A 3,000-year-old Egyptian emmer wheat genome reveals dispersal and domestication history

Michael F Scott[1*†], Laura R Botigué[2*], Selina Brace[3], Chris Stevens[4], Victoria E Mullin[3], Alice Stevenson[4], Mark G Thomas[1,5], Dorian Q Fuller[4], Richard Mott[1†]

1 Genetics Institute, Research Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK

2 Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Campus UAB, Bellaterra, 08193 Barcelona, Spain

3 Department of Earth Sciences, Natural History Museum, London, SW7 5BD, UK

4 Institute of Archaeology, University College London, 31-34 Gordon Square, London WC1H 0PY, UK

5 Research Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK.

* These authors contributed equally to this work.

† Corresponding Authors: m.f.scott@ucl.ac.uk, r.mott@ucl.ac.uk

Keywords: archaeobotany, ancient DNA, museum specimens, emmer wheat

Tetraploid emmer wheat (*Triticum turgidum* subsp. *dicoccon)* is a progenitor of the world's most widely grown crop, hexaploid bread wheat (*T. aestivum*), as well as the direct ancestor of tetraploid durum wheat (*T. turgidum* subsp. *turgidum*). Emmer was one of the first cereals domesticated in the old world, cultivated from around 9700 BCE in the Levant[1,2] and subsequently in South-Western Asia, Northern Africa, and Europe with the spread of Neolithic agriculture[3,4]. Here we report whole genome sequence from a museum specimen of Egyptian emmer wheat chaff, ^{14}C-dated to the New Kingdom 1,130 – 1,000 BCE. Its genome shares haplotypes with modern domesticated emmer at shattering, seed size, and germination loci, and within other putative domestication loci, suggesting these traits share a common origin prior to emmer's introduction to Egypt. Its genome is otherwise unusual, indicating genotypes potentially lost among modern emmer. Genetic similarity with modern Arabian and Indian emmer landraces connects ancient Egyptian emmer with early South-Eastern dispersals, while inferred gene flow with wild emmer from the Southern Levant signals a later connection. Our results show the importance of museum collections as sources of genetic data to uncover the history and diversity of ancient cereals.

Ancient DNA sequences can reveal dispersal and domestication histories. In crops, exome sequencing of barley[5] from 4000 BCE suggests population continuity for Southern Levantine barley. In maize, targeted capture and whole genome sequencing of samples from up to 4000 BCE showed that some domestication alleles were not yet fixed[6–8]. Sorghum genomes as old as 195CE reveal a decline in genetic diversity since this date[9]. Thus far in wheat,

archaeogenetic studies have primarily targeted single genes, such as glutenin[10,11]. Despite their importance, we are not aware of any analyses of whole genome sequence from ancient wheat.

The quintessential cereal domestication trait is non-shattering; spikes retain their seeds, making harvest easier but hindering natural dispersal[12]. Variation for shattering in emmer wheat is largely explained by quantitative trait loci (QTL) on chromosomes 3A and 3B containing homologues of the barley brittle rachis gene[13–15]. Loss of dormancy and increased seed size are also associated with domestication[16–18]. In wild emmer, each spikelet contains two seeds, one of which is smaller and remains dormant for over a year after shedding[19]. The paired grains in domesticated emmer and durum spikelets are the same size and germinate readily[17]. Both traits are controlled by a QTL on chromosome 4B[20].

Archaeobotanical evidence suggests that emmer was domesticated over several millennia, with non-shattering fixed across the Northern and Southern Levant from 6300BCE[1,3,21]. Emmer was cultivated in Egypt from the earliest settlements (5500-4500 BCE)[22,23]. Hulled emmer wheat is harder to process than free-threshing durum and bread wheats but may be preferred for cultural reasons, hardiness, or because the grain is better protected from pests during storage[24]. Free-threshing tetraploid durum and hexaploid bread wheats were increasingly cultivated in Egypt as cultural practices shifted following Alexander the Great's conquest in 332BCE[24,25]. Today, emmer cultivation is rare in Egypt but it remains an important crop in Ethiopia, Yemen, and parts of India[26].

We sequenced a museum specimen of emmer wheat and compared its genome against the wild emmer wheat reference and modern exonic variants[13], and addressed five questions: Do museum crop specimens – here stored in suboptimal conditions for decades – still contain useful endogenous DNA? What are the biogeographical relationships between ancient Egyptian and modern emmer wheat, and the likely history of its dispersal to and from Egypt? Does this sample contain haplotypes absent from sequenced modern accessions, due to either incomplete sampling of the modern emmer gene pool or the historical loss of alleles? Is there evidence of gene flow from wild emmer? Lastly, does ancient Egyptian emmer preferentially resemble modern domesticated accessions at loci associated with domestication?

We obtained whole genome sequence from an accession of ancient Emmer wheat chaff (hereafter UC10164), which was collected by Brunton and Caton-Thomson from the Hememiah North Spur archaeological site in the 1920s[27] and is now stored at the UCL Petrie Museum of Egyptian Archaeology (Figure 1a). AMS ^{14}C dating of two seeds placed this accession in the New Kingdom's Late Ramesside period, Dyn. 20, 1,130 – 1,000 cal. BCE, representing a mature ancient Egyptian agricultural period (Extended Data 1). Two specimens were chosen for sequencing (hereafter S1 and S2), both of which had the non-shattering domestication trait based on visual inspection (Figure 1b). S1 had a high endogenous content: 66% of 861M reads were alignable to the Zavitan v2 modern wild emmer wheat reference genome[13], including ambiguous and duplicate alignments. S2 yielded lower endogenous content (only 33% of reads were alignable) and was sequenced to lower depth (59.3M reads), and hence was only used for comparison with S1.

[FIG 1 ABOUT HERE]

In order to mitigate variant-calling errors arising from low complexity regions, we attempted to call genotypes only at 1.6M Single Nucleotide Polymorphism (SNP) sites segregating among 64 modern wild and domestic emmer accessions, identified using exome capture[13]. We obtained 0.48x coverage of these SNP sites after excluding sequences <35bp, alignments with mapping quality scores less than 30, and duplicate alignments. We then applied further quality control filters to mitigate biases from the alignment of short reads[28] and required at least two alignments to cover each site in S1, resulting in 99,078 called SNP genotypes, on which we based our analyses.

