nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-019-0921-3

The origins and adaptation of European potatoes reconstructed from historical genomes

Rafal M. Gutaker[1], Clemens L. Weiß[1], David Ellis[2], Noelle L. Anglin[2], Sandra Knapp[3], José Luis Fernández-Alonso[4], Salomé Prat[5] and Hernán A. Burbano[1*]

Potato, one of the most important staple crops, originates from the highlands of the equatorial Andes. There, potatoes propagate vegetatively via tubers under short days, constant throughout the year. After their introduction to Europe in the sixteenth century, potatoes adapted to a shorter growing season and to tuber formation under long days. Here, we traced the demographic and adaptive history of potato introduction to Europe. To this end, we sequenced 88 individuals that comprise landraces, modern cultivars and historical herbarium samples, including specimens collected by Darwin during the voyage of the *Beagle*. Our findings show that European potatoes collected during the period 1650–1750 were closely related to Andean landraces. After their introduction to Europe, potatoes admixed with Chilean genotypes. We identified candidate genes putatively involved in long-day pre-adaptation, and showed that the 1650–1750 European individuals were not long-day adapted through previously described allelic variants of the *CYCLING DOF FACTOR1* gene. Such allelic variants were detected in Europe during the nineteenth century. Our study highlights the power of combining contemporary and historical genomes to understand the complex evolutionary history of crop adaptation to new environments.

Potato (*Solanum tuberosum* L.) originated from the South American Andes[1] and spread globally during post-Columbian times. The establishment of potato cultivation in Europe represented a milestone in the widespread geographical success of this crop. Today, 82% of potatoes cultivated worldwide are grown in Eurasia. The earliest historical records of potato in mainland Europe date back to the late sixteenth century, in Spain[2]. While potato planting started in the seventeenth century, cultivation did not gain momentum until between the eighteenth and nineteenth centuries[3], making potato the main staple crop in Ireland[4] by the middle of the nineteenth century. The spread of potatoes was constrained not only by consumer acceptance but also by the distinct environmental conditions in Europe as compared to their original habitat. It is largely accepted that the post-domestication dispersal of crops outside of their native range required extensive adaptation to the new environments[5–7], in particular when crops were moved along latitudinal gradients[8]. In those instances, geographic expansion required an adjustment of plant development to different day length and temperature cues[9]. In the case of short-day-dependent Andean potatoes in Europe, tuberization would happen only in the short days of late autumn[10] (Fig. 1a), which are followed by freezing temperatures that kill the plants before proper storage of nutrients in the tubers is achieved. Hence, overcoming the short-day dependency for tuberization was presumably the most important adaptation to European conditions. Modern potato cultivars behave as facultative short-day plants since they tuberize in long days, but tuber differentiation is still accelerated following transfer of the plants to shorter day lengths.

To date, natural variation in genes controlling tuberization under long days has been identified only for a single gene, *CYCLING DOF FACTOR1* (*StCDF1*). The StCDF1 protein promotes tuberization through unblocking of the *SELF PRUNING 6A* (*SP6A*) pathway (Supplementary Fig. 1). Andean variants of StCDF1 are regulated in long days, resulting in StCOL1-mediated suppression of the tuberization pathway. Transposon-mediated truncations stabilize StCDF1, allowing tuberization independent of day length[11]. These truncated StCDF1 forms are gain-of-function repressing variants with a dominant effect on *SP6A* up-regulation, which ultimately results in tuberization under long days. Where and when these adaptive StCDF1 variants originated is not known. They could have arisen de novo in Europe, or from standing variation segregating in the potato's Andean native range. Andean potatoes also spread southwards in prehistoric times to the lowlands of south-central Chile, where they adapted to climatic conditions and long-day requirements similar to those in Europe[12] (Fig. 1a). This suggests the possibility that potato cultivation in Europe intensified after the introduction of, and potential introgression from, pre-adapted Chilean landraces. Indeed, based on chloroplast and microsatellite data, present-day European cultivars are genetically similar to Chilean landraces[13]. Furthermore, a study using a single chloroplast marker on historical herbarium specimens showed an increase in frequency of the Chilean chloroplast between the eighteenth and nineteenth centuries in Europe[3].

Results

Sampling and sequencing of historical and modern genomes. Here we used historical specimens to elucidate the origins and adaptation of potatoes across 350 years of their evolution in Europe, while increasing the power of our inference using genome-wide variants. We sequenced a total of 88 samples (Supplementary Tables 1 and 2) that included 29 historical herbarium specimens spanning the years 1660–1896, obtained from various European museums. This set included three Chilean historical samples (CHS) and 26 European historical samples (EHS). We also sequenced

[1]Research Group for Ancient Genomics and Evolution, Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tuebingen, Germany. [2]International Potato Center, Lima, Peru. [3]Department of Life Sciences, Natural History Museum, London, UK. [4]Departamento de Biodiversidad y Conservación, Real Jardin Botánico RJB-CSIC, Madrid, Spain. [5]Departamento de Genética Molecular de Plantas, Centro Nacional de Biotecnología-CSIC, Madrid, Spain. *e-mail: hernan.burbano@tuebingen.mpg.de

