the individual is inbred it only has to be sampled once since the two alleles have to have the same type when the individual is inbred. The probability difference between these two scenarios is not high when the allele is very common – since sampling a very common allele twice is not much more unlikely than sampling it once. However, if the allele is very rare the difference is substantial – as sampling a rare allele twice is much less likely than sampling it once. We estimate standard error using a block jackknife approach as described previously in the contamination estimation section (also SI).

The above inbreeding estimation method assumes that the allele frequencies from the population the Saqqaq individual belonged to are known. However, we do not know these allele frequencies. Instead we estimate the allele frequencies from the Siberian

population. The resulting estimate of inbreeding can therefore not be interpreted as the inbreeding coefficient in any strict sense, but instead constitutes a measure of both the inbreeding in the Saqqaq genome and the variation that is explained by differences between the Saqqaq individual and the reference population. Unfortunately it is not possible to quantify how much each of these two constituents contributes with to the total estimate.

Estimation of inbreeding across the genome

Using the conditional probabilities the above approach can also be used in a hidden Markov model (HMM) framework as described in [34]. This will provide the probability of being inbred along the genome (Fig. S13). Besides giving us information about where in the genome Saqqaq is inbred, this approach has the advantage that it takes into account the varying distance between the markers and the length distribution of IBD tracts.

We condition on there being no heterozygous loci and use the emission probabilities from Table S9. The transition probabilities described in [34], see Table S10, will still hold for dense genotype data sets, and can be considered an approximation for less dense SNP data sets. Density is here relative to the recombination breakpoints occurring in the pedigree along the edges of cycles relating the individual to its ancestors. As the method assumes no LD between markers, we removed strong LD in the data. This was done by estimating the squared correlation coefficient (r^2) between pairs of markers in the sample consisting of the Saqqaq and the Siberian population. We performed the estimation in a 50 SNP sliding window so that pairwise LD for a single SNP was estimated for 49 SNPs on either sides. Then the SNPs were iteratively removed if they showed an $r^2>0.5$ with any of the adjacent SNPs. Genotyping errors

are accounted for by assuming a fixed error rate of 0.01, using an approach described in [35].

Even though the method only uses the homozygous sites the tracts we have inferred coincide with the regions where we do not observe heterozygous genotypes (see fig. S13 for the inferred tracts, and fig. S14 for a map of heterozygous sites). This would only happen if the tracts correctly inferred and are truly IBD.

In the Saqqaq genome we observe multiple long IBD tracts (>10Mb) that are far longer than the extent of LD. Thus these tracts cannot be explained by local variation in population differences between the Saqqaq population and the Siberian population but instead show that the runs of homozygosity observed in the Saqqaq individuals is due to inbreeding.

Estimation of divergence times

The estimation of divergence times is complicated by several issues, the two most important being uncertainty in genotype calls, especially a potential for under-calling homozygotes, and the ascertainment bias [36,37] observed in the SNP genotyping data analysed from the reference populations (see Ancestry Analyses section SI for a description of these data). To take these biases into account our method only uses one allele from the Saqqaq genome and conditions on the observed allele frequency spectrum in the reference population. We use a modification of the method of Nielsen *et al.* [38]. This method estimates divergence times between pairs of populations assuming all variants arose before the divergence of the populations. As the SNPs analysed here all are SNPs that have been ascertained in a global population,

primarily based on Europeans and Africans, it is highly unlikely that SNPs in our data set are specific to the Saqqaq or the population we compare the Saqqaq to.

The method is modified in two ways from its original version. First, we assume that the distribution of allele frequencies in the ancestor follows the familiar $1/x$ distribution arising in equilibrium under a standard coalescence model with mutations occurring according to an infinite sites model. Secondly, we base all inferences on the conditional distribution to avoid the effect of ascertainment biases and to increase robustness towards parametric assumptions affecting the site frequency spectrum. Let y be the allele observed in the Saqqaq individual (by randomly sampling one of the alleles in a site), and let $\mathbf{x} = (x_1, x_2, \ldots, x_{2n})$ be the site frequency spectrum in the population to which the Saqqaq is being compared. Then we calculate the likelihood function for the population size scaled divergence time, t, as

$$L(t) \propto p(y \mid t, \mathbf{x}) = \frac{p(y, \mathbf{x} \mid t)}{p(\mathbf{x} \mid t)}, \qquad (2)$$

and optimise this likelihood function numerically to obtain maximum likelihood estimates. We likewise obtain confidence intervals (CIs) using standard asymptotic theory, i.e. we construct a 95% CI by finding the value of t that would reduce the likelihood by 1.96 log likelihood units from the maximum value. Both the denominator and numerator in Eq. 2 are calculated as in Nielsen *et al.* [38], but with the mentioned modification regarding assumptions of distributions of allele frequencies at split times.

Using complete mitochondrial genomes of Chukchi (n=6) [39], Bayesian estimates of N_e were obtained with Beast v1.5.2 [40]. The substitution model HKY was chosen by the likelihood ratio test in Modeltest v3.7 [41]. For the molecular clock, we used a normal distributed clock rate prior with 95% highest posterior densities of 7.35×10^{-8} - 1.16×10^{-7}, based on the range estimated in a study by Endicott & Ho [42]. All other parameters were given uniform distributions. A strict molecular clock was applied with the constant size demographic model. The MCMC chains were run with 10^7 iterations and trees were sampled every 10^3 iterations. The first 10% of iterations were discarded as burn-in. Log-files were analysed in Tracer v1.4.1 [40], and effective sample sizes were used to evaluate convergence. We obtained an effective population size estimate of roughly 350 individuals (95% Credible Interval: 293-810). A mtDNA effective population size of 350 corresponds to an effective diploid population size of $2N_e$ = 1,400 (95% Credible Interval: 1172-3240).

The estimates obtained are scaled in terms of the effective population size of the population to which the Saqqaq is compared. We note that because we only use a single Saqqaq allele, our estimates are not functions of the effective population size of the Saqqaq population. The estimates obtained are shown in Table S11.

Ancestry analyses

A number of algorithms that aim to capture the genetic structure in large multilocus genotyping datasets have been introduced in recent years. They can be grouped under common nominator as *structure-like analyses* [43] owing to the first widely used ready-made program called *structure* [44,45]. These type of analyses construct a specified-by-the-researcher number (K) of ancestral populations, as defined by allele

frequencies at all studied loci, and simultaneously assign ancestry proportions (probabilities) to each individual in the study. A given individual may contain ancestry signal from one or more ancestral populations. The latter could be interpreted as admixture between two or more source populations. However, the term "ancestry (proportions)", used in this context, should not be interpreted as reflecting direct phylogeny. The model of genetic admixture between a chosen number of hypothetical "pure ancestral populations" – *K* - employed in the *structure*-like programs, is understandably simplified as far as demographic histories of real populations are concerned and *per se* do not allow to discriminate between: i) *similarity* due to co-ancestry or recent admixture, and ii) *dissimilarity* due to ancient split or intense drift. However, keeping that in mind, one can make reasonable use of the programs to reveal empirical structure within the data [43].

The original algorithms used Bayesian approach [44,45] but more recently likelihood-based methods [46-49] have gained foothold partly because of advantages in computational speed allowing the usage of bigger datasets produced particularly with the commercial genotyping chips (Illumina, Affymetrics). Of the maximum likelihood methods we chose a recently introduced approach assembled into program *ADMIXTURE* [49]. This algorithm is faster and, due to more stringent convergence criterion, more accurate, than alternatives like *Frappe* used in Li et al. [50].

We used the same dataset of 492 reference samples from 35 extant Asian and American populations as for the PC analyses (see Table S12). We used *PLINK* 1.05 [51] to perform data management and quality control operations. Altogether 95,502

SNPs were found in common between the different versions of the Illumina 660K or 610K genotyping chips and the Saqqaq genome high confidence SNPs.

We explored *K* values from 2 to 10 running *Admixture* 100 times at each *K* and assessed convergence by studying the Log-likelihood scores (LLs). We note that for all values of *K*, the maximum difference of LLs within the fraction of runs (5) yielding the highest LLs, was minimal (up to 0.14 LL units). Though not definitive, this observation is indicative that convergence was reached for all values of *K* tested. Plots for each *K* is shown in Figure S15. In further analyses we chose to concentrate on $K = 5$ because the ancestry components appearing at higher values of *K* were predominantly restricted to a single population and as such less informative.

Comparative human genome data

The human reference genome, build 36.3 was downloaded (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.3/Assembled_chromosomes). The Korean [52] genome data was retrieved from KOBIC (ftp://ftp.kobic.kr). The Asian genome [24] data was retrieved from the YanHuang database (http://yh.genomics.org.cn). The Venter genome [53] data was retrieved from the J. Craig Venter Institute (http://huref.jcvi.org). The SNP set from the Watson genome [54] was retrieved from Baylor College of Medicine (http://www.bcm.edu). The Yoruban genome [55] sequence was retrieved from the Illumina web site (http://www.illumina.com/iGenome). The Chimpanzee genome [56] was retrieved from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Pan_troglodytes). The sequences for the human microbiome related bacteria were retrieved from NCBI – the Human Microbiome Project (ftp://ftp.ncbi.nih.gov/genomes/HUMAN_MICROBIOM).

Metagenomic analysis of the sequence reads

To estimate the species composition of the metagenomic DNA extracted from the Saqqaq hair, we employed a "chain-mapping" method based on Bowtie [57] and BLAST [58] where reads enter a mapping pipeline with a defined set of reference organisms. Reads which cannot be aligned to a reference organism during one mapping round get passed on to the next species level where again the non-mapped reads are further passed on. The mapping chain used in this project had the following taxonomic order, only reads not mapped by *Sesam* initially were processed: 1) Human Reference Genome (Build 36.3), 2) Human sequences with unknown location, 3) Chimpanzee genome, 4) NCBI completed microbial genomes 5)Human microbiome, 6) GenBank non-redundant nucleotide database.

First, reads that were not mapped by Sesam, were aligned against the human reference genome using Bowtie with the options "-n 2 -e 70 -l 34 -m 1 –best --strata" and reporting both aligned, unaligned and alignments with more than one reportable alignment. Reads that could not be mapped against the reference genome were subsequently mapped against the unknown human sequences followed by mapping of the new unmapped reads against the Chimpanzee genome. Reads that did not align against any of these primate databases were mapped against the NCBI set of completed microbial genomes and any overrepresented microbes were assembled individually. After mapping to bacteria associated with the human microflora, the rest of all sequence reads that have not been aligned were compared to the GenBank non-redundant database using *blastall* [58]. Results are shown in Table S13 and visualised in Fig 2b.

Functional SNP assessment

SNPs shown to be associated with various phenotypes were curated from literature and HGMD [59] professional release 2009.2. Only SNPs for which experimental evidence was available for association were included, however, not all SNPs were shown to be causative. SNPs with multiple validations were preferred. SNP consequence on gene models was determined using the Ensembl [60] v. 54 API. Gene ontology associations were obtained from Biomart-Ensembl v. 54 [61]. All SNPs are summarised with rs number and posterior probability in Table S14.

Several findings can possibly be interpreted in favour of adaptations to the cold Arctic climate: The Saqqaq non-synonymous homozygous variant (C/C) change in TP53 on chromosome 17 suggests he possessed the more active form of p53 (Arg variant), which is proposed to protect against winter temperature stress through efficient regulation of metabolism [62]. Furthermore, a panel of genes involved in metabolic syndrome, as well as close interaction partners have recently been genotyped to study association with cold and heat stress [63]. Of the 34 SNPs associated with high significance to cold adaptation, the Saqqaq genome shows the non-reference allele for 12 including non-synonymous mutations in EPHX2, the leptin receptor LEPR, and T55A in FABP2 that are part of the thermogenesis pathway [64] or intracellular metabolism and transport of long-chain fatty acids [63]. For the latter, the Saqqaq allele type (C/C) increases strongly in population frequency with latitude [63] and is believed to protect against cold temperatures by increasing BMI and increasing the fuel for heat production [65]. Additionally, on chromosome 1, 7, and 11 we find non-synonymous changes which affect proteins connected to vasoconstriction (Gene Ontology - a common mechanism to reduce heat loss), higher percentage fat mass, and increased body mass index [66,67].