Multiple lines of evidence indicate these data are from ancient material and are reliable: First, the small fragment sizes and the deamination patterns, assessed using MapDamage[29], are characteristic of authentically ancient DNA (Extended Data 2). After trimming bases potentially affected by deamination, UC10164 does not show an excess of deamination-related substitutions, falling within the distribution of the modern samples (Supplementary Table 5). In addition, to rule out contamination by modern hexaploid bread wheat (*T. aestivum*), we assessed cross-mapped alignments against the bread wheat reference[30]. 99.4% of S1 alignments with mapping quality scores of at least 30 were to the A or B subgenomes, which derive from emmer wheat (Extended Data 2) and only 0.6% to the D genome which is absent from emmer. 7.6% of S2 alignments were to the D subgenome. We called low confidence genotypes from S2 without filtering on coverage depth and found 184 sites that overlapped with S1, of which 172 (93.5%) matched the genotype calls from S1.

Thus, concordance between S1 and S2 is higher than between S1 and any modern accession (mean 80.7%, sd 2.6%, maximum 87.4%).

The number of heterozygous calls in UC10164 (S1) was consistent with emmer's outcrossing rate of less than 1%[31]. Among modern accessions, heterozygosity at called sites was between 0.4% and 4.1% (mean 1.4%, sd 0.6%). We estimated heterozygosity in UC10164 to be 1.2%, using only the 3,160 SNP sites covered by four sequences. Finally, as described below, we found large genomic segments, sometimes extending over tens of Mb, over which UC10164 shares a haplotype with one or more wild and/or domesticated modern emmer accessions. That long haplotypes of UC10164 are almost identical to a modern accession suggests our genotype calls are accurate.

We next sought to place UC10164 in its historical dispersal context, specifically its relation to the early Eastwards dispersal of emmer into the Indus valley by ~6000 BCE and the Southern Arabian peninsula by 2500 BCE[32,33]. Emmer was introduced in the Nile valley from around 4500 BCE probably via the Southern Levant[24] (Extended Data 1). Modern wild emmers divide into two subgroups (Northern Levant and Southern Levant) and modern domesticated emmer into four genetic subgroups of domesticated emmer (Mediterranean, Caucasus, Eastern European, and Indian Ocean)[13].

Based on the 99k SNPs, we used several methods to map genome-wide similarity between UC10164 and modern accessions, which all confirm that UC10164 is genetically closest to domesticated accessions and specifically to the domesticated Indian Ocean subgroup. First, identity by state at these SNPs shows UC10164 is most concordant (86.4%-87.4%) with the

Indian Ocean subgroup, compared to other domesticated accessions (mean 81.7%, sd 1.8%) or wild accessions (mean 79.1%, sd 1.6%), Supplementary Table 2. Second, UC10164 is closest to the Indian Ocean accessions in a Principal Components Analysis (PCA) visualization of genetic similarity among accessions, Figure 2b. Third, phylogenetic analysis (Figure 3a) shows UC10164 branches closest to (but not within) the Indian Ocean clade. Finally, ADMIXTURE[34] source population inference indicates that UC10164 shares most ancestry with the Indian Ocean subgroup (Extended Data 3b). The Indian Ocean subgroup consists of one accession from each of Oman and India, and two from Turkey (Figure 2a). Thus, UC10164 resembles domesticated emmers that dispersed to the East and South from the Levant.

[FIGURE 2 ABOUT HERE]

Overall, the genotype of UC10164 is distinct from modern emmers. It is most concordant (87.4%) with the modern accession PI319868, from the Indian Ocean subgroup, Supplementary Table 2. Only one modern domesticate is less concordant with other accessions (PI352347, with maximum concordance of 86.6% with PI355454, both of which are in the Mediterranean subgroup; Extended Data 4). UC10164 falls closest to, but outside, the cluster of Indian Ocean accessions on the three main principal components (Figure 2b) and it is outside their phylogenetic clade (Figure 3a). Furthermore, UC10164 has many unique haplotypes, defined using sliding windows of 50 SNPs (Figure 2c). These might represent lost alleles or possibly incomplete sampling of modern emmer (few of the sequenced modern emmers are in the Indian Ocean subgroup, and none are from Africa).

Emmer cultivation all but disappeared in the Nile valley since the Roman era[35,36] and so it is likely much ancient Egyptian emmer diversity has been lost.

Compared to modern domesticates, UC10164 has high concordance with wild Southern Levant emmers (Extended Data 4), falls closer to them in the PCA (Figure 2b) and shares ancestry with them as inferred from ADMIXTURE source population inference (Extended Data 3b). We tested for gene-flow between wild Southern Levant emmer wheat and UC10164 using a four-population test[37]. As with our phylogenetic analysis (Figure 3a, Extended Data 3a), we constructed an outgroup genotype using reads from the diploid species *T. urartu* and *T. speltoides*, which are likely progenitors of emmer[38], to call variants on the A and B subgenomes, respectively. Then, using D statistics, we compared the frequency of derived alleles that each wild Southern Levant accession shares with UC10164 versus the domesticated Indian Ocean subgroup[37]. Five wild Southern Levant emmers share a significant excess of alleles with UC10164, with Z scores between 2.076-4.775 ($P<0.05$ against the null hypothesis that UC10164 and the Indian Ocean subgroup share an equal fraction of derived alleles with these Southern Levant wild accessions), calculated using block jackknife sizes ranging from 100kb-50mb, (Supplementary Table 3). We conclude that gene flow likely occurred between the ancestors of UC10164 and Southern Levant wild emmers (Figure 3b).