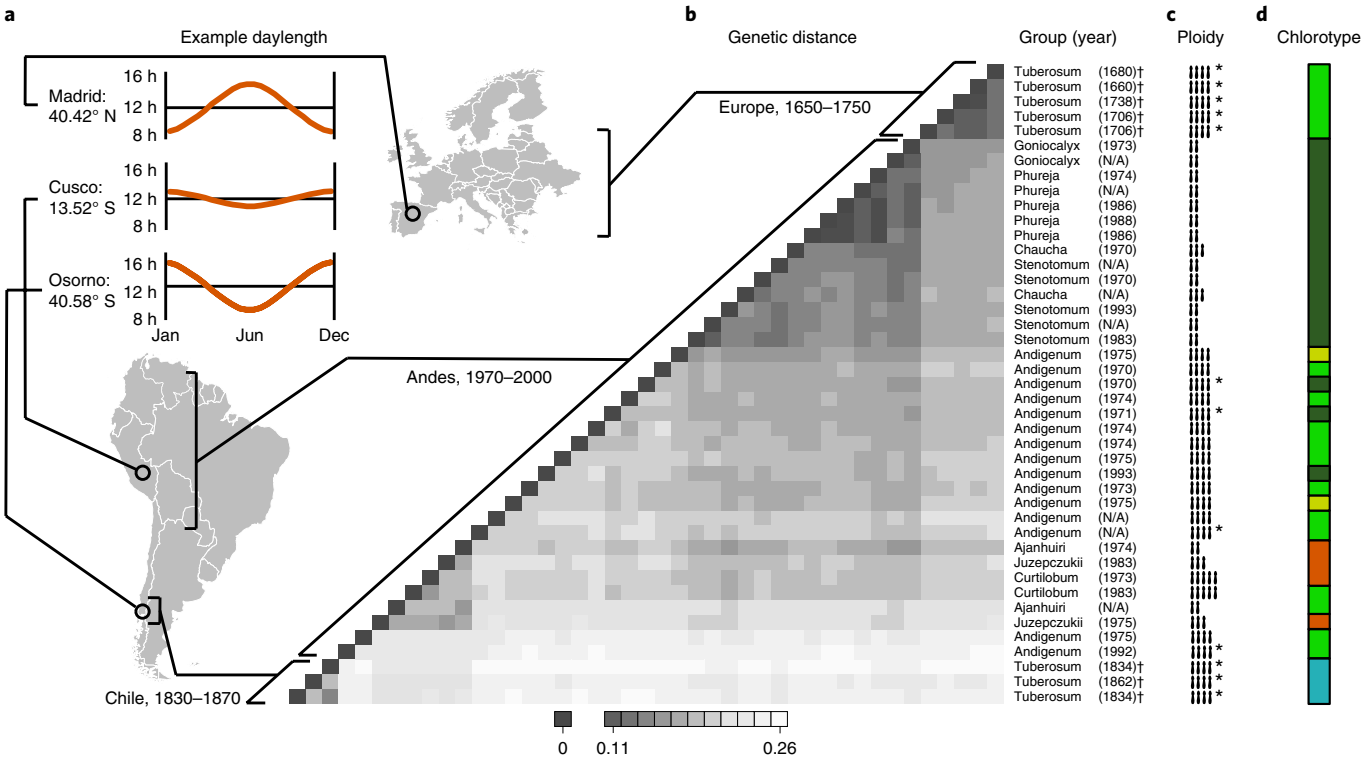

Fig. 1 | Relationship of the oldest European potatoes (collected 1650–1750) to historical and contemporary South American potatoes. a, Map of South America and Europe displaying cities representative of three latitudinal zones of potato cultivation. The graphs show annual day length fluctuation for each city. **b**, Heatmap illustrating pairwise genetic distances between South American landraces and the oldest European potatoes (annotated with potato group names, collection years and an obelus indicating historical herbarium specimen) based on 35,659 nuclear SNPs. The grey scale indicates identity-by-state pairwise genetic distance. **c**, Levels of ploidy in various potato groups as measured by flow cytometry or estimated based on nuclear SNP allelic frequencies (annotated with asterisks). **d**, Chlorotype classification derived from a maximum likelihood tree based on complete chloroplast genome assemblies. The oldest European potatoes carry a chloroplast type that segregates at high frequency only within tetraploid Andean landraces.

43 potatoes constituting a 'mini-core' subset of South American modern samples (SMS) representing landrace diversity. To include cultivars, we also sequenced 16 European modern samples (EMS).

Since potato species have relatively large genomes (~840 Mb)[14] and their ploidy level ranges from diploid to hexaploid[15], we adopted an array-based targeted re-sequencing approach that allowed us to obtain sufficient sequence coverage for accurate single nucleotide polymorphism (SNP) calling. The array targeted the whole chloroplast genome and ~4.3 Mb of the nuclear genome (Supplementary Table 3). The targeted nuclear fraction consisted of SNPs segregating in transcriptomes of elite cultivars[16–18] and 334 genes associated with response to changes in day length[19,20], including *StCDF1* (ref. [11]). Sequencing of unrepaired herbarium-derived libraries revealed damage patterns typical of ancient DNA[21] (Supplementary Fig. 2). To generate historical datasets virtually devoid of DNA damage, we prepared chemically repaired DNA libraries[22]. After removal of PCR duplicates, on average 20% of historical and 12% of contemporary (Supplementary Table 4) unique reads mapped to our nuclear targets (Supplementary Fig. 3), attaining a mean gene coverage of 47× and 70×, respectively (Supplementary Fig. 4).