Disease risk was investigated in a simplistic fashion by assessing the number of known disease risk SNPs affected in the Saqqaq genome. This analysis is similar to what has been reported earlier [24], but with a larger phenotype list. The phenotypes studied were Alcohol Tolerance, Alcoholism, Nicotine Dependence, Hypertension, Parkinson's, Alzheimer's, Hypolactasia, Diabetes and Obesity. Results are shown in Fig. S16 and indicate an overall disease risk that is close to the Korean individual.

Supplementary References

[1] Meldgaard, M. Ancient Harp Seal Hunters of Disko Bay. Subsistence and Settlement at the Saqqaq Culture Site Qeqertasussuk (2400 – 1400 BC), West Greenland. *Meddelelser om Grønland*, Man & Society (Danish Polar Center, Copenhagen 2004).

[2] Gilbert, M. T. P. *et al.* Paleo-eskimo mtDNA genome reveals matrilineal discontinuity in greenland. *Science* **320**, 1787–9 (2008).

[3] Meldgaard, J. *Nationalmuseets Arbejdsmark*, 83 – 96 (1983).

[4] Grønnow, B. in Threads of Arctic Prehistory: Papers in Honour of William E. Taylor Jr., edited by Morrison, D., Pilon, J.L. Canadian Museum of Civilization, Mercury Series. Archaeological Survey of Canada Paper 149: 197 – 238. (1994)

[5] Brunner, H. and Coman, K. The Identification of Mammalian Hair. Inkata Press, Melbourne, Australia, (1974).

[6] DeGaetano, D.H. *et al.* Fungal Tunneling of Hair from a Buried Body. JFSCA 31, (4), 1048- 1054, (1992).

[7] Wilson, A.S. *et al.* Yesterday's hair-human hair in archaeology. Biologist, 48, (5), 213-217, (2001)

[8] Ramsey, C. B. *et al.* Improvements to the pre-treatment of bone at Oxford. *Radiocarbon* **46** 155-63 (2004).

[9] Brown, T.A. *et al.* Improved collagen extraction by modified Longin method. *Radiocarbon* **30**, 171-177 (1988).

[10] O'Connell, T.C. and Hedges, R.E.M. Isotopic comparison of hair and bone: Archaeological analyses. *J. Archaeol. Sci.* **26**, 661-665 (1999).

[11] Ramsey, C. B. and Hedges, R.E.M. Hybrid ion sources: Radiocarbon measurements from microgram to milligram. *Nuclear Instruments and Methods in Physics Research B* **123**, 539-545 (1997).

[12] Reimer, P.J and Reimer, R.W. A marine reservoir correction database and on-line interface. *Radiocarbon* **43**, 461-3 (2001).

[13] Krog H. and Tauber H. C-14 chronology of late- and post-glacial marine deposits in north Jutland. Danmarks geologiske Undersøgelse, Årbog 1973, 93-105 (1974).

[14] Tauber, H. ^{14}C Activity of Arctic Marine Animals. In: Radiocarbon Dating: Proceedings of the Ninth International Conference, edited by Berger, R. and Seuss, H.E. *University of California Press, Berkeley*, 447-452 (1979).

[15] Olsson, I. U. Content of ^{14}C in marine mammals from northern Europe. *Radiocarbon* **22**, 662-675 (1980).

[16] McNeely R. *et al.* Canadian marine reservoir ages, preliminary data assessment, Open File 5049, pp. 3. Geological Survey Canada (2006).

[17] Wilson A. S. *et al.* Stable isotope and DNA evidence for ritual sequences in Inca child sacrifice. *P. Natl. Acad. Sci. USA* **104**, 16456-16461 (2007).

[18] Buchardt B. *et al.* Fingernails and diet: Stable isotope signatures of a marine hunting community from modem Uummannaq, North Greenland. *Chem. Geol.* **244**, 316-329 (2007).

[19] Coltrain J. B. *et al.* Sealing, whaling and caribou: the skeletal isotope chemistry of Eastern Arctic foragers. *J. Archaeol. Sci.* **31,** 39-57 (2004).

[20] Gilbert, M. T. P. *et al.* Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science* **317**, 1927–30 (2007).

[21] Edgar, RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-97 (2004).

[22] Karafet, T. M. *et al.* High levels of Y-chromosome differentiation among native siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Hum Biol* **74**, 761–89 (2002).

[23] Becksette M. *et al.* Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinf.* **7**, 389-413 (2006).

[24] Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–5 (2008).

[25] Rhead B. *et al.* The UCSC genome browser database: update 2010. *Nucleic Acids Res*, doi:10.1093/nar/gkp939 (2009)

[26] Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–37 (2003).

[27] Reich, D. *et al*. Reconstructing Indian population history. *Nature* **461**, 489-494 (2009).

[28] Karafet T. *et al.* New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res* **18**, 830-8 (2008).

[29] Zlojutro, M. *et al.* Mitochondrial DNA and Y-chromosome variation in five

eastern Aleut communities: evidence for genetic substructure in the Aleut population. *Ann Hum Biol* **36**, 511–26 (2009).

[30] Malhi, R. S. *et al.* Distribution of Y chromosomes among native north americans: a study of Athapaskan population history. *Am J Phys Anthropol* **137**, 412–24 (2008).

[31] Bosch, E. *et al.* High level of male-biased Scandinavian admixture in greenlandic inuit shown by Y-chromosomal analysis. *Hum Genet* **112**, 353–63 (2003).

[32] Patterson N. *et al.* Population Structure and Eigenanalysis. *PLoS Genet* **2**, e190 (2006).

[33] Lindsay, B. Composite likelihood methods. In: Statistical Inference from Stochastic Processes, edited by Prabhu, N.U. American Mathematical Society, *Contemporary Mathematics*, **80**, 221–239 (1988).

[34] Leutenegger, A. L. *et al.* Estimation of the inbreeding coefficient through use of genomic data. *Am. J. Hum. Genet.* **73**, 516–523 (2003).

[35] Albrechtsen, A. *et al.* Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet. Epidemiol.* **33**, 266–274 (2009).

[36] Nielsen, R. Estimation of Population Parameters and Recombination Rates using Single Nucleotide Polymorphisms. *Genetics* **154**, 931-942 (2000).

[37] Clark, A. G. *et al*. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res*. **15**, 1496-1502 (2005).

[38] Nielsen, R. *et al.* Maximum-Likelihood Estimation of Population Divergence Times and Population Phylogeny in Models without Mutation. *Evolution* **52**, 669-677 (1998).

[39] Derenko, M. et al. Phylogeographic analysis of mitochondrial DNA in northern Asian populations. *Am. J. Hum. Genet.* **81**, 1025-41 (2007).

[40] Drummond, A.J. and Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Bio.* **7**, 214-21 (2007).

[41] Posada, D. and Crandall, K. A. MODELTEST: testing the model of DNA substitution. *Bioinf*. **14**, 817-8 (1998).

[42] Endicott, P. and Ho, S.Y.W. A Bayesian evaluation of human mitochondrial substitution rates. *Am. J. Hum. Genet.* **82**, 895-902 (2008).

[43] Weiss, K. M. and Long, J. C. Non-Darwinian estimation: my ancestors, my

genes' ancestors. *Genome Res* **19**, 703-10 (2009).

[44] Pritchard, J. K. *et al.* Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-59 (2000).

[45] Falush, D. *et al.* Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics* **164**, 1567-1587 (2003).

[46] Purcell, S. and Sham, P. Properties of structured association approaches to detecting population stratification. *Hum Hered* **58**, 93-107 (2004).

[47] Tang, H. *et al.* Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol* **28**, 289-301 (2005).

[48] Zhu, X. *et al.* A classical likelihood based approach for admixture mapping using EM algorithm. *Hum Genet* **120**, 431-45 (2006).

[49] Alexander, D. H. *et al.* Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-64 (2009).

[50] Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-4 (2008).

[51] Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

[52] Ahn, S.-M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* **19**, 1622–9 (2009).

[53] Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254 (2007).

[54] Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–6 (2008).

[55] Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–9 (2008).

[56] Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87 (2005).

[57] Langmead, B. *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).

[58] Altschul, S. F. *et al.* Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).

[59] Stenson, P. *et al.* The Human Gene Mutation Database: 2008 update. *Genome*

Med. **1**, 13 (2009).

[60] Hubbard, T. J. P. *et al.* Ensembl 2009. *Nucleic Acids Res.* **37**, D690-697 (2009).

[61] Smedley, D. *et al.* BioMart--biological queries made easy. *BMC Genomics* **10**, 22-33 (2009).

[62] Hong, S. *et al.* Winter temperature and UV are tightly linked to genetic changes in the p53 tumor suppressor pathway in Eastern Asia. *Am. J. Hum. Genet.* **84**, 534-41 (2009).

[63] Hancock, A. *et al.* Adaptations to Climate in Candidate Genes for Common Metabolic Disorders. *PLoS Genet*. **4**, e32 (2008).

[64] Dulloo, A. G. *et al.* Leptin directly stimulates thermogenesis in skeletal muscle. *FEBS Lett*. **515**, 109-13 (2002).

[65] Baier, L. J. *et al.* An amino acid substitution in the human intestinal fatty acid binding protein is associated with increased fatty acid binding, increased fat oxidation, and insulin resistance. *J. Clin. Invest.* **95**, 1281-7 (1995).

[66] Zhao, L-J. *et al.* Polymorphisms of the tumor necrosis factor-alpha receptor 2 gene are associated with obesity phenotypes among 405 Caucasian nuclear families. *Hum Genet.* **124**, 171-7 (2008).

[67] Shiri-Sverdlov, R. *et al.* Identification of TUB as a novel candidate gene influencing body weight in humans. *Diabetes* **55**, 385-9 (2006).

[68] Stuiver, M. and Polach, H.A. Discussion: Reporting of ^{14}C data. *Radiocarbon* **19**, 355-63 (1977).

[69] Reimer P. *et al.* IntCal04 terrestrial radiocarbon age calibration, 0-26 cal kyr BP. *Radiocarbon* **46**, 1029-58 (2004).

[70] Yamamoto, F. *et al.* Molecular genetic basis of the histo-blood group ABO system. *Nature* **345**, 229-33 (1990).

[71] Yamamoto, F. *et al.* Human histo-blood group A2 transferase coded by A2 allele, one of the A subtypes, is characterized by a single base deletion in the coding sequence, which results in an additional domain at the carboxyl terminal. *Biochem. Biophys. Res. Commun*. **187**, 366-74 (1992).

[72] Iida, R. *et al.* Genotyping of five single nucleotide polymorphisms in the OCA2 and HERC2 genes associated with blue-brown eye color in the Japanese population. *Cell Biochem Funct* **27**, 323–7 (2009).

[73] Soejima, M., and Koda, Y. Population differences of two coding SNPs in

pigmentation-related genes SLC24A5 and SLC45A2. *Int. J. Legal. Med.* **121**, 36–9 (2007).

[74] Prodi, D. A. *et al.* EDA2R is associated with androgenetic alopecia. *J. Invest. Dermatol.* **128**, 2268–70 (2008).

[75] Ellis, J. A. *et al.* Baldness and the androgen receptor: the AR polyglycine repeat polymorphism does not confer susceptibility to androgenetic alopecia. *Hum. Genet.* **121**, 451–7 (2007).

[76] Yoshiura, K. *et al.* A SNP in the ABCC11 gene is the determinant of human earwax type. *Nat. Genet.* **38**, 324–30 (2006).

[77] Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–8 (2007).

[78] Branicki, W. *et al.* Association of the SLC45A2 gene with physiological human hair colour variation. *J. Hum. Genet.* **53**, 966–71 (2008).

[79] Shi, H. *et al.* Winter temperature and UV are tightly linked to genetic changes in the p53 tumor suppressor pathway in Eastern Asia. *Am. J. Hum. Genet*. **84**, 534-41 (2009).

Table S1. Radiocarbon determinations from the Qeqertasussuk site. Radiocarbon ages are expressed in years BP after Stuiver and Polach [68]. Stable isotope ratios are expressed in ‰ relative to vPDB and nitrogen to AIR. Mass spectrometric precision is ±0.2‰ for carbon and ±0.3‰ for nitrogen. Wt. used is the amount of bone pretreated and the yield represents the weight of gelatin or ultrafiltered gelatin in milligrams. %yield is the wt.%collagen which should not be <1wt.% at ORAU. This is the amount of collagen extracted as a percentage of the starting weight. %C is the carbon present in the combusted gelatin. CN is the atomic ratio of carbon to nitrogen. At ORAU, for bone collagen this is acceptable if it ranges between 2.9—3.5, for hair keratin the range is slightly wider (3.0-3.8).