[FIGURE 3 ABOUT HERE]

We next asked whether UC10164 shared a common history of selection under domestication with modern domesticates. We initially focused on well-characterized QTL

for domestication traits. For shattering, QTLs on chromosomes 3A and 3B contain putative loss-of-function insertion/deletion mutations in the *TtBtr1-A* and *TtBtr1-B* genes[13]. At these QTLs, all domesticated accessions are very similar, while genetic diversity among wild accessions remains high (Figure 4), indicative of selective sweeps. From the size of these regions (4Mb and 5.5Mb), we estimated selection coefficients[39,40] in the ranges of 0.002-0.020 and 0.003-0.027 assuming crossover rates of 0.1-0.5cM/Mb[13], a selfing rate of 0.99-0.995[31]. A QTL on chromosome 4B has major effects on grain size and seed dormancy[20] and a 3Mb (509-512Mb) signal of a selective sweep among domesticated accessions, implying a selection coefficient in the range of 0.0007-0.0060 (assuming a lower crossover rate of 0.05-0.2cM/Mb in this region[13]). While the density of genotyped SNPs in the ancient specimen is lower, exaggerating the variance in minor allele frequency, at 98 of 99 SNPs called within these three QTLs, UC10164 has the allele carried by the majority of domesticated accessions (Figure 4b, Extended Data 5).

We then examined a wider set of loci previously hypothesised to have been selected during domestication in modern accessions[13], but with unknown phenotypic effects. We used the most extreme 5% of loci (2Mb sliding windows), as determined by domesticated:wild F_{ST} (n=505), domesticated:wild nucleotide diversity π_D/π_W (n=503), and Tajima's D in domesticated emmer (n=505). These comprise 1,155 unique loci after accounting for overlaps between selection scans. The concordance of UC10164 with domesticated accessions outside these loci resembled the genome-wide average, with a markedly elevated concordance between UC10164 and the four Indian Ocean accessions (Figure 4a). However, within the 1,155 loci, UC10164 is significantly ($P<0.001$, based on a locus randomization test) more concordant with all the modern domesticated than modern wild

accessions. Of the 1,155 loci, only seven overlap with selective sweeps for shattering, seed size, and germination. Thus, the shared selection history between UC10164 and modern domesticated emmers extends well beyond these well-characterized QTLs.

[FIGURE 4 ABOUT HERE]

Across loci associated with selection under domestication, UC10164 is enriched for alleles characteristic of modern domesticated emmer and at QTL for key domestication traits – shattering, seed dormancy and seed size – all modern domesticated emmers share a haplotype with UC10164. Therefore, selection at these loci probably occurred during domestication in the Near East, between 9700 and 6300 BCE and prior to the introduction of emmer into Egypt (5500-4500 BCE), consistent with archaeological evidence[21,41].

UC10164 most closely resembles modern domesticated emmer from India, Oman, and Turkey. RFLPs and karyotypes[42,43] suggest modern landraces from Yemen and Ethiopia should also resemble UC10164. Our data indicate a connection between early eastward and southward dispersals of emmer, distinct from northward and westward dispersals. This connects the arrival of cereal agriculture across the Iranian Plateau and into the Indus valley by 6000 BCE with its dispersal into the Nile Valley around 4500 BCE[32,33,44], and then into northern Sudan within a few centuries[45] (Extended Data 1).

We found evidence of gene flow between wild Southern Levant emmers and ancient Egyptian emmer, possibly during cultivation within the range of wild emmer prior to its introduction to Egypt[24,25], or during a later period of Egyptian interaction with or occupation

of the Levant (e.g., 1300-1185 BCE)[46]. Hybridization between modern wild and domesticated emmer growing together in the Southern Levant has been proposed[42,43,47,48], and similar signals are found in this study ('Mediterranean' accessions PI355454 and PI52347, Extended Data 3b). These results highlight the geographically extensive nature of wild progenitor contributions to crop diversity[47,48].

Emmer wheat cultivation in Egypt was extensive in antiquity but has since dramatically declined[24]. It is therefore not unexpected that UC10164 is relatively distinct compared to sequenced modern emmer. This may represent incomplete sampling (particularly from Africa and Arabia) or extinct ancient variation. As has been recently suggested[49], ancient alleles lost in modern domesticates might be targeted for re-introduction from the wild to boost crop improvement.

Wheat genomic resources[13,30] now permit the analysis of whole-genome sequence data, in this case from emmer chaff harvested over 3,000 years ago. Genomes of older, Neolithic, wheats might determine when domestication alleles accumulated and/or pinpoint gene flow events between crops and wild relatives[48]. Importantly, material excavated about 90 years ago and since then stored without climate control can yield usable DNA, which accentuates the great potential of archaeobotanical museum specimens for genetic analysis.

Online Methods

Plant material, radiocarbon dating and morphological analysis

The emmer wheat samples were excavated from the Hememiah North Spur archaeological site in Egypt from 1921 onwards[27]. The plant material in this study was collected from the west side of the site. Even though the seeds were initially ascribed to the Predynastic Badarian period, the authors report intrusive burials from the Old Kingdom at the site. Furthermore, it is doubtful that extensive cultivation occurred in this area in the Badarian[44]. The plant material was originally identified as emmer wheat by Dr. John Percival and subsequently confirmed by Dorian Fuller and Chris Stevens in the UCL archaeobotany laboratory.

The accession used in the present study comprises uncarbonized chaff stored at the Petrie Museum of Egyptian Archaeology, University College London, under collection number UC10164 (Figure 1a). Two samples were chosen (S1 and S2) (Figure 1b), both of which have rough disarticulation scars visible below and above the internode (Figure 1b). These are the most reliable diagnostic elements for domesticated forms of emmer wheat[50] and indicate that the specimens came from ears that did not readily shatter. Replicated Accelerator Mass Spectrometry ^{14}C dating was performed on two further specimens from the accession at the Beta Analytic Inc. laboratory in Miami, Fl. USA and the results are given in Supplementary Table 4. The two-sigma calibration of the combined dates is 1130-1000 Cal. BCE as calculated in OxCal v.3.10[51] using IntCal13[52].

Archaeobotanical database

The archaeobotanical evidence for the occurrences of emmer wheat over time in the Middle East, Egypt and around the Indian Ocean was compiled from the Old World Crops Database generated at UCL as part of the European Research Council research project on "Comparative Pathways to Agriculture" (ERC #323842). These data track the dispersal of emmer wheat eastwards from the Levant prior to the occurrence of these cereals in the Nile Valley[33]. The distribution of emmer based on median archaeological ages is shown in Extended Data 1.