The origins of 1650–1750 European potatoes. First, we set out to resolve relationships between all cultivated potato species derived from South America. SMS include tetraploid Chilean landraces of *Solanum tuberosum*, di-, tri- and tetraploid Andean landraces (including *S. andigenum*, *S. chaucha*, *S. goniocalyx*, *S. phureja* and *S. stenotomum*[23], now recognized as part of a more broadly defined *S. tuberosum*[15]), and bitter cultivated potato species (*S. curtilobum*, *S. juzepczukii* and *S. ajanhauri*[15]) (Supplementary Table 2). Based on chlorotypes and nuclear SNPs, we found substantial genetic distance between *S. tuberosum* and bitter potatoes (Fig. 1b and Supplementary Figs. 5 and 6). Within *S. tuberosum* we observed a clear distinction between Chilean and Andean landraces and, within the latter, a further division between diploids and tetraploids.

Because several *Solanum* species and ploidy forms are cultivated in South America, it was imperative to assess the genetic make-up and ploidy level of the potatoes that were initially introduced to Europe. For this we focused on the five oldest EHS (collected 1650–1750), which showed a high degree of genetic homogeneity (Fig. 1b). We established that these were tetraploid based on distributions of allele frequencies[24], as were a subset of Andean landraces and all Chilean potatoes (Fig. 1c and Supplementary Fig. 7). Furthermore, the oldest EHS carried a chlorotype that segregates at high frequency in tetraploid Andean landraces but is absent in diploid Andean landraces (Fig. 1d). Despite being tetraploid and carrying a chloroplast present mainly in tetraploid Andean landraces, based on nuclear genomic distances, the oldest EHS are similar to a subset of the diploid Andean landraces (Fig. 1b). Given that polyploidization is common in potato species[15], we speculate that this finding could be driven by absence of the tetraploid descendants of the EHS-related diploids. Taken together, our analyses suggest that the oldest European potatoes were tetraploids derived from the Andes, a geographic source that is consistent with the trading routes operating at the time[25].

The complex population history of European potatoes. We investigated potato evolution within Europe by analysis of all EHS and