OxA	Sample identification	Material	Species	Radiocarbon age BP	Wt. used	Yield	%Yld	%C	$\delta^{13}C$	$\delta^{15}N$	CN
18746	Qt 85 FC 85.0/251.0:4	bone	*R. tarandus*	3505 ± 29	590	19.6	3.3	39.9	-19.4	1.3	3.4
18747	Qt 85 F8 10.0/23.5: 76 3287	bone	*R. tarandus*	3713 ± 28	474	40.2	8.5	42.7	-17.6	3.5	3.3
18748	Qt 85 FB 10.0/23.5: 95 5115	bone	*R. tarandus*	3664 ± 28	436	31.6	7.2	43.1	-17.6	2.8	3.3
18749	Qt 86 PC 85/261: 11	bone	*R. tarandus*	3628 ± 28	362	40.1	11.1	43.2	-18.0	0.1	3.3
20656	Hair sample	hair	*sapiens*	4044 ± 31	25.8	21	81.2	45.3	-13.9	19.3	3.7

Table S2. Calibrated ages BC from the Qeqertasussuk site. The results are calibrated against the INTCAL04 curve [69]. See SI text for details of reservoir correction.

	Calibrated age range BP (68.2% prob.)		Calibrated age range BP (95.4% prob.)	
	From	To	From	To
OxA-18746	3833	3723	3860	3695
OxA-18747	4139	3989	4148	3979
OxA-18748	4079	3927	4085	3905
OxA-18749	3977	3901	4073	3854
OxA-20656	4035	3728	4170	3600

Table S3. Expected and measured carbon and nitrogen isotopic values of standards run alongside the samples at Oxford and Bradford.

Standard	Expected $\delta^{13}C$ (‰)	Measured $\delta^{13}C$ (‰)	Expected $\delta^{15}N$ (‰)	Measured $\delta^{15}N$ (‰)
Oxford	-26.1	-26.4	-1.8	-1.7
		-26.6		-1.8
		-26.4		-1.6
		-26.1		-2.0
Bradford	-27.8	-27.7	1.0	1.1
		-27.5		1.1

Table S4. 4x4 error matrices. For each reaction a multiple alignment was generated (using MUSCLE v3.7, [21]), majority count identified correct base and all differences were counted to generate these matrices. Gaps and N's are omitted from the table.

HiFi tag 03 RT1HF (17871 sequences)				
	A	C	G	T
A	0.9976	0	0.0011	0
C	0.0007	0.9901	0	0.0071
G	0.0095	0	0.9875	0
T	0.0001	0.0022	0	0.9971
HiFi tag 04 RT2HF (6402 sequences)				
	A	C	G	T
A	0.9959	0.0002	0.0023	0.0003
C	0.001	0.9531	0.0002	0.0352
G	0.0259	0	0.9672	0.0003
T	0.0004	0.0034	0	0.9951
HiFi tag 05 WR1HF (5661 sequences)				
	A	C	G	T
A	0.9954	0.0003	0.0029	0.0006
C	0.0005	0.9817	0.0006	0.0139
G	0.0088	0	0.9896	0.0002
T	0.0006	0.0032	0.0001	0.9955
Phusion tag 09 RT1PH (12662 sequences)				
	A	C	G	T
A	0.999	0	0.0001	0.0001
C	0.0003	0.9986	0	0.0007
G	0.004	0	0.9948	0.0003
T	0.0001	0.0023	0	0.9967
Phusion tag 10 RT2PH (7279 sequences)				
	A	C	G	T
A	0.996	0.0003	0.0011	0.0013
C	0.0017	0.9942	0	0.0022
G	0.0028	0	0.9937	0.0012
T	0.001	0.0037	0	0.9946
Phusion tag 11 WR1PH (13281 sequences)				
	A	C	G	T
A	0.9987	0	0.0005	0.0002
C	0.0008	0.9968	0.0002	0.0018
G	0.0004	0	0.9994	0
T	0.0002	0.0014	0	0.9983

Table S5. Summary Statistics. For each chromosome and the full genome, we calculate the coverage (percent of nucleotides covered), the average depth, and the number of high confidence SNPs.

Chromosome	Coverage	Average depth	All SNPs	High-confidence SNPs
chr1	79.08%	21.58	178660	30707
chr2	84.49%	18.86	174727	28795
chr3	84.93%	17.13	138013	19536
chr4	82.88%	15.29	145200	16548
chr5	84.10%	17.15	120601	18104
chr6	84.82%	17.57	127304	18367
chr7	83.87%	19.88	129227	20214
chr8	84.55%	18.52	109176	17060
chr9	69.72%	21.18	96875	16057
chr10	83.99%	21.59	120182	21479
chr11	86.01%	21.61	111926	19235
chr12	87.10%	19.74	103456	15184
chr13	71.66%	16.05	75212	10646
chr14	72.07%	19.97	68041	11067
chr15	69.17%	21.71	66345	12651
chr16	78.22%	28.18	86198	16008
chr17	89.21%	30.65	69020	13387
chr18	85.49%	18.56	59537	9992
chr19	81.47%	37.34	62706	9923
chr20	87.23%	26.70	56700	11126
chr21	61.98%	22.69	34035	4785
chr22	63.28%	34.55	38287	8927
chrX	69.04%	8.97	34247	3071
chrY	19.60%	12.13	4024	243
mtDNA	94.07%	3,802.81	40	39
Full genome	79.31%	19.94	2209739	353151

Table S6. Results of contamination analyses with standard errors.

Variable	**Estimate**	**Info**
F	0.0035/3	All reads
F	0.000886/3	Quality cut off
O	1286/392470	All reads
O	888/344819	Quality cut off
$\hat{C}$	0.00846±0.0021	All reads
$\hat{C}$	0.011±0.0021	Quality cut off

Table S7. Observed private minor alleles. AA denotes a genotype with two of the major allele and aa a genotype with two minor alleles. The two first columns denote whether the minor allele exists in Asia or Europe respectively

Exists in Asia	**Exists in Europe**	**AA**	**Aa**	**aa**
No	No	1096	0	1
Yes	No	1152	27	4
No	Yes	2341	0	0
Yes	No	76389	22202	10336

Table S8. SNPs overlapping with Y chromosome consortium (YCC) SNPs [22]. YCC SNP name, position on Y chromosome, Reference Allele, match or no match to reference (y/n), Saqqaq genotype determined, 1 – posterior probability, sequencing depth.

SNP	Position	Reference allele	match	Saqqaq	1-pp	Depth
47z	3496441	G	y	GG	3.58E-108	38
50f2(P)	21906454	C	y	CC	4.47E-18	5
IMS-JST002611	7606725	G	y	GG	2.35E-47	17
M103	20395525	C	y	CC	1.33E-34	11
M105	20325878	C	y	CC	1.61E-15	4
M108.2	20392322	T	y	TT	4.75E-24	7
M110	20366484	T	y	TT	2.36E-21	6
M12	7643479	G	y	GG	3.40E-26	8
M123	20223973	C	y	CC	4.14E-21	6
M127	20222660	G	y	GG	1.23E-25	8
M129	20177104	C	y	CC	2.06E-74	27
M136	20353140	C	y	CC	1.84E-47	15
M138	20172242	G	y	GG	8.38E-47	15
M145=P205	20176595	C	y	CC	3.09E-36	12
M146	20238638	T	y	TT	8.75E-20	6
M150	20328906	C	y	CC	4.57E-39	13
M152	20327455	C	y	CC	6.30E-44	14
M156	20176614	T	y	TT	7.14E-30	10
M157	20327241	A	y	AA	2.53E-24	7
M158	20175753	C	y	CC	6.43E-27	8
M159	20210827	A	y	AA	3.98E-23	7
M16	20069688	C	y	CC	8.27E-43	14
M162	8680322	T	y	TT	5.89E-24	7
M168	13323384	T	y	TT	5.78E-18	5
M174	13463673	T	y	TT	1.16E-20	6
M179	13348093	C	y	CC	1.38E-23	7
M185	13414252	C	y	CC	1.60E-65	22
M188	13434262	C	y	CC	3.50E-52	17
M192	13523655	C	y	CC	2.79E-85	31
M194	13523943	T	y	TT	5.65E-36	11
M195	13525929	A	y	AA	1.39E-35	12
M203	14100930	G	y	GG	6.31E-62	23
M204	14100594	T	y	TT	4.31E-44	18
M208	14085596	C	y	CC	1.42E-49	16
M211	14054007	C	y	CC	5.77E-30	9
M214	13981318	T	y	TT	2.58E-43	14
M22	13299556	A	y	AA	2.65E-51	20
M226	14100840	C	y	CC	1.31E-101	35
M227	14100839	C	y	CC	3.05E-112	39
M236	2709695	C	y	CC	3.21E-65	23
M27	20199033	C	y	CC	1.11E-32	10
M272	21148162	A	y	AA	3.81E-21	6

M274	21147188	C	y	CC	8.27E-91	34
M28	20189226	T	y	TT	7.35E-43	15
M288	2709693	G	y	GG	6.93E-70	24
M3	17605756	G	y	GG	4.18E-24	7
M300	21158585	G	y	GG	4.88E-47	15
M31	20199141	G	y	GG	1.83E-47	15
M314	21162466	A	y	AA	1.30E-15	4
M323	20327105	C	y	CC	1.42E-35	12
M324	2881785	G	y	GG	1.81E-32	10
M329	2935526	G	y	GG	5.62E-30	9
M34	20200103	G	y	GG	7.19E-32	10
M343	2947823	A	n	CC	5.60E-36	14
M346	2947154	A	y	AA	1.37E-20	6
M347	2937477	C	y	CC	3.23E-40	13
M35	20201090	G	y	GG	1.16E-52	17
M365	2948677	A	y	AA	5.87E-35	11
M367	2948627	A	y	AA	3.73E-87	31
M368	2948631	A	y	AA	1.20E-78	30
M369	2948476	G	y	GG	1.08E-20	6
M370	2948597	C	y	CC	3.26E-128	45
M387	2800273	T	y	TT	3.03E-21	6
M41=P210	2723888	C	y	CC	6.10E-64	21
M427	17601990	G	y	GG	1.55E-32	11
M428	17601952	C	y	CC	1.04E-26	8
M450	7608914	G	y	GG	9.24E-69	24
M49	20328113	T	y	TT	1.78E-46	15
M5=P73	20069333	C	y	CC	1.06E-20	6
M50	20328059	T	y	TT	6.41E-61	20
M51	20328250	G	y	GG	2.34E-38	12
M54	20328651	G	y	GG	5.28E-173	65
M56	20332273	A	y	AA	1.60E-21	6
M58	20195691	G	y	GG	2.21E-15	4
M59	20328163	A	y	AA	4.10E-28	9
M61	20210836	C	y	CC	1.75E-38	12
M63	20330025	G	y	GG	1.04E-32	10
M64.2	20362770	A	y	AA	8.21E-16	4
M65	20365652	A	y	AA	3.09E-84	32
M66	20340960	A	y	AA	2.05E-26	8
M70	20353268	A	y	AA	3.31E-15	4
M71	20353834	C	y	CC	6.67E-18	5
M8	7351533	G	y	GG	2.03E-15	4
M83	20332674	C	y	CC	3.06E-114	40
M89	20376700	T	y	TT	1.35E-21	6
M90	20383716	C	y	CC	1.38E-15	5
M93	20361892	C	y	CC	8.76E-24	7
M98	20238621	C	y	CC	1.48E-26	8
N1	13360919	C	y	CC	5.46E-21	6
N2	20359275	T	y	TT	7.26E-27	8
N4	20397919	A	y	AA	1.65E-15	4