DNA extraction and sequencing

DNA extraction was carried out within the UCL clean-room ancient DNA research facilities, in which no previous extractions of wheat had been performed. Sample S1 was briefly immersed in 0.5% bleach whereas sample S2 was not, both samples were thoroughly rinsed and crushed to powder. They were then immersed in 2% CTAB buffer (containing 1% PVP) and incubated at 37°C for 6 days. DNA was extracted first with chloroform:isoamyl-alcohol 24:1 and purified using the DNEasy Plant Mini Kit (Qiagen) with a single modification to binding buffer amount (3x elutant volume as opposed to suggested 1.5x) and second with 300 μl of acetone instead of AW2 to reduce material loss.

Genomic sequencing libraries were prepared at the Natural History Museum, London, as in[53], and partially UDG treated as described in[54]. UDG treatment removes nucleotide misincorporations that are associated with aDNA. After partial UDG treatment, a small fraction of these misincorporations should be retained in the terminal nucleotides of each

fragment, which can be used to authenticate the ancient origin of the DNA. The same library preparation procedure was performed for extraction and library preparation controls. Tapestation and Qubit results indicated that there were negligible amounts of DNA present in these controls. Sequencing was performed at the UCL Institute of Neurology High Throughput Sequencing centre on an Illumina NextSeq 500. Controls were spiked into a sequencing lane for an unrelated pooled library. Samples S1 and S2 were barcoded and initially pooled with other libraries, yielding 2x77million and 2x54million 75bp reads, respectively. Sample S1 was then re-sequenced twice without pooling, yielding a further 2x378million and 2x361million 75bp reads (2x861million reads in total). Sequencing and alignment statistics are summarised in Supplementary Table 1.

Alignment

Adapters were removed from the raw reads using AdapterRemoval v2.2.2[55] with options --trimns and --trimqualities to remove low quality and ambiguous bases from the ends of reads, and to collapse overlapping paired reads into a single sequence for each read pair. We only used sequences that could be successfully collapsed for further analysis and discarded sequences <20bp. The length distribution of the resulting 649M S1 sequences and 46M S2 sequences is shown in Extended Data 2b. From the library preparation control, only 14 sequences (3.3%) remained after adapter removal.

These sequences were aligned to the Zavitan v2 reference genome for emmer wheat[13] using bwa aln (with options -l 16500, -n 0.01 -o 2, which set the number of seeds, maximum fraction of missing alignments, and maximum number of gap opens, respectively) and bwa samse[56], as in[57]. Patterns of nucleotide mis-incorporation, particularly driven by cytosine

deamination, are often used to authenticate the ancient origin of next-generation sequencing reads[29,58]. After partial UDG treatment, we expect a small excess of thymine base-calls relative to the Zavitan wild emmer wheat reference genome at the 5' end and a small excess of cytosine base-calls at the 3' end. We confirmed this expected pattern in the first and last 2 bp of the sequence fragments using MapDamage v2.0 software[29], which also shows that the fragment interior is negligibly affected, see Extended Data 2a.

We trimmed the first and last 2 bp of each read from the fastq sequence files using the fastX toolkit 0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/index.html) and re-aligned the reads to the reference. We used GATK[59] (v4.0.5.2) MarkDuplicates, which marked ~37% of all S1 alignments as duplicates that would later be ignored during genotype calling. Alignment statistics after these steps are summarised in Extended Data 2 and Supplementary Table 1.

For the controls, we obtained 2x422 75bp reads from the library preparation control only, of which only fourteen collapsed sequences remained after adapter removal. Only one sequence could be aligned to the emmer wheat reference genome. This alignment would have failed several of our quality control filters (see below): the mapping quality was 0 (threshold 30), the length was 24bp (threshold 35bp), and the alignments were to non-exonic regions. Visual inspection of this sequence indicated that it was likely to be adapter sequence that was not successfully removed. Given that the controls yielded negligible DNA and that the sequences that were obtained were predominantly from adapters, they do not provide evidence for biological contamination.

Genotyping

We only attempted to call genotypes at 1.6M exonic SNP sites previously identified as polymorphic among modern accessions by[13]. These exonic sites are relatively non-repetitive so the proportion of ambiguous alignments is much lower (Supplementary Table1 and Extended Data 2d). Furthermore, these positions were filtered on the basis of heterozygosity across samples to remove variant sites likely to be affected by read mis-alignment, e.g., between homeologs[60].

We used the quality control filters described in[28], which are designed to mitigate biases caused by the alignment of short aDNA fragments to a reference genome with a single allele. Briefly, all reads were realigned to a modified-reference genome, which had a randomly-chosen third nucleotide at each SNP site (neither the reference or non-reference SNP allele). In addition, all the reads aligned to SNP sites were modified such that they carried the other allele and then realigned. This step also required the removal of all alignments with gaps. We then discarded all alignments with a mapping quality score below 30. Reads were only retained for variant calling if they re-aligned to the same position after both modifications (modified reference genome and modified reads).

We called genotypes using GATK HaplotypeCaller in GENOTYPE_GIVEN_ALLELES and EMIT_ALL_SITES mode with --interval-padding 100, using only sequences of minimum length 35bp and maximum length 150bp (--min-read-length and --max-read-length). Calling a single random allele from the alignments at those 99k genotyped sites yielded the same proportion of non-reference alleles (21.94% using a random allele calling method and 21.90% using GATK).

S2 had higher levels of subgenome misalignment (Extended Data 2c), a lower percentage of endogenous alignments and was sequenced to lower depth. Therefore, genotypes called for S2 were only used for comparison with sample S1 and were not filtered further. We only used genotypes called from S1 for our main analysis and only included SNP sites with a coverage depth of at least two aligned sequences and a maximum of 35, filtered using bcftools[61] (v1.9). After these quality filters, we called 99,078 SNP genotypes.

Outgroup Construction

To construct an outgroup genotype at the SNP sites of interest, we downloaded short read sequences for *T. urartu* (SRR4010671 and SRR4010672, representing the emmer A subgenome outgroup) and *A. speltoides* (SAMEA2342530, representing the emmer B subgenome outgroup) from the ENA. We aligned 257M 2x250bp reads (*T. urartu*) and 1.1B 2x100bp reads (*A. speltoides*) to the emmer reference[13] using bwa mem[56]. SNP genotypes were called as above. VCFtools[62] (v0.1.15) was used to discard any SNP sites not on the A subgenome for *T. urartu* or not on the B subgenome for *A. speltoides*. The A subgenome calls from *T. urartu* and B subgenome calls from *A. speltoides* were then combined and SNP sites were discarded if they were covered by less than five or more than 130 reads. After filtering, we obtained an outgroup genotype at 932,461 SNP sites.