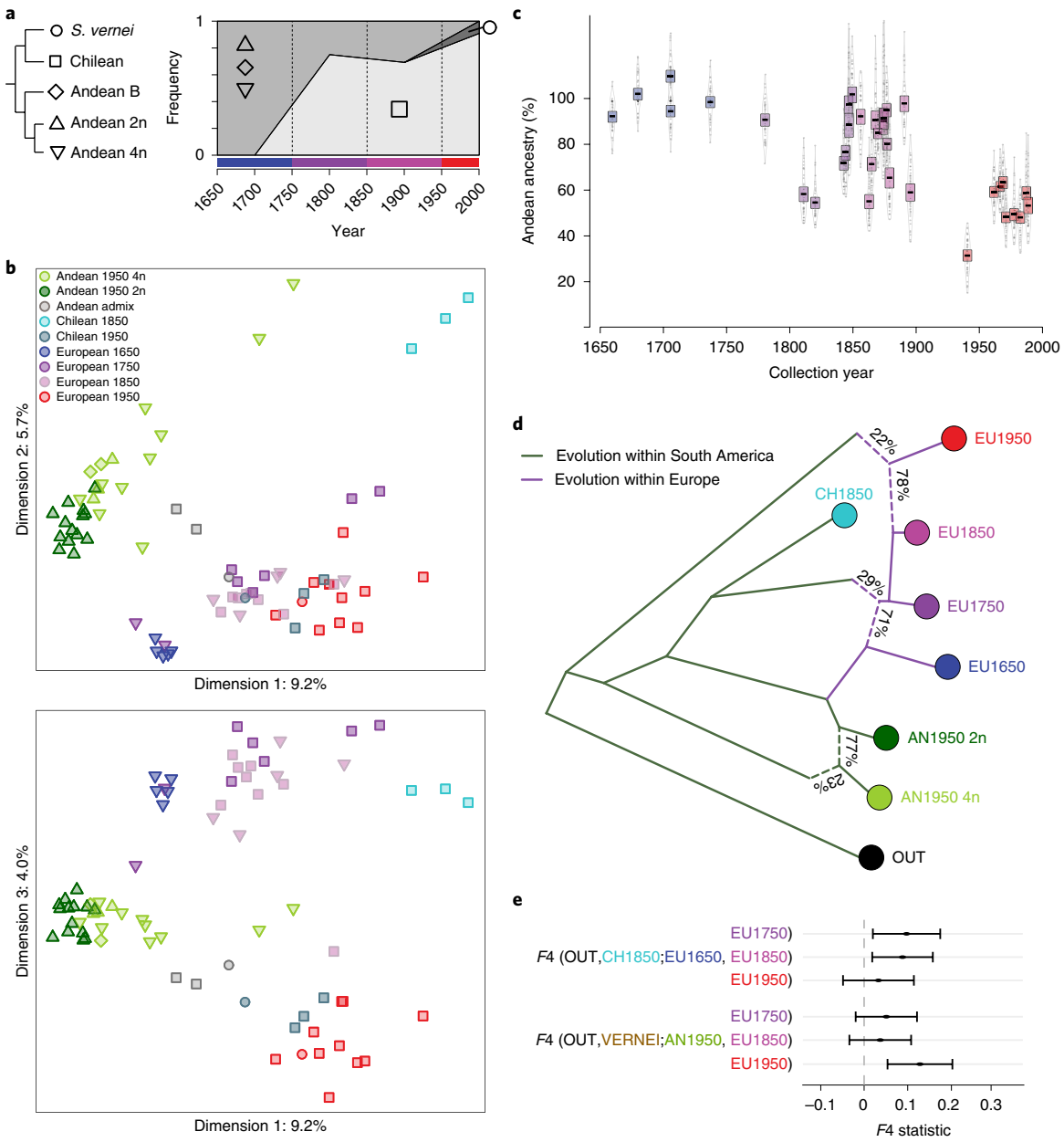

Fig. 2 | Genetic differentiation and admixture of potato populations in Europe. a, Simplified phylogenetic tree of potato chlorotypes and the plot of their frequencies in Europe over time. Andean, Chilean and wild potato chlorotypes are denoted by different symbols that are used in subsequent panels. **b**, Multi-dimensional scaling plot based on 35,659 nuclear SNPs. Genetic distances between potato samples are projected onto the first three (scaled) dimensions. Axis labels indicate the fraction of total variation explained in each dimension. European potato forms a genetic cline along dimension1 consistent with temporal changes in Andean and Chilean ancestry. **c**, Reduction in Andean genome ancestry in European samples over time calculated using f_4 ratio: (OUT,AN1950;X,CH1850)/(OUT,AN1950;EU1650,CH1850), with wild tomato *S. habrochaites* used as an outgroup. **d**, Best model of admixture graph for temporal populations from Europe and their source populations in the Andes and Chile. The graph indicates highly reticulate evolution and common hybridization of various potato populations in Europe. **e**, Population f_4 statistics indicate European potato admixture with Chilean potato in the eighteenth and nineteenth centuries, and with South American wild relatives in the twentieth century (error bars indicate three standard errors based on blocked jackknifing).

EMS samples spanning the years 1650–2000 (Supplementary Fig. 6). Phylogenetic reconstruction of whole-genome assemblies from herbarium chloroplasts allowed assignation of samples to chlorotypes (Supplementary Fig. 8) and investigation of chlorotype frequencies in 100-year intervals (Fig. 2a). We observed a decrease in Andean-related chlorotypes coupled with an increase in Chilean-related chlorotypes starting from the late eighteenth century, a pattern also present when we combined our data to a previously published dataset[20] (Supplementary Fig. 9). The change in chlorotype frequencies was mirrored by concurrent changes in the nuclear genome. European potatoes became progressively more similar to historical Chilean individuals while becoming increasingly distant from their Andean counterparts (Fig. 2b and Supplementary Fig. 10). The decrease in Andean ancestry became more apparent when its proportion in these historical samples was measured based on f_4 ratios (Fig. 2c). The Andean ancestry halved in Europe between