N5	13360931	T	y	TT	2.50E-15	4
P100	20362080	C	y	CC	6.11E-56	19
P101	20362005	C	y	CC	4.73E-89	34
P102	13987410	T	y	TT	1.09E-103	41
P103	13987224	C	y	CC	7.75E-131	46
P104	13987218	C	y	CC	2.47E-122	45
P105	13932355	C	y	CC	2.54E-55	18
P106	13932315	C	y	CC	6.87E-136	53
P107	13329286	G	y	GG	3.35E-92	32
P108	13935640	T	y	TT	1.59E-86	33
P111	13935055	G	y	GG	5.53E-21	6
P113	6800170	A	y	AA	3.70E-23	7
P114	13530048	G	y	GG	1.48E-163	83
P115	13378643	G	y	GG	3.77E-38	13
P116	13379083	A	y	AA	1.05E-132	50
P117	13329086	G	y	GG	3.30E-18	5
P120	13935393	C	y	CC	1.79E-15	4
P121	6799856	C	y	CC	1.09E-57	19
P122	13360125	C	y	CC	3.04E-98	35
P126	19685157	C	y	CC	8.84E-30	9
P127	8650751	C	y	CC	2.51E-29	9
P133	10476739	A	y	AA	1.34E-17	5
P135	20078243	T	y	TT	2.11E-13	5
P138	12709283	C	y	CC	1.31E-15	4
P140	15821368	C	y	CC	2.50E-15	4
P141	7001217	A	y	AA	1.75E-32	10
P142	7278078	A	y	AA	2.45E-18	5
P146	8662414	T	y	TT	1.25E-60	21
P151	8740660	C	y	CC	1.52E-16	5
P158	16002906	T	y	TT	8.06E-102	37
P160	8534188	C	y	CC	1.71E-38	12
P167	10461456	G	y	GG	4.30E-83	28
P169	21327964	C	y	CC	4.82E-26	8
P172	7025214	C	y	CC	4.17E-73	25
P174	14318719	G	y	GG	3.12E-73	26
P180	17110667	G	y	GG	1.12E-34	11
P181	15903504	C	y	CC	2.41E-38	12
P183	7392131	C	y	CC	2.41E-32	11
P184	7278127	T	y	TT	1.19E-20	6
P186	7628567	C	y	CC	2.39E-41	13
P187	9168251	T	y	TT	4.93E-57	20
P188	22043749	G	y	GG	5.35E-100	38
P194	14712373	C	y	CC	3.28E-38	13
P196	14263706	C	y	CC	5.61E-14	4
P207	8679537	A	y	AA	1.48E-38	12
P211	2723706	T	y	TT	6.44E-19	6
P212	3605069	T	y	TT	1.09E-31	10
P22=M104	8679591	G	y	GG	1.19E-64	22
P221	8413706	C	y	CC	6.74E-35	11

P226	8905379	T	y	TT	1.42E-20	6
P227	19869093	C	n	GG	6.16E-07	4
P230	15979505	A	y	AA	1.12E-32	10
P231	10599614	G	n	AA	6.77E-19	8
P238	7831130	A	n	GG	2.80E-54	20
P239	16390623	C	y	CC	7.63E-17	5
P240	13108815	T	n	CC	5.63E-17	8
P243	8745082	G	y	GG	6.03E-29	9
P244	12943106	C	y	CC	4.13E-83	29
P249	8395202	G	y	GG	1.97E-18	6
P254	3604897	C	y	CC	1.79E-18	5
P255	8745037	G	y	GG	5.51E-24	7
P256	8745229	G	y	GG	2.48E-15	4
P257	12942935	A	n	GG	3.54E-40	15
P259	14100867	T	y	TT	1.09E-80	30
P262	6992147	G	y	GG	5.97E-53	17
P263	19625356	A	y	AA	1.55E-15	4
P266	6799737	T	y	TT	2.86E-119	40
P268	22399064	T	y	TT	4.97E-21	6
P269	22399065	A	y	AA	2.44E-21	6
P277	14088608	A	y	AA	4.61E-27	8
P278	8527052	G	y	GG	5.12E-43	14
P280	20302477	G	n	CC	4.25E-08	4
P282	16538054	G	y	GG	9.03E-15	4
P286	16225644	T	n	CC	4.56E-17	7
P287	20531484	G	y	GG	5.76E-82	28
P289	8527081	C	y	CC	8.54E-75	25
P291	8526987	A	y	AA	1.23E-35	12
P293	8835177	G	y	GG	1.17E-35	11
P3	20069972	C	y	CC	4.31E-21	6
P38	12994386	A	y	AA	9.90E-98	35
P40	12994401	C	y	CC	3.18E-94	34
P42	12991736	G	y	GG	1.27E-74	25
P43	20340360	G	y	GG	7.61E-15	4
P45	20340134	G	y	GG	2.46E-32	12
P5	20303744	G	y	GG	3.71E-23	8
P51	13002049	T	y	TT	3.72E-19	6
P57	12997073	T	y	TT	4.57E-15	4
P58	12996674	T	y	TT	2.70E-34	11
P59	12994544	A	y	AA	1.72E-25	8
P60	12994478	T	y	TT	1.07E-54	18
P71	20071465	A	y	AA	6.40E-16	4
P75	20340385	G	y	GG	1.12E-26	8
P76	20340414	G	y	GG	1.25E-24	9
P77	13435595	G	y	GG	2.25E-62	22
P80	6799898	C	y	CC	4.55E-48	17
P81	6799855	G	y	GG	2.80E-51	19
P82	6799825	C	y	CC	3.16E-12	4
P83	6799771	G	y	GG	5.77E-46	15

P84	6799766	G	y	GG	2.16E-49	18
P87	13529112	A	y	AA	2.24E-68	25
P90	13359972	C	y	CC	1.27E-69	24
P91	13359992	C	y	CC	9.52E-61	20
P93	13378895	C	y	CC	6.19E-95	35
P94	13379074	G	y	GG	4.59E-127	47
P95	13379099	G	y	GG	1.07E-160	58
P96	13379136	C	y	CC	6.31E-33	10
P97	13395666	T	y	TT	4.00E-61	22
P98	13395750	C	y	CC	7.58E-125	43
P99	13395779	C	y	CC	6.34E-57	21
PK2	21078607	T	y	TT	1.49E-33	12
PK4	19745349	A	y	AA	3.48E-21	6
RPS4Y711=M130	2794853	C	y	CC	2.16E-41	14
SRY2627=M167	2718270	G	y	GG	4.72E-81	30
SRY4064=M40=P20	2723942	C	y	CC	2.34E-149	55
SRY465	2715179	G	y	GG	1.36E-117	43
SRY9138 =M177	2718868	G	y	GG	4.06E-53	18
U106=M405	8856077	C	y	CC	1.04E-37	12
U175	14763087	G	y	GG	4.19E-21	6
U181	14883819	C	y	CC	2.55E-15	4
U186	14987945	A	y	AA	5.09E-24	7
U247=P253	17348768	C	y	CC	8.83E-16	6
U290	20105445	T	y	TT	9.76E-146	53
USP9Y+3636=M222	13411807	G	y	GG	4.44E-24	7
V13	6902262	G	y	GG	2.46E-35	11
V27	6956050	A	y	AA	1.10E-29	9
V32	6992820	G	y	GG	4.82E-23	7
V6	6992006	G	y	GG	2.38E-38	14
V65	15797065	G	y	GG	1.84E-18	6

Table S9. Probabilities of observing the genotypes given the inbreeding state conditional on that there are no heterozygotes $P(G_i | G_i \neq Aa, X)$,. p_A and p_a are the frequencies of the major and minor allele in locus *i* respectively.

	X=0	X=1
AA	$p_A^2/(p_A^2+p_a^2)$	p_A
aa	$p_a^2/(p_A^2+p_a^2)$	p_a

Table S10. The transition probabilities for the Markov chain. Where t is time (measured in genetic distance), F is the inbreeding coefficient, and e^{-at} corresponds to no change in the co-ancestry over a segment of length t [34].

	$X_{i+1}=0$	$X_{i+1}=1$
$X_i=0$	$e^{-at}+(1- e^{-at})(1-F)$	$(1- e^{-at})F$
$X_i=1$	$(1- e^{-at})(1-F)$	$e^{-at}+(1- e^{-at})F$

Table S11. The estimates of population size scaled divergence times.

	Na-Dene	Chukchis	Koryaks	Han Chinese	Nganassans
ML estimates	0.093	0.043	0.106	0.20	0.089
95% CI lower	0.085	0.035	0.098	0.18	0.080
95% CI upper	0.101	0.051	0.114	0.22	0.097

Table S12. Reference populations used to study the phylogeographic context of the Saqqaq genome. We have used relevant populations from the published data from Li et al. [50] and generated 197 new genome scans with Illumina 660/610K genotyping chips. "Group name" refers to usage of population groupings on Fig 3b, while "nr" refers to the numbers in Fig 3a.

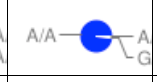
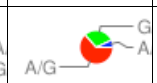
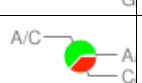
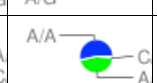
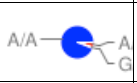
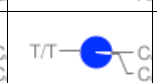
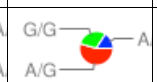
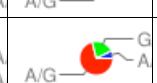
			Region					
	nr	Popualtion	Americas	East Asia	Siberia	Europe	Total	Group name
Li et al	1	Tuscan				7	7	Europeans
-	2	North Italian				12	12	Europeans
-	3	Russian				25	25	Europeans
-	4	French				28	28	Europeans
-	5	Daur		9			9	East Asia
-	6	Hezhen		9			9	East Asia
-	7	Oroqen		9			9	East Asia
-	8	Xibo		9			9	East Asia
-	9	Mongola		10			10	East Asia
-	10	Tu		10			10	East Asia
-	11	Uygur		10			10	East Asia
-	12	Japanese		25			25	East Asia
-	13	Han		44			44	East Asia
-	14	Yakut			25		25	
-	15	Colombian	7				7	South America
-	16	Surui	8				8	South America
-	17	Karitiana	13				13	South America
-	18	Pima	14				14	Middle America
-	19	Maya	21				21	Middle America
Li et al Sum			63	135	25	72	295	
This study	20	Buryats			19		19	Southern Siberia
-	21	Koryaks			17		17	
-	22	Evenkis			16		16	
-	23	Tuvinians			16		16	Southern Siberia
-	24	Nganassans			15		15	
-	25	Chukchis			14		14	
-	26	Altai			13		13	Southern Siberia
-	27	Selkups			10		10	
-	28	Mongolians			9		9	East Asia
-	29	Yukaghirs			9		9	

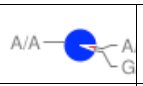
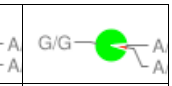
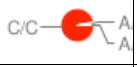
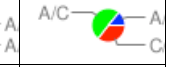
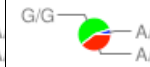
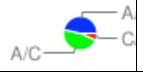
-	30	Dolgans			7		7
-	31	Kets			2		2
-	32	Na-Dene	21				21
-	33	East Greenlanders	10				10
-	34	West Greenlanders	10				10
-	35	Aleutians	9				9
-		Saqqaq	1				1
This study Sum			51		147		198
Total			114	135	172	72	493

Table S13. Reference organisms used in the "chain-mapping" method for estimation of the composition of the metagenomic DNA extracted from the Saqqaq hair. First column describes the reference organism and the last column shows the number of mapped/identified at each step

Human Reference Genome (Build 36.3)	2,701,861,758
Human sequences with unknown location	72,656,547
Chimpanzee genome	25,443,281
NCBI completed microbial genomes	2,544,465
Human microbiome	10,608,836
GenBank non-redundant nucleotide database	185,197,431
Unidentified	501,301,701

Table S14. List of SNP identifiers used for functional assessment. Associations to phenotype were curated from literature, public databases and HGMD Professional. Frequencies from Hapmap Phase-II are represented in pie charts, with the colour of the Saqqaq Allele matching the corresponding genotype in the pie charts. In most cases, the Saqqaq allele is closest to the Asian (Han Chinese and Japanese) populations.