Transition / Transversion ratio

In order to determine whether there was still an excess of C > T and G > A substitutions caused by postmortem damage deaminations, we compared the Transition / Transversion (Ti/Tv) ratio between the ancient and modern samples. We oriented the SNP genotypes against the outgroup genotype, excluding calls that were heterozygous or missing from the

outgroup. The Ti/Tv ratio from the ancient sample (2.14) is the same as the average Ti / Tv of modern samples (2.07 – 2.20, mean 2.15, sd 0.030, Supplementary Table 5). The proportion of C > T and G > A substitutions in the ancient sample is 0.391, which also falls within the range observed in modern samples (0.374 – 0.395, mean 0.386, sd 0.0039), Supplementary Table 5.

Genome-wide population structure

Modern accessions were assigned to subgroups using the information from the phylogeny in[13]. We re-assigned two accessions (PI487264, from Syria, and Mt. Gerizim, from central Israel) from the Northern Levant to the Southern Levant subgroup. The re-assigned Mt. Gerizim accession is found within the range of the other Southern Levant subgroup accessions. Nevertheless, there is one Northern Levant subgroup accession that remains within the range of the Southern Levant accessions (PI428129, from central Lebanon), Figure 2a. Our re-assignment makes the Northern Levant clade monophyletic both in the original phylogeny[13] and in our phylogeny. Based on their locations of origin, we related these genetically-defined subgroups to subspecies groups defined by Vavilov from phenotypic/geographic information: abyssinicum Vav. (Indian Ocean), dicoccum (Mediterranean), and asiaticum Vav., within which there are convarieties serbicum (A. Schulz) Flaksb. (Eastern Europe) and transcaucasicum Flaksb. (Caucasus)[26,63].

We performed a Principal Components Analysis (PCA) using the 99k UC10164 SNP sites. VCFtools and PLINK[64] (v1.90b6.3) were used to prepare calls for import into R[65] (v3.5.1). Any calls that were missing in the modern accessions were replaced with the median call among modern accessions. We performed PCA using just the modern accessions using the R

prcomp() function. The genotype of UC10164 was then projected onto these PCs using the R predict() function. We repeated this procedure excluding modern wild accessions.

We further assessed population structure using ADMIXTURE v1.3.0[34]. Because the ADMIXTURE model does not explicitly consider linkage disequilibrium, we first thinned the genotype calls[34]. We used the --indep-pairwise 200 10 0.9 option in PLINK, which considers sliding windows of 200 SNPs, in steps of 10 SNPs, and removes SNPs that have an R^2 value of more than 0.9 with any other SNP in the window, which left 60,478 SNPs. We ran ADMIXTURE 50 times for K parameters varying from 2 to 7 and chose the run with the highest maximum likelihood. The cross-validation procedure implemented in ADMIXTURE (--cv) suggested that K=5 had the lowest cross-validation error (CV error values: 0.706, 0.671, 0.666, 0.665, 0.672, 0.676).

We used the SNPhylo pipeline[66] (version 20140701) to construct phylogenetic trees, first removing SNPs at which the minimum depth of coverage for any of the modern emmer accessions was less than two, leaving 41,425 SNPs. Next, we removed SNPs with a Minor Allele Frequency (MAF) less than 0.1 or more than 10% missing data, which left 13,105 SNPs. Finally, SNPs were pruned for Linkage Disequilibrium (LD) using a threshold of 0.1, leaving 5,431 SNPs. We then produced a maximum likelihood phylogenetic tree and performed a bootstrap analysis with 1000 bootstraps. To confirm that the overall structure of the resulting phylogenetic tree was not biased by the inclusion of U10164 and by restriction to the subset of SNPs that were called in U10164, we excluded UC10164 and repeated this analysis using the full set of SNP sites. We used the same filtering criteria as above, leaving 252,436 SNPs after filtering for coverage, 72,348 SNPs after filtering on MAF

and missingness, and 10,237 SNPs after pruning for LD. We replicated most nodes in Figure 3a, see Extended Data 3a. Notably, nodes that define the relationship between subgroups all replicated. Thus, this SNP-based method does not support a monophyletic clade of domesticated emmer wheats, as presented in[13]. However, in the phylogeny in[13], the node supporting the clade of domesticated emmer wheats has low bootstrap support (51) such that replication may not be expected.

We calculated D statistics using AdmixTools[37] (v5.1). Genotypes were converted to EIGENSTRAT format using the convertf function and then the qpDstat function was used to calculate the D statistics displayed in Figure 3. Standard errors were calculated using a block jackknife with a specified block size varied from 100kb to 5mb. The resulting standard errors are shown in Supplementary Table 3.

Haplotype structure

We calculated the fraction of similar haplotypes between each pair of samples at sites called in UC10164. We excluded sites at which the focal sample was heterozygous or missing and then analysed all possible overlapping 50-SNP windows moved in intervals of one SNP. We used a 95% threshold to classify haplotypes as 'similar' and confirmed this was representative threshold by repeating this analysis for thresholds of 90% to 99%, over which results were broadly the same. For example, among domesticated accessions, UC10164 has either the highest or second highest fraction of unique haplotypes across this range of thresholds, Figure 2c.

We also examined haplotype sharing by defining haplotype mosaics in UC10164. We excluded the 440 SNP sites at which UC10164 was heterozygous. We then used sliding windows of 50 SNPs (moved by steps of 25 SNPs) and plotted the fraction of calls that differ between the ancient Egyptian accession and each modern emmer wheat accession (Extended Data 5). We then determined the best estimate of the genomic mosaic carried by the ancient Egyptian accession in terms of the modern genomes, using a dynamic programming algorithm akin to the Viterbi path from a hidden Markov model. Our algorithm calculates a mosaic of genotypes from modern accessions that minimizes the number of differences from the genotype of the ancient Egyptian accession. To prevent inferring excessive mosaic breakpoints due to sequencing errors, the algorithm has a transition penalty for changing haplotype (equivalent to 2.5 SNP differences in the analysis shown). Furthermore, in some regions, UC10164 might carry a haplotype absent from any modern emmer wheat accession. Therefore, we introduced a dummy accession that differs from the ancient accession by a fixed threshold amount (7.5% in the analysis shown) at every site. That is, where UC10164 is inferred to carry this dummy haplotype, its average dissimilarity to any modern accession in the dataset exceeds this threshold across the region.