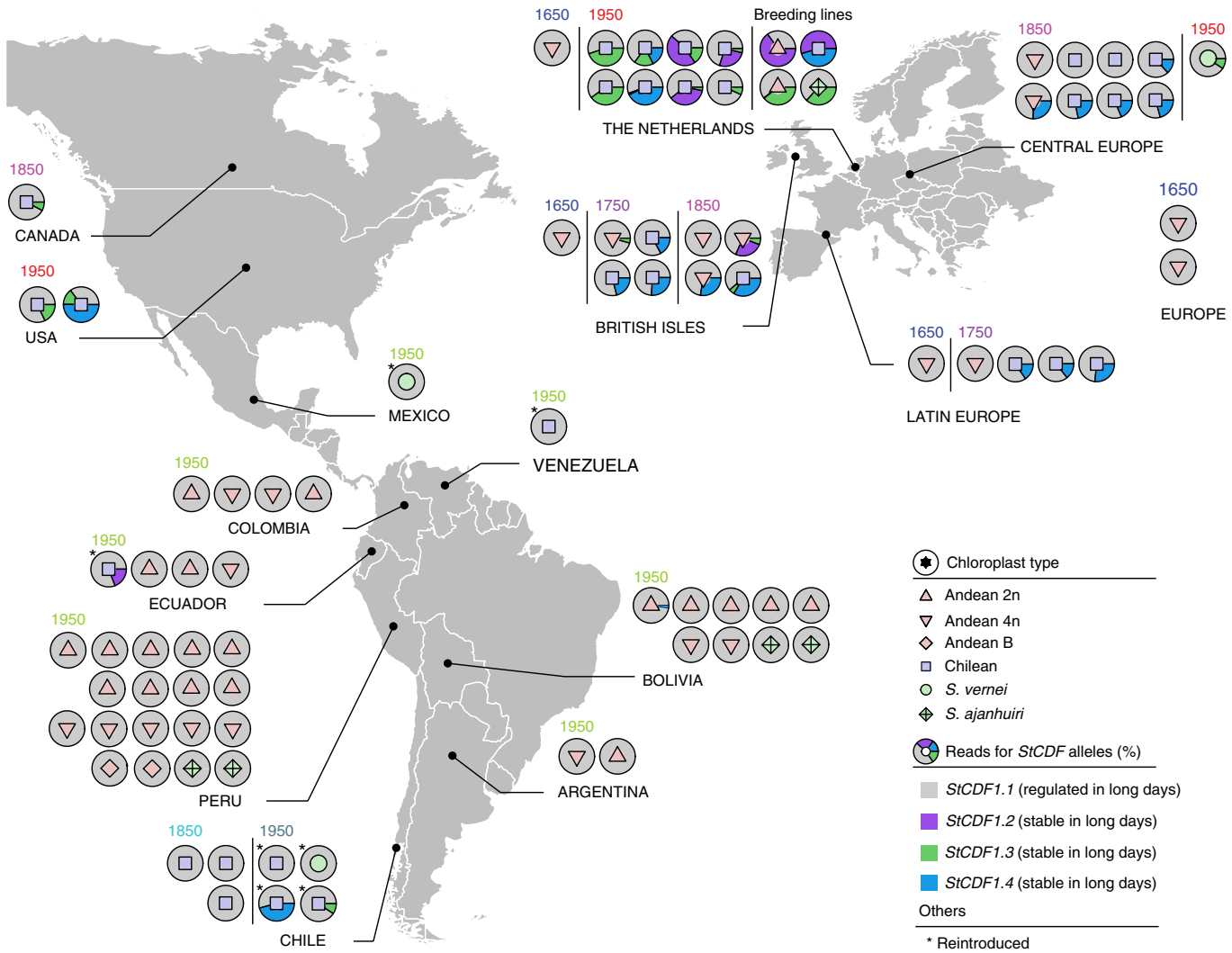

Fig. 3 | Geographic and temporal distribution of stabilizing insertions in the *StCDF1* gene. Individual potato samples are represented by circles and placed on the map relative to the place and time of their collection. Within circles, central shapes represent chlorotypes, while pie charts represent the number of DNA reads supporting the presence of *StCDF1* alleles. For South America, we observed no presence of the *StCDF1.2*, *1.3* or *1.4* allelic variants, other than in individuals, for which chlorotype and nuclear SNPs indicated admixture with European potato (marked by asterisks). In Europe, stabilizing insertions are present starting from the eighteenth century onwards, but are absent in the earliest European potatoes. All diploid breeding lines that were previously used to discover the role of *StCDF1* in long-day adaptation contained the stabilizing insertions.

the seventeenth and twentieth centuries. This steady decrease was disrupted in the years 1846–1891 with the resurgence of Andean ancestry, which coincides with the potato late blight epidemic in 1845–1847 that triggered the Irish potato famine[26,27]. This shift in ancestry suggests that farmers at the time may have reintroduced older potato stocks to overcome the famine caused by losses of pathogen-susceptible crops.

Subsequently, to investigate the dynamics of European potato ancestry, we grouped EHS into four temporal populations in 100-year intervals (Supplementary Table 1) and modelled their relationship employing an admixture graph framework (Fig. 2d)[28]. In agreement with our results based on genetic distance (Fig. 1b), the oldest EHS (collected 1650–1750) were found to have descended directly from an ancestor of Andean landraces. Within the next 100 years, potatoes in Europe admixed with newly introduced Chilean potatoes (Fig. 2d,e). In addition, twentieth-century European potatoes were found not to have descended directly from their nineteenth-century admixed predecessors, but rather to have received gene flow from a wild potato species (Fig. 2d). Wild species such as *S. vernei* and *S. demissum* were used in twentieth-century breeding programmes to introduce resistance to plant pathogens[20,29]. Using available genomic resources for *S. vernei*[30], we validated that this admixture signal in Europe can be attributed to modern breeding with wild *Solanum* species (Fig. 2e).

European potatoes also had their imprint on Andean and Chilean diversity, probably through reintroductions coupled with admixture. Within SMS, we identified three individuals carrying European ancestry based on chlorotype and principal component analysis (PCA) nuclear ancestry deconvolution[31,32] (Supplementary Figs. 8, 10 and 11). This PCA-based method also revealed that all sampled contemporary Chilean potatoes are very similar to modern potatoes in Europe, but very different from the historical Chilean samples collected by Darwin and Isern in 1834 and 1862, respectively. Moreover, one contemporary Chilean sample carried a chlorotype from a wild relative used in potato breeding. To formally ascertain the ancestry of contemporary Chilean samples, we included them in the extended admixture graph model which indicated 83% European ancestry (Supplementary Fig. 12). In sum,