SNP	Reference	1-pp	Saqqaq Allele	HapMap frequencies			
				CEU	*HCB*	*JPT*	*YRI*
rs8176719[1]	66	NA	-/-	NA	NA	NA	NA
rs8176750[2]	67	8.84E-34	G/G	NA	NA	NA	NA
rs12913832[3]	68	5.13E-10	A/A				
rs7495174[3]	68	1.14E-25	A/G				
rs4778241[3]	68	8.24E-06	A/A				
rs1129038[3]	68	8.69E-40	C/C	NA	NA	NA	NA
rs1426654[4]	69	8.78E-12	G/G				
rs1385699[5]	70	8.57E-05	T/T				
rs6152[5]	71	4.12E-21	G/G	NA	NA	NA	NA
rs1528133[6]	63	1.50E-08	A/C				
rs2272383[6]	63	1.42E-27	A/G				
rs2272382[6]	63	8.40E-60	A/G				
rs5746059[7]	62	1.64E-05	A/A				
rs17822931[8]	72	8.36E-09	T/T				
rs3827760[9]	73	5.61E-12	C/C				
rs16891982[10]	74	3.87E-13	C/C				
rs1042522[11]	75	2.75E-11	G/G				
rs13222385[12]	59	3.21E-05	A/G				
rs751141[12]	59	2.13E-07	T/T				
rs1800404[12]	59	2.51E-14	A/A				

rs1426654[12]	59	8.78E-12	G/G				
rs2570932[12]	59	0.0015097	C/C				
rs12946618[12]	59	0.0013802	A/A				
rs12946115[12]	59	2.44E-06	C/C				

[1]Blood Group: not type O. [2]Blood Group: A1 subtype. [3]Brown eyes. [4]Not European light skinned. [5]Increased risk of baldness. [6]Higher body mass index. [7]Higher percentage fat mass in Caucasian and Chinese samples. [8]Dry earwax, common in Asian people. [9]Thick hair and Shovel shaped upper front teeth. [10]More likely to have black hair (in European cohort study). [11] Cold adaptation: non-synonymous change in TP53. [12] Cold adaptation: Metabolic genes.

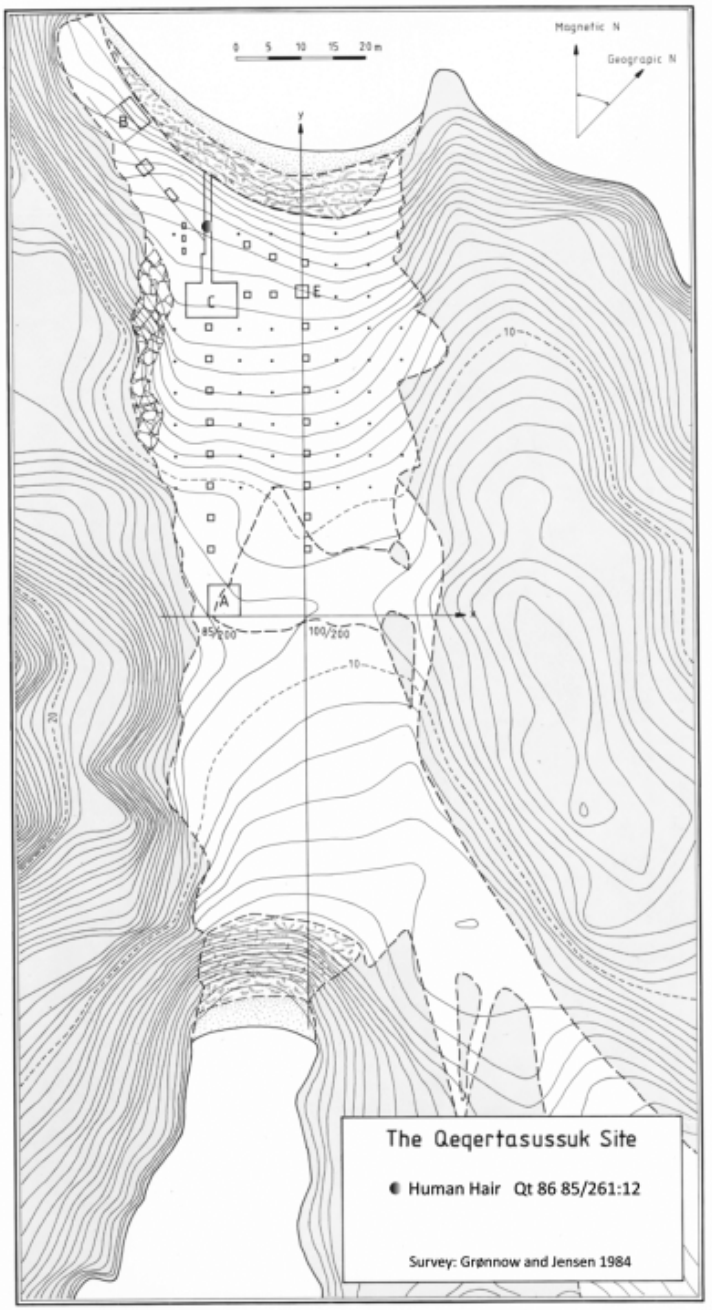


Fig. S1. Plan of the excavation areas at Qeqertasussuk. Ten percent of the site was excavated and its stratigraphy and extent was mapped in two main excavation units and through an extensive net of test pits and trenches. The position of the human hair Qt 86 85/261: 11 is marked with a dot.

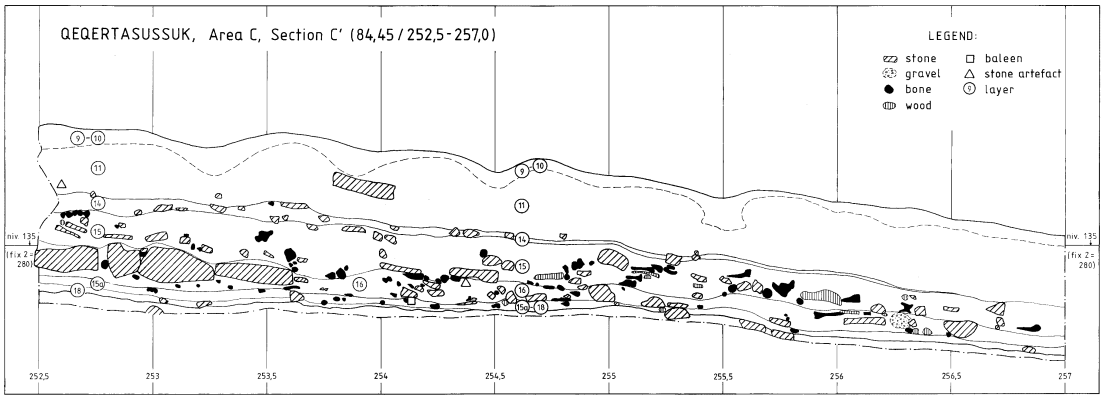


Fig. S2. Layer 18 represents the earliest traces of habitation at Qeqertasussuk ca. 2500 BC and layer 11 the youngest (ca. 1400 BC). It was in layer 15 dated to 2100 – 1900 BC that the human hair tuft was found.

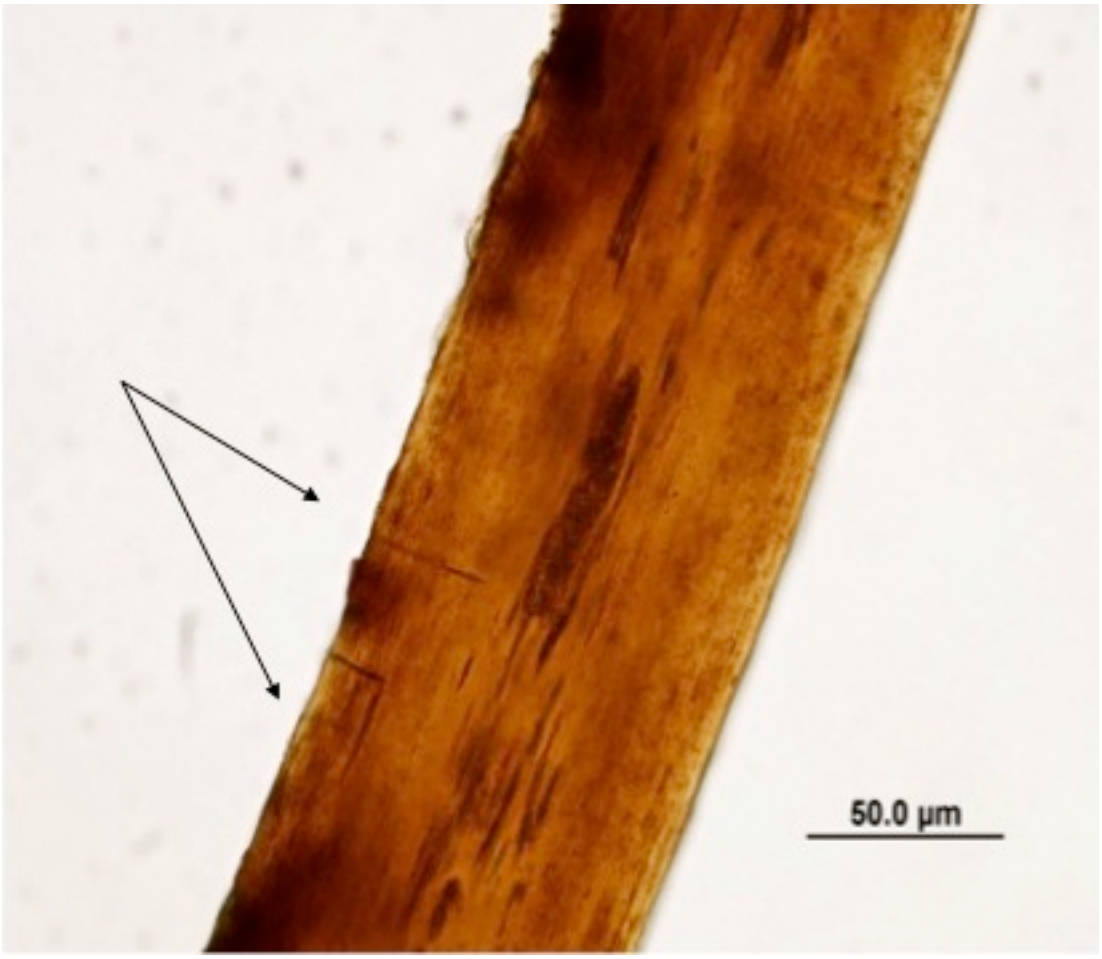


Fig. S3. The characteristic signs of the early stages of fungal tunnelling were apparent on the Saqqaq hairs; this figure is an illustration of this phenomenon.

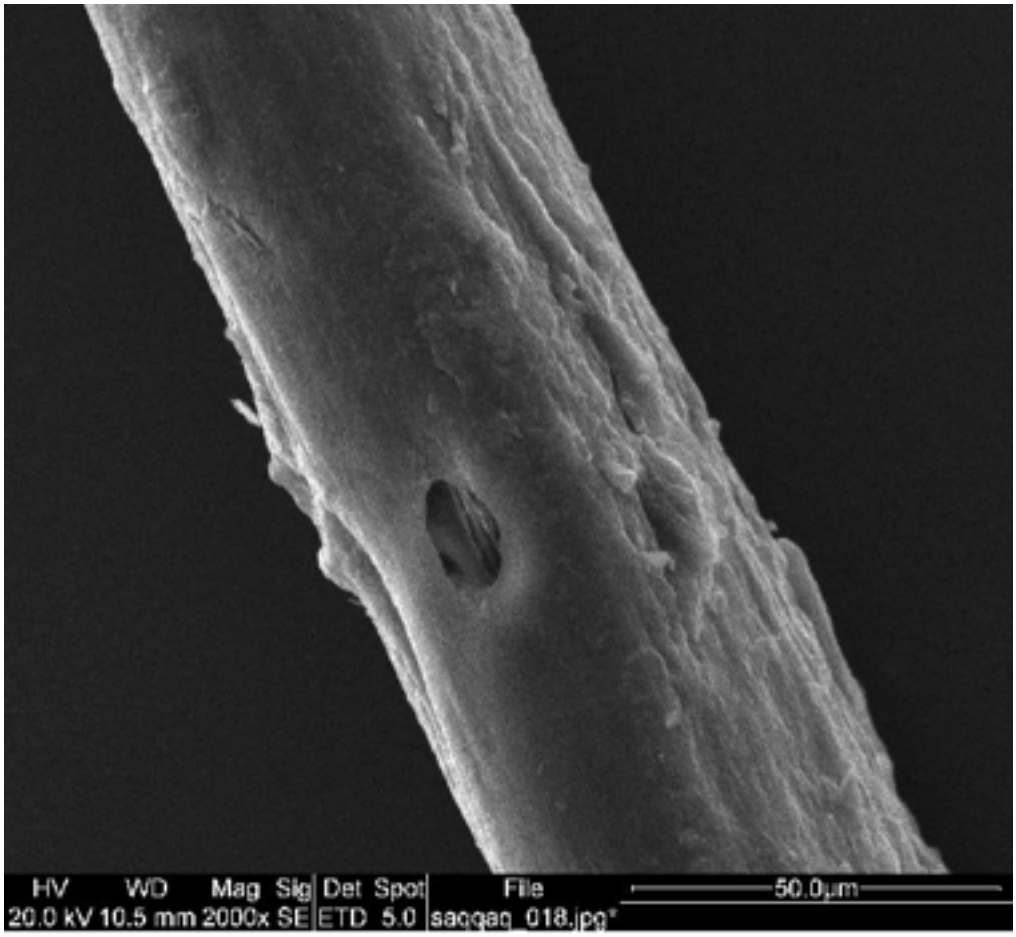


Fig. S4. Scanning Electron Microscopy photo showing lesion typical of fungal tunnelling.