Known Domestication Loci

We examined two QTLs that were associated with shattering on chromosomes 3A and 3B[13]. By comparing the emmer reference with a durum wheat sequence, mutations in TtBtr1-A and TtBtr1-B that appear to cause loss of function were identified by [13]. Specifically, the domesticated allele of TtBtr1-A has a 2bp deletion and the domesticated 'Svevo' allele of TtBtr1-B has a 4kb insertion. In order to have a full non-shattering phenotype both

mutations need to be present as homozygotes. Insertions and deletions are difficult to detect using alignments of short reads obtained from ancient samples. Nevertheless, we examined the haplotype of UC10164 using the called SNPs in the region of these mutations, Figure 4.

We also examined a QTL associated with grain size and seed dormancy in which the causal gene is unknown[20]. In this case, flanking sequence for the markers either side of the QTL peak were obtained from cerealsDB (www.cerealsdb.uk.net/cerealgenomics/CerealsDB). We aligned these flanking sequences to the emmer reference using BLASTN[67] (version 2.6.0). As putative physical map positions for the variation detected in the mapping population, we considered hits to chromosome 4B only. Marker IWB72369 had three significant hits on chromosome 4B and is therefore ambiguously localised. We excluded the blast hit for IWB72369 that was outside the chromosome 4B blast hit for the next marker in the genetic map, IWB43529. This left two possible locations for IWB72369: the resulting physical map positions and plausible range for the QTL peak are plotted in Figure 4b.

Selective Sweeps

Within these QTL, there are regions where all domesticated accessions share a haplotype, indicative of a selective sweep. We defined selective sweep regions using 1Mb sliding windows, within which we required each domesticated accession to carry the 'domesticated allele' (major allele among modern domesticated accessions) at 95% of SNPs. For each QTL, we then chose the longest continuous region with high haplotype similarity. Within these loci, the fraction of 'non-domesticated' genotype calls carried by modern domesticated

accessions was 0.14%-1.18% (mean 5.0%, sd 0.28%). The ancient sample carries the 'non-domesticated' allele at 1/99 SNPs called within these loci.

We estimated selection coefficients s from the size of selective sweep regions as $s \approx \frac{cd}{0.01}$, which is a function of the recombination rate (c) and the distance between a selected site sites and hitchhiking neutral sites (d)[40,68]. Because neutral sites will hitchhike on either side of the selected locus, we used half of the length of the region that shows high similarity among all domesticated accessions to estimate d. We used a range of plausible recombination rate parameters (cM/Mb) taken from Figure S5 in[13], which we converted to recombination rates using Haldane's map function[69]. We then calculated the "effective recombination rate" (c^*) by adjusting a range of plausible selfing rates (η) by using equation 9.44 in[40], $c^* \approx c\left(1 - \frac{\eta}{2-\eta}\right)$, which was used to estimate s.

Concordance between Accessions

UC10164 tends to have elevated concordance with modern domesticated accessions (versus modern wild accessions) within regions identified as putative 'outliers' in tests of selection[13], Figure 4a. As a test statistic, we used the average difference in concordance with UC10164 between domesticated and wild emmer wheats. To assess significance, we used randomization procedure that retained the position of the outlier windows relative to one another. We 'circularized' the genome and then performed 1,000 permutations of the positions of the outlier windows by random rotation[70]. In all three types of outliers, the true average difference was the most extreme of 1000 replicates). Thus, within loci putatively

associated with selection under domestication, UC10164 is significantly enriched for the alleles that are present in modern domesticated accessions.

Data Availability

Sequence data are deposited in the ENA with the study accession number PRJEB31103. The genotype calls are also provided as source data to Figure 2. The database of archaeobotanical observations is provided as source data to Extended Data 1.

Author Contributions

L.R.B., M.F.S., R.M., D.Q.F., C.S., A.S., and M.G.T. designed and coordinated the study. M.F.S. designed and performed data analysis. L.R.B., S.B., V.E.M. performed experiments. C.S. obtained image data. M.F.S. and R.M. co-ordinated sequencing. D.Q.F. co-ordinated carbon dating. M.G.T. supervised access to the ancient DNA laboratory. D.Q.F., A.S., C.S. collated archaeobotanical data. M.F.S., R.M. wrote the manuscript. All authors have edited and approved the manuscript.

Competing Interests Statement

The authors declare no competing interests.

Corresponding Authors

Correspondence and material requests should be directed to M.F.S (m.f.scott@ucl.ac.uk) and R.M (r.mott@ucl.ac.uk).

Acknowledgements

M.F.S. and R.M. are supported by the RCUK BBSRC grant BB/M011585/1. R.M. is also supported by RCUK BBSRC grant BB/P024726/1. L.R.B. is supported by the Spanish Ministry of Economy and

Competitiveness Severo Ochoa Programme for Centres of Excellence in R&D 2016-2019 (SEV-2015-0533), and CERCA Programme, Generalitat de Catalunya. M.G.T. is supported by a Wellcome Trust Senior Research Fellowship, grant 100719/Z/12/Z. D.F. and C.S. are supported by the ERC ComPag project, grant #323842. V.E.M. is partially supported by the RCUK NERC Grant NE/P012574/1. We thank Yoan Diekmann, Delia O'Rourke and Anna Garnett for helpful discussions.