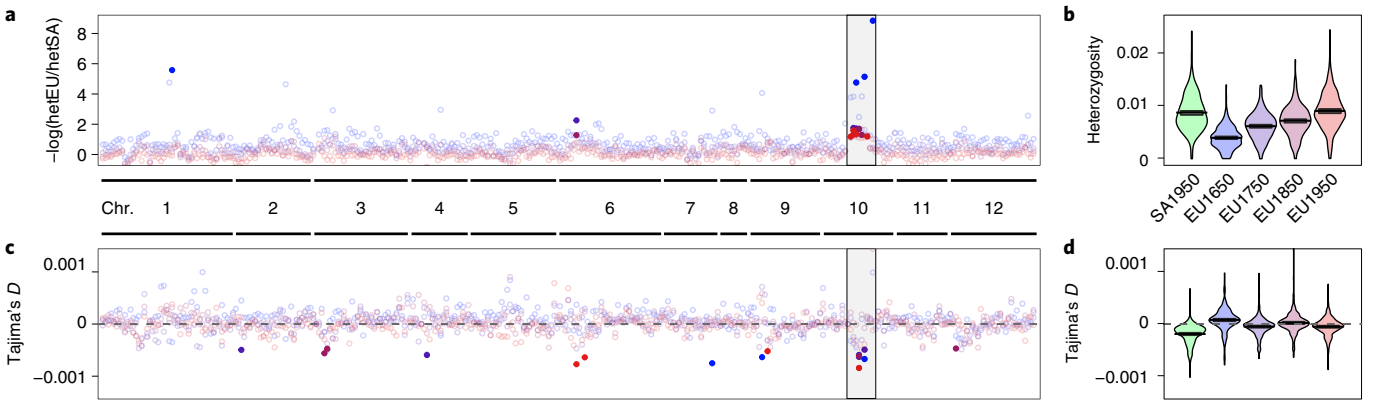

Fig. 4 | Potential signatures of selection in genes linked to photoperiod. a, Loss of heterozygosity in European temporal populations as compared to the Andean population. The 1% outliers of empirical distributions are represented by filled opaque circles. Throughout all the centuries in Europe (blue, seventeenth; violet, eighteenth; purple, nineteenth; red, twentieth), heterozygosity has not recovered on chromosome 10 in the region enriched with gibberellin synthesis genes (grey shading). Chr., chromosome. hetEU, heterozygosity in European temporal populations. hetSA, heterozygosity in tetraploid Andean landraces. **b**, Distribution of heterozygosity in all tested genes. Heterozygosity dropped visibly due to a bottleneck associated with the early introduction of potato to Europe, and was subsequently recovered by gene flow from various South American potato populations. **c**, Tajima's *D*-statistic in European temporal populations. Many genes in the gibberellin pathway (grey shading) are characterized by an excess of rare variants (1% outliers to empirical distributions), which could indicate positive selection. **d**, Distribution of Tajima's *D* in all tested genes. Elevated Tajima's *D* characteristic of population contractions is present in the earliest introduced European population. Over time, populations in Europe stabilized based on equilibrium between rare and common variants.

our analyses suggest the nineteenth-century reintroduction of European potato to Chile (Supplementary Figs. 8 and 10–12).

Adaptation to long-day tuberization in European potato. Taking into account the admixture-driven population turnover in Europe, we sought to investigate the origin of long-day-adaptive alleles in the *StCDF1* gene. The adaptive insertions could have entered the European potato population (1) from standing variation of Andean landraces, (2) from a de novo mutation in Europe or (3) through admixture with pre-adapted Chilean landraces or wild species. To test these scenarios we quantified in each sample the reads that support the presence of an undisrupted allele (*StCDF1.1*), and those that support alleles disrupted by either a putative transposon (*StCDF1.3*) or a seven-base pair (bp) insertion (*StCDF1.2*)[11]. In 18 samples a substantial number of short reads could not be assigned to any of the known alleles, thus leading to the discovery of a previously unknown 7-bp insertion variant (*StCDF1.4*; Supplementary Fig. 13). As described for *StCDF1.2*, the 7-bp insertion in *StCD1.4* causes a frameshift mutation resulting in loss of the C-terminal domain, hence conferring long-day adaptation through StCDF1 stabilization[11].

None of the alleles with adaptive insertions were found in the oldest European samples of Andean descent (Fig. 3 and Supplementary Table 5). Similarly, we did not find any of these variants in Andean landraces (excluding those of European ancestry; Supplementary Figs. 8, 10 and 11). Starting from 1810 in Europe, each of these adaptive alleles was observed, coinciding with the onset of Chilean admixture (Fig. 2c). Since, by all indications, contemporary Chilean varieties were reintroduced from Europe, we used three historical herbarium specimens as being representative of Chilean ancestry. We found no evidence for adaptive insertions in *StCDF1* in historical samples from the coastal lowlands of Chile. However, we noticed an elevated number of segregating sites in exons and introns and, within exons, an inflated number of missense mutations (Supplementary Fig. 14). Although multiple missense mutations have previously been reported in some wild species and modern cultivars[30], their functional importance remains to be determined. In summary, our analyses suggest that it is very unlikely that long-day adaptation originates from Andean standing variation. Although our sampling of historical Chilean specimens is not dense, and does not allow us to reject a Chilean origin of adaptive insertions, based on our data we speculate that the insertions could have arisen de novo in Europe and were rapidly 'fixed' due to their dominant inheritance and potential breeding advantage.

The lack of adaptive insertions in *StCDF1* within the oldest European potatoes (collected 1650–1750) led us to investigate other potentially adaptive genes. To that end, we calculated two summary statistics, loss of heterozygosity and Tajima's *D*[33], independently for each of our temporal populations in Europe and for each photoperiod response gene (Fig. 4). We found an area on chromosome 10 encompassing eight photoperiod response genes with substantial reduction in heterozygosity as compared to Andean landraces (the top 1% tail of the empirical distribution) (Fig. 4a and Supplementary Table 6). Despite the general trend of increasing heterozygosity in Europe (Fig. 4b), the pattern of reduced diversity on chromosome 10 was stable across 350 years of potato evolution, which could indicate ongoing selection. Loci in this putatively selected region correspond to gibberellic acid synthesis and a MADS-box gene targeted by SP6A florigen[34]. Two of our candidates were previously linked to tuberization promotion through synthesis of gibberellins[35]. Furthermore, two genes within the same region were characterized by an unusually high number of rare alleles (the top 1% tail of the empirical distribution) (Fig. 4c). Such a negative Tajima's *D* is particularly remarkable given that the oldest EHS have, in general, elevated Tajima's *D*, probably due to population contraction associated with the introduction of the potato to Europe (Fig. 4d). Although with time-averaged Tajima's *D* approaches zero, we found one locus on chromosome 10 that shows a consistent surplus of rare alleles, suggesting selection on nearby loci. A mutation in the orthologue of this gene (*AGL62*) in *Arabidopsis thaliana* facilitates the formation of viable hybrid seeds through inhibition of endosperm development[36]. Mutations in *AGL62* could hence have played an important role in potato interploidy hybridization[15,37].