Fig. S5. This figure illustrates the type of debris attached to many of the Saqqaq hair shafts, many of which exhibited protuberances similar to the one marked in the figure. This figure additionally shows the outline of the cuticle edges and the exposure of the cortex.

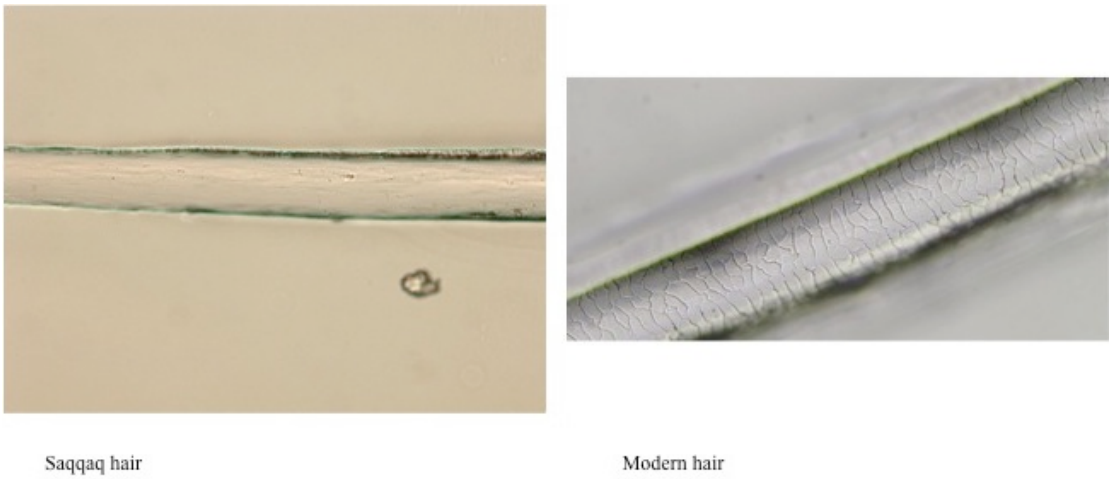


Fig. S6. Photographic images of scale cast patterns derived from Saqqaq hair [A] and modern hair [B] which support the premise that keratinophilic fungi have 'peeled' the cuticle from the hair shaft to reveal the cortex.

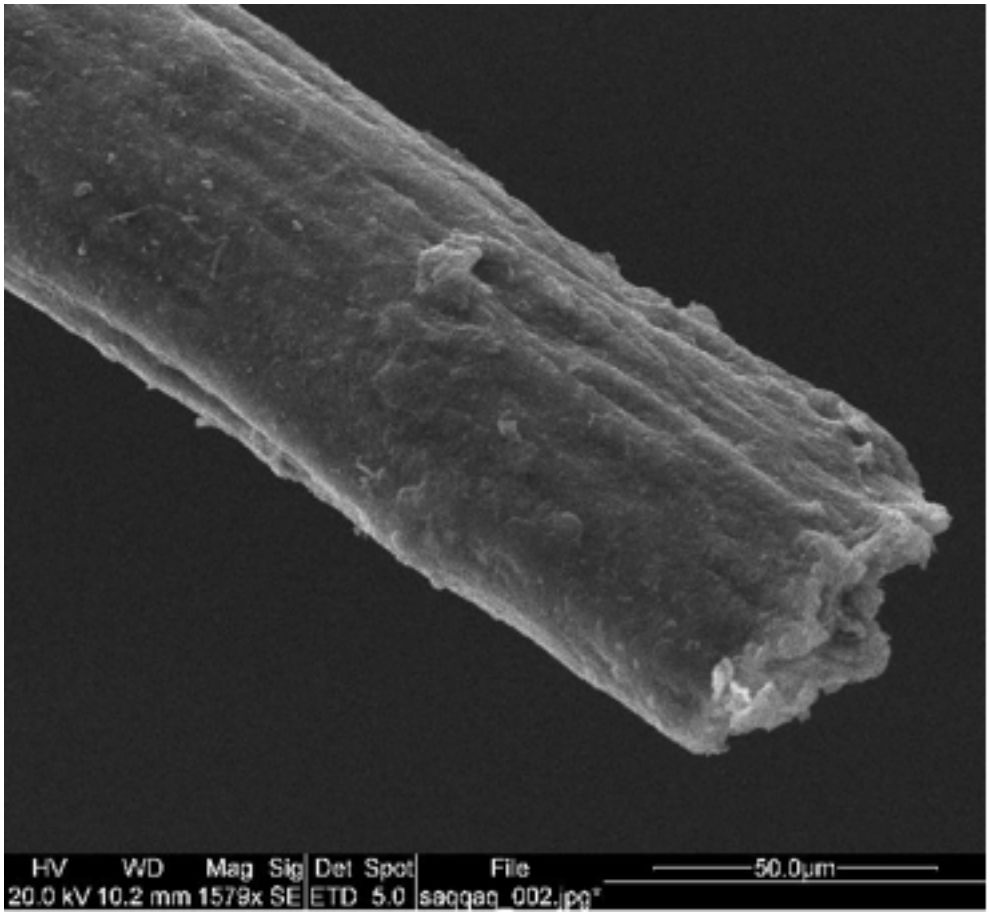


Fig. S7. Scanning Electron Microscopy image showing that samples had undergone post-depositional alteration and that there was no clear evidence of a surviving cuticle.

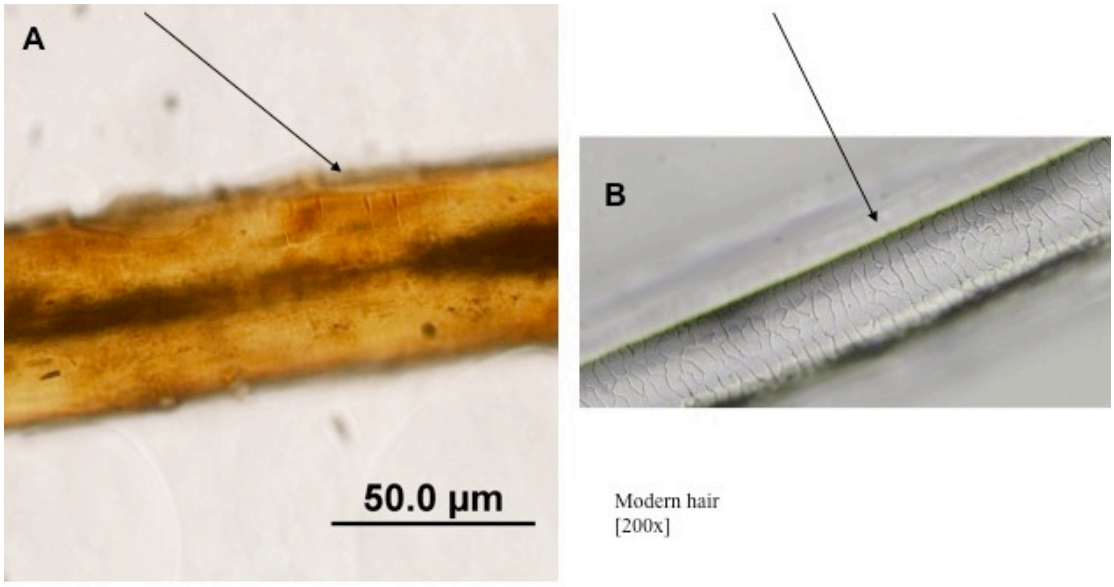


Fig. S8. The edges of the scale pattern from the modern hair are similar in appearance to the pattern present on marked site on the Saqqaq hairs, which appear to be remnants of the cuticle.

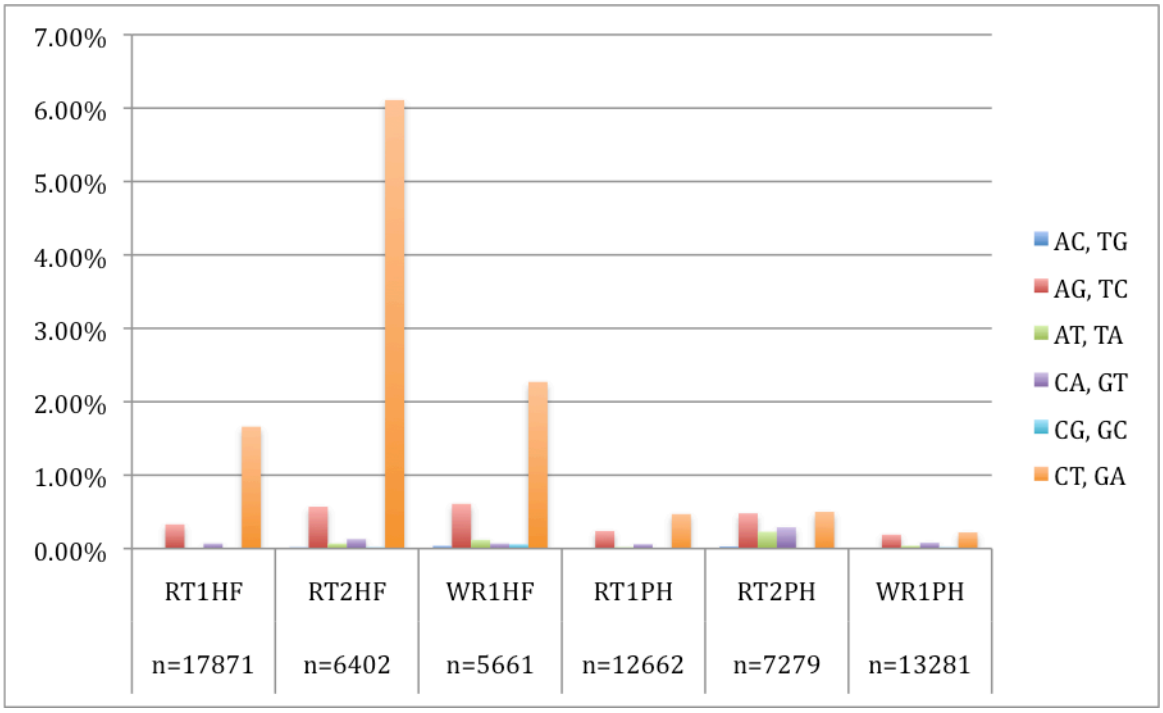


Fig. S9. Error rates for each of the 6 complementary pairs of errors. Numbers of sequences matching the 6 different tags are shown below each individual/PCR. RT, Rangifer Tarandus, WR, Woolly Rhino, HF, HiFi, PH, Phusion.