References

1. Arranz-Otaegui, A., Colledge, S., Zapata, L., Teira-Mayolini, L. C. & Ibáñez, J. J. Regional diversity on the timing for the initial appearance of cereal cultivation and domestication in southwest Asia. *Proc. Natl. Acad. Sci.* **113**, 14001–14006 (2016).
2. Fuller, D. Q., Willcox, G. & Allaby, R. G. Early agricultural pathways: Moving outside the 'core area' hypothesis in Southwest Asia. *J. Exp. Bot.* **63**, 617–633 (2012).
3. Fuller, D. Q. & Lucas, L. Wheats: origins and development. in *Encyclopedia of Global Archaeology* 7812–7817 (2014).
4. McClatchie, M. *et al.* Neolithic farming in north-western Europe: Archaeobotanical evidence from Ireland. *J. Archaeol. Sci.* **51**, 206–215 (2014).
5. Mascher, M. *et al.* Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley. *Nat. Genet.* **48**, 1089–1093 (2016).
6. Ramos-Madrigal, J. *et al.* Genome Sequence of a 5,310-Year-Old Maize Cob Provides Insights into the Early Stages of Maize Domestication. *Curr. Biol.* **26**, 3195–3201 (2016).
7. Vallebueno-Estrada, M. *et al.* The earliest maize from San Marcos Tehuacán is a partial domesticate with genomic evidence of inbreeding. *Proc. Natl. Acad. Sci.* **113**, 14151–14156 (2016).
8. Kistler, L. *et al.* Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. *Science* **1313**, 1309–1313 (2018).
9. Smith, O. *et al.* A domestication history of dynamic adaptation and genomic deterioration in sorghum. *Nat. Plants* **5**, 369–379 (2018).
10. Palmer, S. A., Smith, O. & Allaby, R. G. The blossoming of plant archaeogenetics. *Ann. Anat.* **194**, 146–156 (2012).
11. Bilgic, H., Hakki, E. E., Pandey, A., Khan, M. K. & Akkaya, M. S. Ancient DNA from 8400 Year-Old Çatalhöyük Wheat: Implications for the origin of neolithic agriculture. *PLoS One* **11**, 1–18 (2016).
12. Purugganan, M. D. & Fuller, D. Q. The nature of selection during plant domestication. *Nature* **457**, 843–848 (2009).
13. Avni, R. *et al.* Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**, 93–97 (2017).
14. Nalam, V. J., Vales, M. I., Watson, C. J. W., Kianian, S. F. & Riera-Lizarazu, O. Map-based analysis of genes affecting the brittle rachis character in tetraploid wheat

(Triticum turgidum L.). *Theor. Appl. Genet.* **112**, 373–381 (2006).
15. Pourkheirandish, M. *et al.* Evolution of the Grain Dispersal System in Barley. *Cell* **162**, 527–39 (2015).
16. Fuller, D. Q. Contrasting patterns in crop domestication and domestication rates: Recent archaeobotanical insights from the old world. *Ann. Bot.* **100**, 903–924 (2007).
17. Harlan, J. R., de Wet, J. M. J. & Price, E. G. Comparative Evolution of Cereals. *Evolution (N. Y).* **27**, 311–325 (1973).
18. Salamini, F., Özkan, H., Brandolini, A., Schäfer-Pregl, R. & Martin, W. Genetics and geography of wild cereal domestication in the near east. *Nat. Rev. Genet.* **3**, 429–441 (2002).
19. Horovitz, A. The soil seed bank of wild emmer. in *The Proceedings of International Symposium on In situ Conservation of Plant Genetic Diversity* (eds. Zencirci, N., Kaya, Z., Anikster, Y. & Adams, W. T.) 185–188 (Central Research Institute for Field Crops, 1998).
20. Nave, M., Avni, R., Ben-Zvi, B., Hale, I. & Distelfeld, A. QTLs for uniform grain dimensions and germination selected during wheat domestication are co-located on chromosome 4B. *Theor. Appl. Genet.* **129**, 1303–1315 (2016).
21. Allaby, R. G., Stevens, C., Lucas, L., Maeda, O. & Fuller, D. Q. Geographic mosaics and changing rates of cereal domestication. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, (2017).
22. Crawford, D. J. Food: Tradition and change in hellenistic Egypt. *World Archaeol.* **11**, 136–146 (1979).
23. Caton-Thompson, G. & Gardner, E. W. *The desert Fayum*. (Royal Anthropological Institute of Great Britain and Ireland, 1934).
24. Nesbitt, M. & Samuel, D. From stable crop to extinction? The archaeology and history of the hulled wheats. in *Hulled Wheats. Promoting the Conservation and Use of Underutilized and Neglected Crops.* (eds. Padulosi, S., Hammer, K. & Heller, J.) (1996).
25. Wetterstrom, W. Foraging and farming in Egypt: The transition from hunting and gathering to horticulture in the Nile valley. *Archaeol. Africa Food, Met. towns* 165–226 (1993).
26. Zaharieva, M., Ayana, N. G., Hakimi, A. Al, Misra, S. C. & Monneveux, P. Cultivated emmer wheat (Triticum dicoccon Schrank), an old crop with promising future: A review. *Genet. Resour. Crop Evol.* **57**, 937–962 (2010).
27. Brunton, G. & Caton-Thompson, G. *The Badarian civilization and predynastic remains near Badari*. (British School of Archaeology in Egypt: London, 1928).
28. Günther, T. & Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet.* **15**, 1008302 (2019).
29. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).
30. IWGSC. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018).
31. Golenberg, E. M. Outcrossing rates and their relationship to phenology in Triticum dicoccoides. *Theor. Appl. Genet.* **75**, 937–944 (1988).
32. Fuller, D. Q. Agricultural origins and frontiers in South Asia: A working synthesis. *J. World Prehistory* **20**, 1–86 (2006).
33. Stevens, C. J. *et al.* Between China and South Asia: A Middle Asian corridor of crop dispersal and agricultural innovation in the Bronze Age. *Holocene* **26**, 1541–1555