Discussion

The introduction of potatoes to Europe has been documented using evidence mainly derived from historical documents and literature. It is likely that after initial arrival in the late sixteenth century[3],

different genotypes of tuber-bearing potatoes from various parts of South and Central America were brought to Europe. In our study of the last 350 years of European potato evolution, we characterized the genetic diversity of historical potatoes using genome-wide markers. We show that the evolution of cultivated potato was shaped by at least two admixture events, from (1) distinct Chilean potatoes during the eighteenth century and (2) wild *Solanum* species used in breeding programmes during the twentieth century[29].

Our results suggest that the first potatoes introduced to Europe were probably pre-adapted through changes in the gibberellin pathway, and did not carry the gain-of-function alleles in the *StCDF1* gene that confer adaptation to a European climate. The emergence of the *StCDF1* adaptive variants coincided with the influx of Chilean ancestry into European potato during the eighteenth century. Although this is suggestive of a Chilean origin of the adaptive variants, we did not find such variants in historical Chilean specimens. Therefore, we propose that native CHS were adapted through different molecular mechanism. Due to the scarcity of historical Chilean samples, we cannot rule out that the adaptive variants were segregating at low frequency in Chile during pre-Columbian times. In spite of that, our findings suggest the possibility that the *StCDF1* adaptive variants could have emerged de novo in Europe. Similar to previous studies on human population history[38], we show here that elucidation of the complex origins of different potato lineages was possible only through the use of historical museum specimens. Without these, our inferences would be confounded by population replacements and reintroductions. Although population replacement in Chile was inferred from only three herbarium individuals due to the limited availability of historical samples, we argue that this inference is reliable since our genome-wide SNPs represent genetic blocks contributed by many genetic ancestors of current and historical potato populations.

Our analyses of admixture provide further insight into the origin and integration of novel genetic variants in modern domesticated potato. These results revealed that Europe became a melting pot of multiple potato varieties introduced from various parts of South America. We argue that the rapid and wide geographical expansion of potatoes during post-Columbian times favoured hybridization beyond what would have been possible in their native range and under limited human-driven mobility. We speculate that hybridization propensity in Europe may, in addition, have been promoted by putatively selected mutations in the AGL62-like gene. There is mounting evidence for post-domestication admixture from wild relatives into crops such as rice[39,40] and maize[41,42] after mid- and short-range expansions. In potato, long-range expansion combined with hybridization propensity and post-domestication admixture resulted in extensive diversity in Europe.

Methods

DNA extraction, library preparation and targeted capture. *Historical DNA.* Extraction of DNA for all herbarium specimens (Supplementary Methods 1.1 and 1.2 and Supplementary Table 1) was carried out in the clean-room facility at the Institute of Archaeological Sciences, University of Tuebingen. To avoid contamination by exogenous DNA, state-of-the-art precautions were taken when extracting ancient DNA—the use of protective gear by experimenters and of separate hoods for handling tissue, reagents and DNA extracts; and sterilization with UV light of all equipment, surfaces and hoods after each extraction round. A modified extraction method[43], which combines a *N*-phenacylthiazolium bromide lysis buffer with MinElute Purification columns (Qiagen), was utilized on the set of herbarium samples extracted in this study (Supplementary Table 1). Each batch of processed samples (Supplementary Table 4) was accompanied by at least one negative control, which contained no plant tissue.

Genomic libraries were constructed in a clean-room facility using the above-mentioned precautionary measures. Samples were processed in the same batches as for DNA extraction (Supplementary Methods 1.3 and Supplementary Table 4) accompanied by at least one negative control, which contained the extraction blank, and one additional negative control for library preparation without DNA extract. We used a library preparation method tailored for mutiplexed targeted capture and sequencing[44]. This protocol included modifications that render it suitable for handling of ancient DNA, such as replacement of magnetic beads by silica-column purification[45]. Libraries were quantified in a real-time quantitative PCR (qPCR) reaction with a DyNAmo HS SYBR Green kit (Thermo Fisher Scientific) in LightCycler 96 (Roche). These libraries were subsequently indexed with two barcoded primers during ten cycles of amplification[46] with AccuPrime Pfx polymerase (Thermo Fisher Scientific). Library amplification was carried out in a laboratory located in a separate building and distant from the clean-room facility. Amplified libraries were purified from PCR reagents using MinElute Purification columns and quantified again using PCR with reverse transcription. The final amplification step was carried out by employing a number of cycles adjusted to the number of molecules present in the library based on qPCR measurements (cycle number ranged from three to five for samples and from 12 to 15 for extraction blanks). Libraries were pooled in equimolar concentrations and sequenced on the Illumina MiSeq platform using MiSeq Reagent Kit v.2, 300 cycles (Illumina).