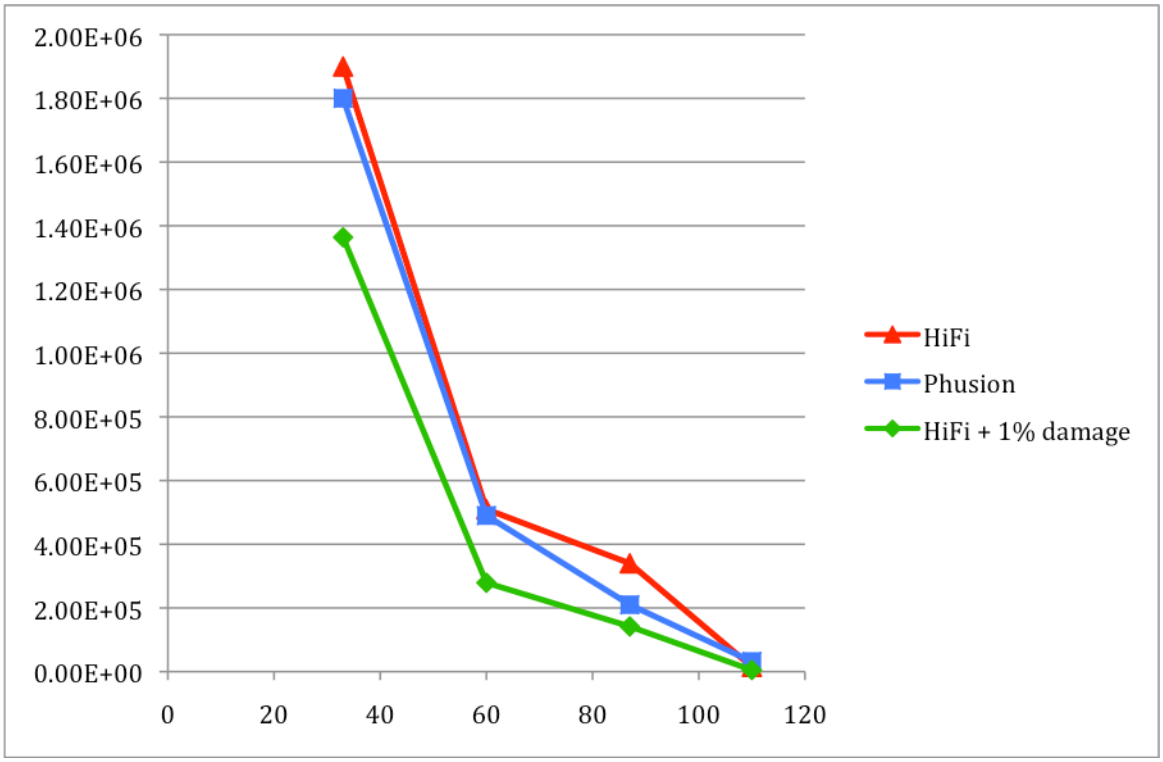


Fig. S10. Approximation of number of starting templates, for each enzyme at different lengths (excl primers). Damage is estimated as H*(1-D)^L, where H is HiFi measurements, D is damage rate (here 1%) and L is length of PCR product exclusive primers.

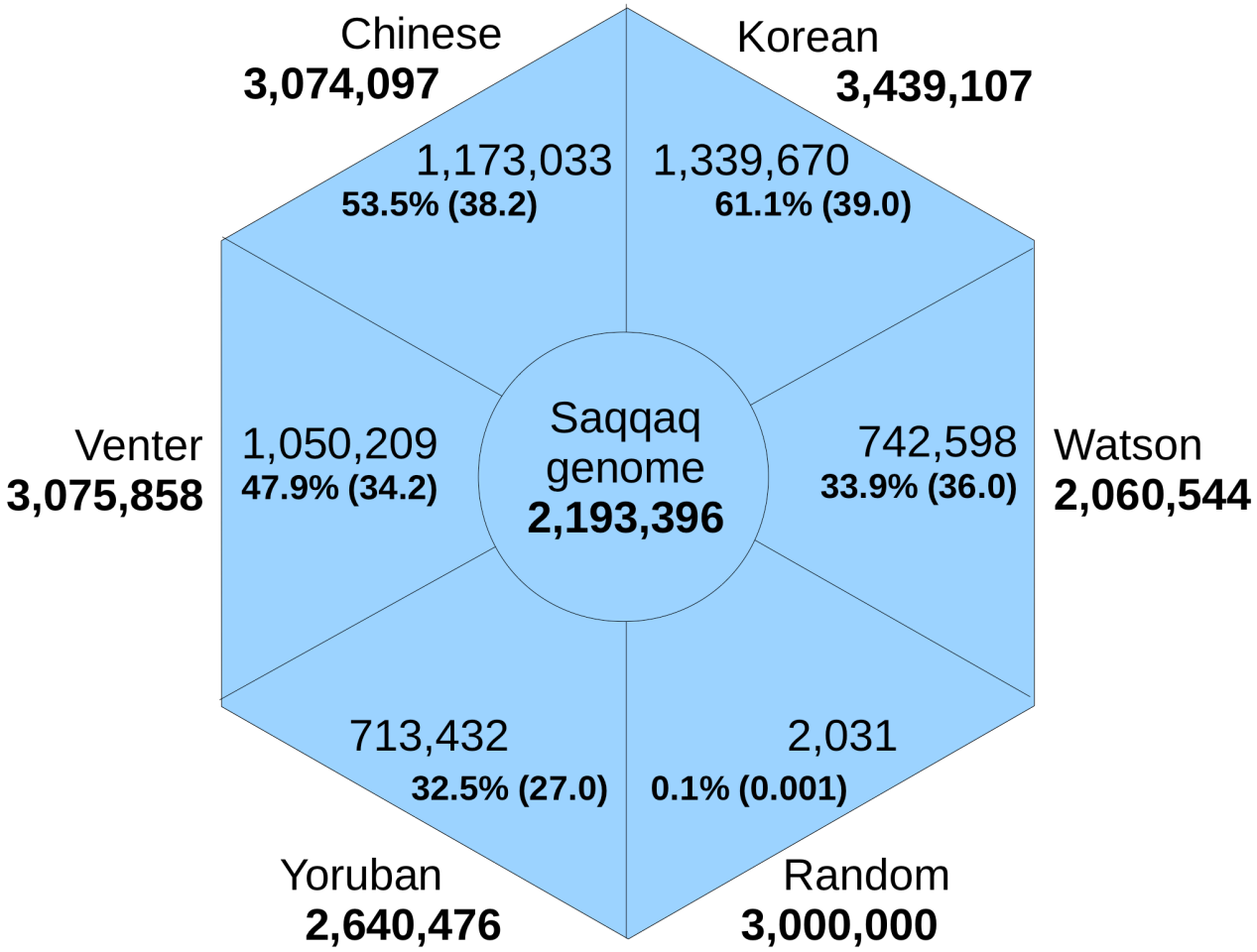


Fig. S11. The number of all SNPs overlapping between the Saqqaq genome and five different human genomes (Venter, Watson, Asian, Korean and Yoruban) and a random SNP vector. Values are given in total number of overlap, percentage relative to the Saqqaq genome and in parenthesis the percentage normalised to the total number of SNPs found in the other genome. The Saqqaq genome shares most SNPs with the Korean genome, tightly followed by the Asian genome.

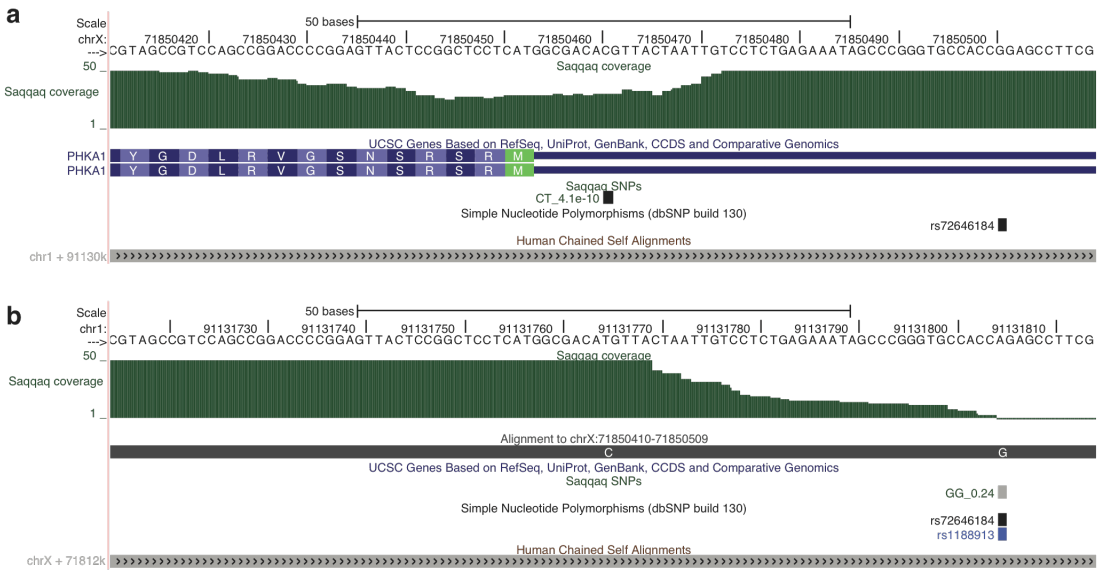


Fig. S12. Representative example of high confidence heterozygote SNP in chromosome X that can be attributed to the repetitive structure of the genome. a) Region of chromosome X overlapping the first coding exon of a gene (PHKA1), with high read depth, and a high confidence CT SNP call (note that one minus the posterior probability is given). The region shows strong homology (similarity) to a region on chromosome 1 (see Human Chained Self Alignments), depicted in (b). b) The homologous region on chromosome 1 is identical to the region shown in (a) as seen from the alignment (black), apart from two positions. The first of these corresponds to the heterozygote SNP call, where the reference has a T. Reads from this region that wrongly map to the region on chrX would thus cause nucleotide variation and a erroneous heteroxygote SNP call. The second deviating position corresponds to a known SNP in dbSNP (rs1188913). The Saqqaq genome is likely homozygote for G at the known SNP, explaining the low read depth in this region and the high read depth in the corresponding region on chromosome X. The figure was made using the UCSC Genome Browser [24].

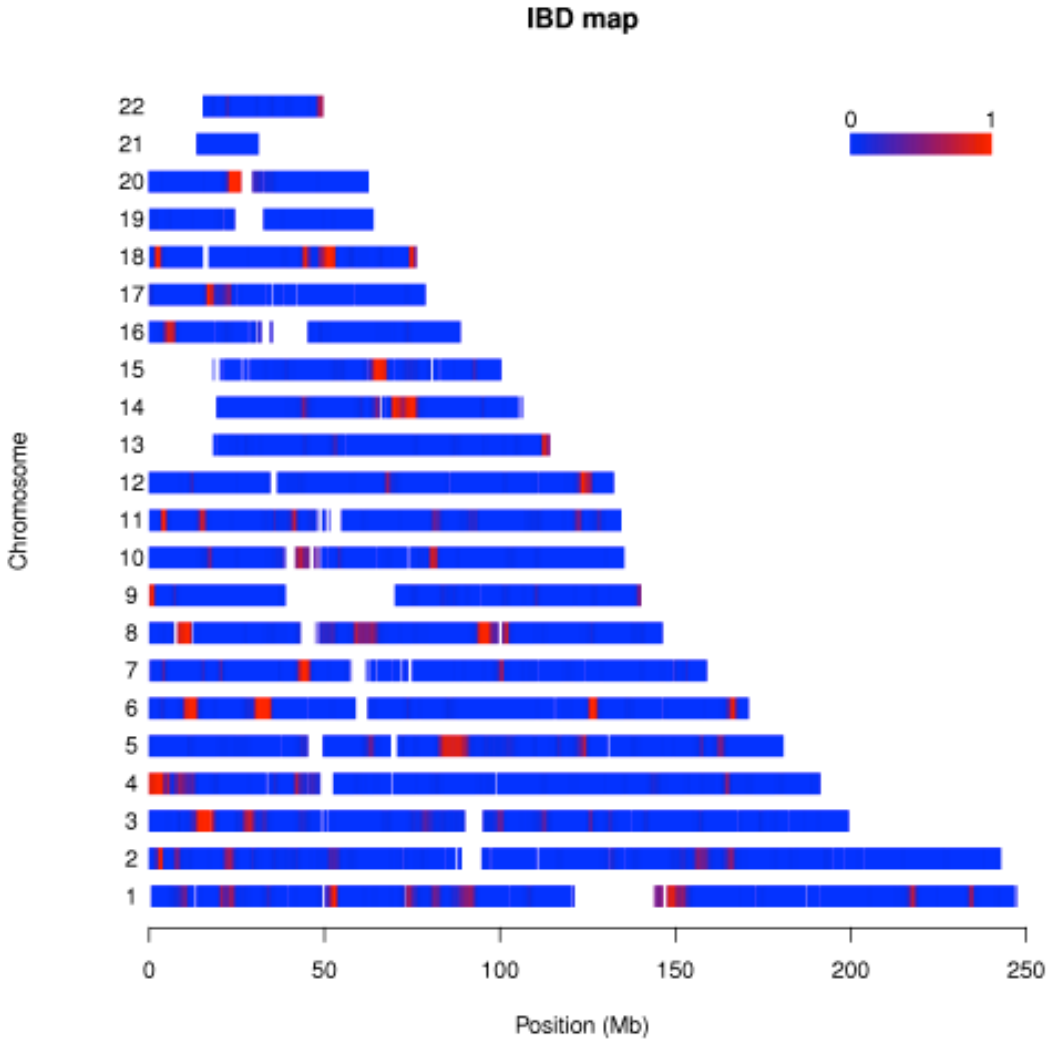


Fig. S13. Indentity by Descent across the genome. The probability of being inbred is shown for all high confidence genotypes for the Saqqaq genome. Only homozygous autosomal genotypes were used and the allele frequencies used in the estimation were obtained from the Siberian populations. An error rate of 1% was assumed.