(2016).
34. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
35. van der Veen, M. *Consumption, Trade and Innovation: Exploring the Botanical Remains from the Roman and Islamic Ports at Quseir al-Qadim, Egypt*. *Journal of African Archaeology Monograph Series* **6**, (2012).
36. Murray, M. A. Cereal production and processing. in *Ancient Egyptian Materials and Technology* (eds. Nicholson, P. T. & Shaw, I.) 505–536 (Cambridge University Press, 2000).
37. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–93 (2012).
38. Marcussen, T. *et al.* Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345**, 1250092 (2014).
39. Olsen, K. M. *et al.* Selection under domestication: Evidence for a sweep in the rice waxy genomic region. *Genetics* **173**, 975–83 (2006).
40. Walsh, B. & Lynch, M. *Evolution and Selection of Quantitative Traits*. (Oxford University Press, 2018).
41. Fuller, D. Q., Lucas, L., Gonzalez Carretero, L. & Stevens, C. From intermediate economies to agriculture: trends in wild food use, domestication and cultivation among early villages in southwest Asia. *Paleorient* **44**, 59–74 (2018).
42. Badaeva, E. D. *et al.* Chromosomal passports provide new insights into diffusion of emmer wheat. *PLoS One* **10**, 1–25 (2015).
43. Luo, M. C. *et al.* The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. *Theor. Appl. Genet.* **114**, 947–959 (2007).
44. Wengrow, D., Dee, M., Foster, S., Stevenson, A. & Ramsey, C. B. Cultural convergence in the Neolithic of the Nile Valley: A prehistoric perspective on Egypt's place in Africa. *Antiquity* **88**, 95–111 (2014).
45. Fuller, D. & Hildebrand, E. Domesticating Plants in Africa. in *The Oxford Handbook of African Archaeology* (eds. Mitchell, P. & Lane, P.) 507–525 (Oxford University Press, 2013).
46. Hasel, M. G. *Domination and resistance : Egyptian military activity in the southern Levant, ca. 1300-1185 B.C.* (Brill, 1998).
47. Civáň, P., Ivaničová, Z. & Brown, T. A. Reticulated origin of domesticated emmer wheat supports a dynamic model for the emergence of agriculture in the fertile crescent. *PLoS One* **8**, e81955 (2013).
48. He, F. *et al.* Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nat. Genet.* **51**, 896–904 (2019).
49. Di Donato, A., Filippone, E., Ercolano, M. R. & Frusciante, L. Genome Sequencing of Ancient Plant Remains: Findings, Uses and Potential Applications for the Study and Improvement of Modern Crops. *Front. Plant Sci.* **9**, 441 (2018).
50. Zohary, D., Hopf, M. & Weiss, E. *Domestication of Plants in the Old World*. (Oxford University Press, 2012).
51. Bronk Ramsey, C. Bayesian analysis of radiocarbon dates. **51**, 337–360 (2009).
52. Reimer, P. J. *et al.* Intcal13 and marine13 radiocarbon age calibration curves 0-50,000 years cal bp. *Radiocarbon* **55**, 1869–1887 (2013).

53. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **6**, (2010) doi: 10.1101/pdb.prot5448.
54. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial uracil – DNA – glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20130624. (2015).
55. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 1–7 (2016).
56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
57. Meyer, M. *et al.* A high coverage genome sequence from an archaic denisovan individual. *Science* **338**, 222–226 (2013).
58. Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci.* **104**, 14616–14621 (2007).
59. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
60. Jordan, K. W. *et al.* A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol.* **16**, 1–18 (2015).
61. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
62. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
63. Vavilov, N. I. *Origin and geography of cultivated plants*. (Cambridge University Press, 1989).
64. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
65. R Core Team. R: A Language and Environment for Statistical Computing. (2016).
66. Lee, T. H., Guo, H., Wang, X., Kim, C. & Paterson, A. H. SNPhylo: A pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* **15**, 1–6 (2014).
67. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 1–9 (2009).
68. Kaplan, N. L., Hudson, R. R. & Langley, C. H. The Hitchhiking Effect Revisited. *Genetics* **123**, 887–899 (1989).
69. Haldane, J. B. S. The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet.* **8**, 299–309 (1919).
70. Cabrera, C. P. *et al.* Uncovering networks from genome-wide association studies via circular genomic permutation. *G3* **2**, 1067–75 (2012).

Figure Legends

Figure 1 (a) The accession (UC10164) of emmer wheat husks as stored (photo courtesy of the Petrie Museum of Egyptian Archaeology, UCL), and (b) the specimens that were used for sequencing with rough disarticulation scars, diagnostic of the non-shattering phenotype, circled.

Figure 2 The geographical (a) and genetic (b, c) relationships between UC10164 and modern accessions of emmer wheat (n=64). The filled areas in (a) designate the geographical area enclosed by the domesticated accessions in each subgroup, while the inset shows a zoom of the region from which most wild accessions originate. (b) The first three principal components are shown for PCA using all samples (left panels) and using the subset of domesticated samples only (right panels). UC10164 clusters with the modern domesticated emmers, which are all closer to wild Northern Levant emmers than they are to wild Southern Levant emmers. Of the modern domesticated emmers, UC10164 is closest to the Indian Ocean subgroup (green). (c) The fraction of haplotypes in each accession that are 'unique'. Haplotypes are defined using 50-SNP sliding windows. 'Unique' haplotypes have <95% genotypic similarity with all other accessions. (b) and (c) use the 99,078 SNP sites called in UC10164.

Figure 3 Phylogenetic analysis of UC10164 and modern emmer wheat accessions (n=64). (a) Maximum Lilkelihood tree with bootstrap support displayed on nodes where it is less than 100. (b) D statistics for each wild accession in the Southern Levant subgroup with the phylogeny: (outgroup, (Southern Levant accession, (Indian Ocean subgroup, UC10164))). Displayed standard errors were calculated using a jackknife with blocks of 5Mb. Positive D statistics indicate an excess of derived alleles are shared between the wild Southern Levant accession and UC10164. The red dashed arrow in (a) indicates a putative gene flow event that could lead to the observed pattern of derived allele sharing.

Figure 4 Haplotype sharing between UC10164 and modern domesticated accessions within loci associated with selection under domestication. (a) elevated concordance between UC10164 and domesticated accessions within regions that show putative signatures of selection under domestication, as defined by F_{ST}, domesticated-wild nucleotide diversity π_D/π_W, and Tajima's D, compared to other loci. (b) the minor allele frequency of all samples within 100 SNP sliding windows (moved in 50 SNP intervals) in the regions containing QTL identified for key domestication traits: shattering (chromosomes 3A and 3B) and grain size/seed dormancy (chromosome 4B). The position of putative loss-of-function mutations in TtBr1-A and TtBr1-B genes is labelled. The range on chromosome 4B shows the maximum extent of plausible positions found for the markers (72369 and 73477) that flank the QTL peak on chromosome 4B. All three cases contain regions in which all modern domesticated accessions (red) and UC10164 (black) have reduced diversity and carry the major allele, which is here defined relative to modern domesticated accessions.