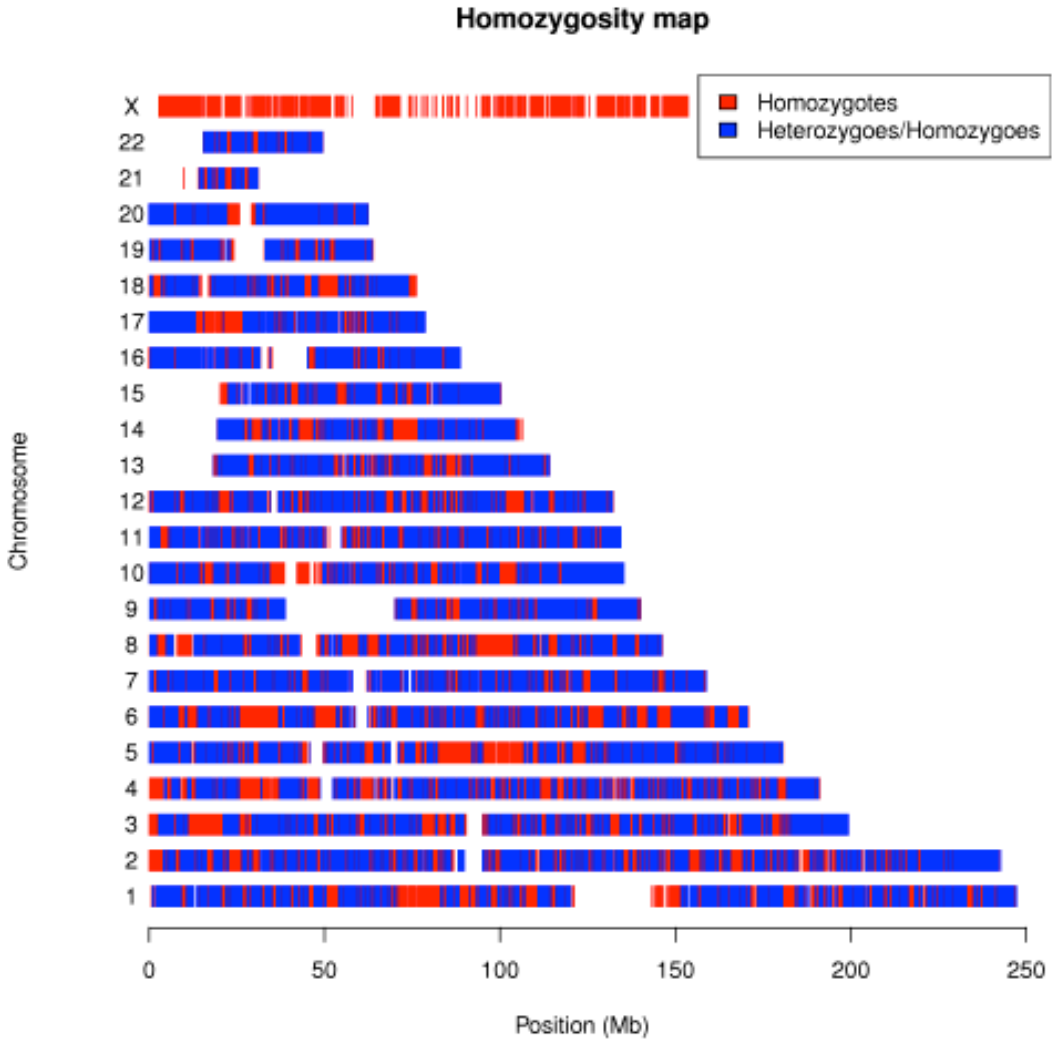


Fig. S14. Heterozygous and homozygous high confidence genotype calls are plotted across the genome. The heterozygous genotypes are super imposed on the homozygous calls. Thus the blue tracts contain both heterozygous and homozygous genotype calls while the red tracts contain homozygous genotypes exclusively.

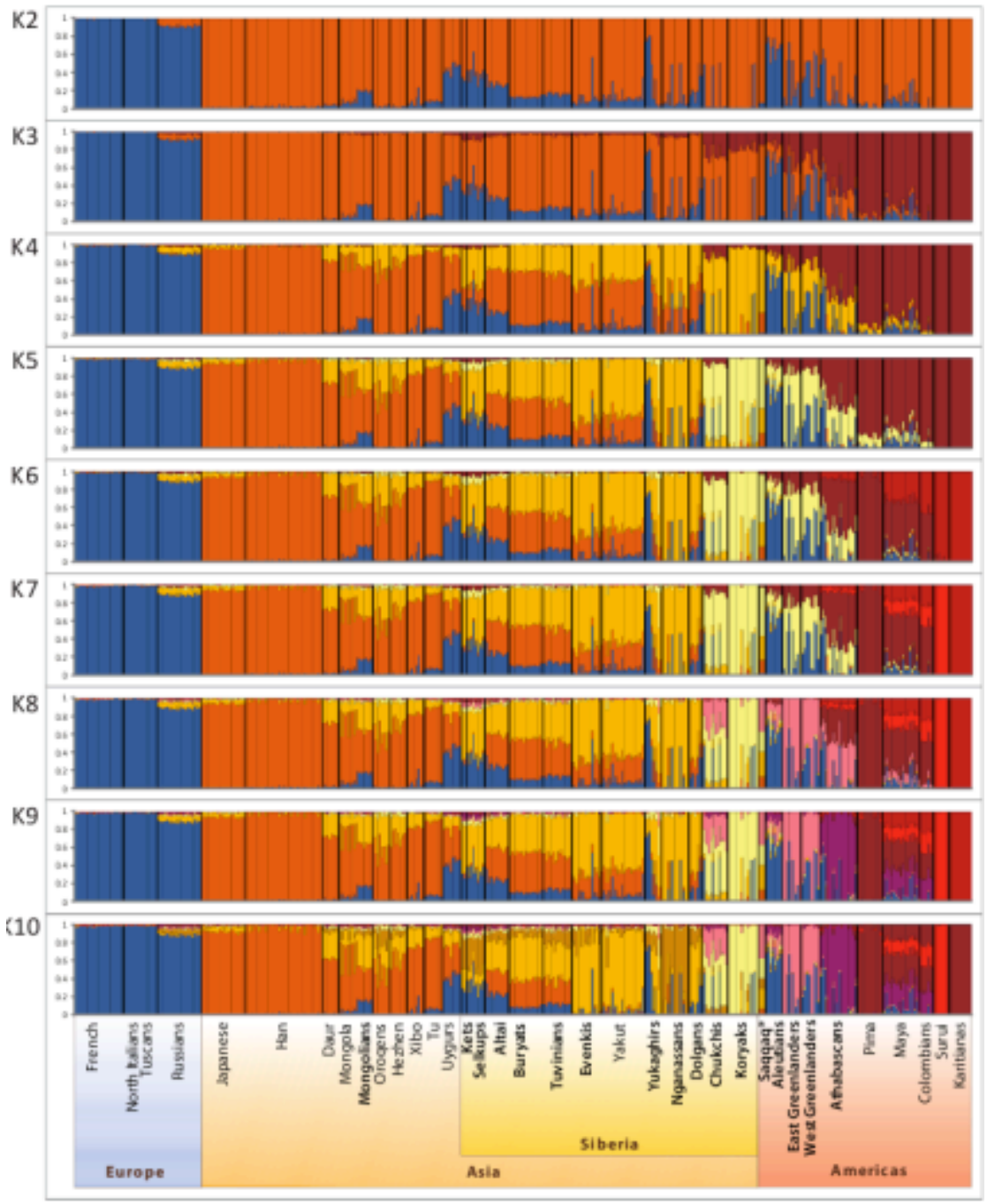


Fig. S15. Individual ancestry proportions of the studied individuals at K = 2 to K = 10. From the 100 runs of the *Admixture* [49] program the runs that yielded the highest Log-likelihood scores are plotted.

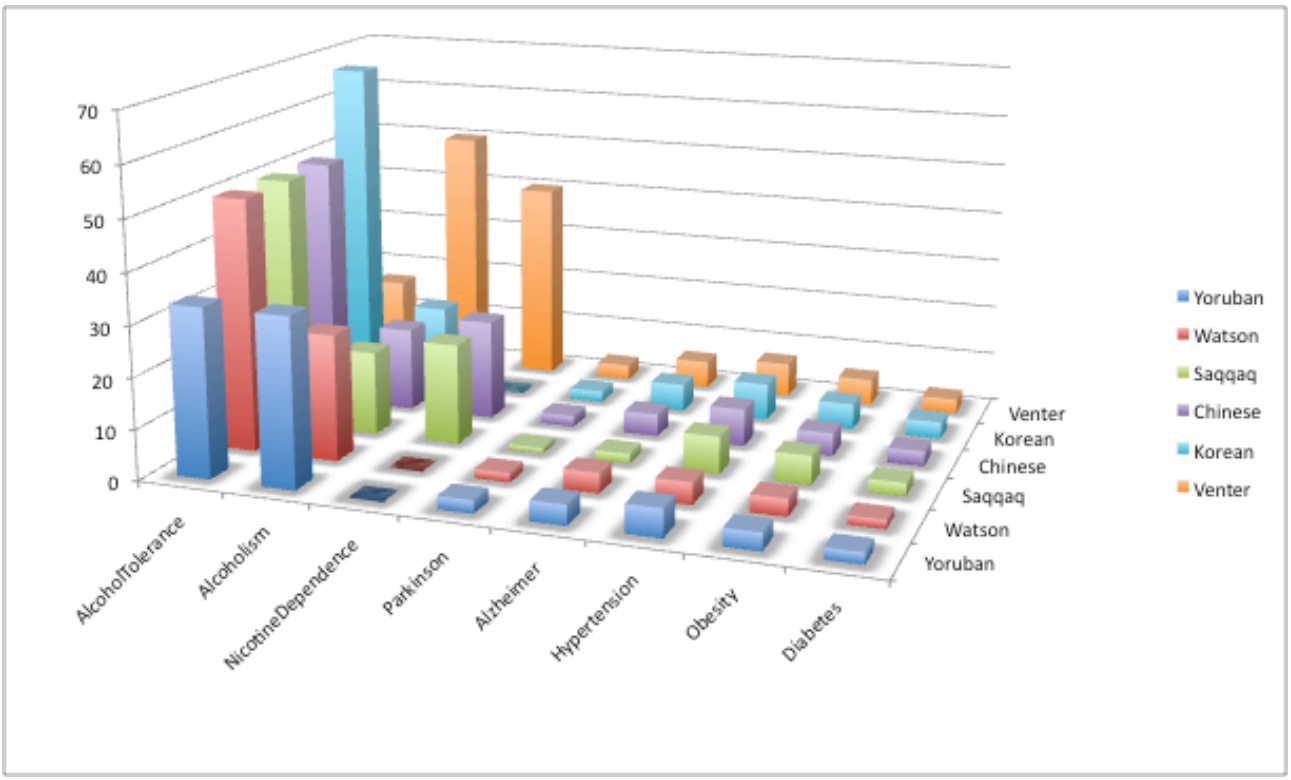


Fig. S16. Comparison of alleles identified that increase the risk to specific complex diseases. The curated set of SNPs associated with complex diseases was extracted from the Human Gene Mutation Database Professional version 2009.2 (30/6/2009) and extracted from literature where experimental evidence was available. The height of the bars show the number of gene-associated SNPs found in each genome.