We additionally prepared genomic libraries with the protocol extended by the use of uracil-DNA glycosylase (UDG) treatment (Supplementary Methods 1.4). These libraries were prepared as described above, but during the blunting step the USER enzyme (New England Biolabs) was added and the time and temperature of incubation were adjusted accordingly[22]. UDG-treated libraries were pooled in equimolar concentrations and subjected to targeted capture.

Modern DNA. Potato samples representing contemporary diversity in South America and Europe (Supplementary Methods 1.1 and 1.5 and Supplementary Table 2) were snap-frozen in liquid nitrogen. Leaf and stem tissue (<20 mg dried material,<100 mg fresh material) was placed in tubes with one tungsten bead (3 mm diameter) and disrupted using TissueLyserII (Qiagen). The snap-freezing and disruption procedure was carried out twice. Ground tissue was subjected to DNA extraction using DNEasy Plant Mini Kits (Qiagen) following the manufacturer's protocol, with extended elution incubation of 10 min.

Extracted DNA for contemporary samples was sheared in a Covaris instrument with the following duty cycle: 10%; intensity, 5.0; cycles, 200; duration, 45 s. Shearing efficiency was visualized using BioAanalyzer. Fragmented DNA was used to construct genomic libraries following a published protocol, with magnetic bead clean-ups between each enzymatic reaction[44] (Supplementary Methods 1.6).

Targeted DNA capture. Enrichment was carried out on a custom-designed SureSelect DNA Capture Array with one million features (Agilent Technologies) and design ID 084050. Bait probes were designed to cover a wide range of genomic features, including 334 genes related to photoperiod response and the entire chloroplast genome (Supplementary Methods 1.7 and Supplementary Table 3). Bait probes of length 60 bp were tiled to cover nuclear targeted regions starting every 3 bp. Bait probes for chloroplast capture were tiled starting every 5 bp. All probes were filtered based on 15-mer analysis to exclude repetitive sequences[47,48]. The hybridization procedure for both historical and contemporary samples was carried out following a published protocol[47]. Enriched libraries were pooled in equimolar concentrations and sequenced on the Illumina HiSeq 3000 platform using a HiSeq Reagent Kit v.2, 300 cycles (Illumina) in paired-end mode.

Authentication of ancient DNA in herbarium specimens. The Illumina output format was converted to fastq, and de-multiplexing was carried out based on the recognition of two 8-bp indices using bcl2fastq2 provided by Illumina (Supplementary Methods 2.1). Adaptors were removed from raw reads using Skewer v.0.1.120 (ref. [49]) and paired reads were merged using Flash v.1.2.11 (ref. [50]). Afterwards, merged reads were mapped to the potato reference genome PGSC v.4.03 (ref. [14]) using bwa *mem* v.0.7.10 (ref. [51]), then sorted utilizing samtools v.1.2 (ref. [52]). Finally, PCR duplicates were removed using picard tools v.2.3.0.

Ancient DNA-associated damage[21,53] (Supplementary Methods 2.2) was quantified in mapped reads (after removal of PCR duplicates) with MapDamage2.0 (ref. [54]). Subsequently, the length distribution of mapped sequences after removal of PCR duplicates was plotted in R v.3.4.1 (ref. [55]) and the median of this distribution calculated. The fraction of potato endogenous DNA was calculated as the proportion of the number of mapped reads to the potato reference genome over the total number of sequenced and merged reads. Library complexity was estimated as the average number of unique reads overlapping a single base across the entire reference genome. For this calculation we used qPCR estimation of the number of molecules in unamplified libraries, multiplied by the fraction of endogenous sequences and median fragment size and then divided by the total length of the genome (Supplementary Table 4).

Bioinformatic processing of non-damaged libraries. *Preprocessing of UDG-treated libraries.* De-multiplexing, trimming, merging and mapping of sequencing reads were carried out as described in the preceding section. Removal of PCR duplicates was carried out with the DeDup programme of the EAGER pipeline[56] (Supplementary Methods 2.3). The total number of merged reads was calculated for each library (Supplementary Fig. 3). From that number we calculated the proportion of reads that were (1) successfully mapped to the reference genome, (2) uniquely mapped (without PCR duplicates), (3) mapped to targeted regions of the nuclear genome and (4) mapped to the chloroplast genome (Supplementary Fig. 3 and Supplementary Table 4).

Supplementary sequencing data for cultivated potato[30] (experiment No. SRX2646034) and for wild tomato[57] (experiment Nos. ERX384399–ERX384404) were downloaded from the Short Read Archive (Supplementary Methods 2.5). Subsequently, short reads were processed as described in the preceding section.

Analyses of capture efficiency. The number of reads that mapped to targeted regions of the nuclear and chloroplast genomes was used to calculate average coverage and enrichment coefficient (Supplementary Methods 2.4, Supplementary Fig. 3 and Supplementary Table 4). The coefficient of enrichment was calculated as the proportion of on-target reads relative to the total number of mapped reads.

Assembling chloroplast genomes. Reads sequenced from enriched libraries were mapped to the potato reference chloroplast genome using bwa *mem* v.0.7.10 (ref. [51]). After sorting using samtools v.1.2 (ref. [52]), files were converted to unaligned fasta format (Supplementary Methods 2.6). SPAdes v.3.5.0 (ref. [58]) was used to build contigs from chloroplast fasta files. These contigs were aligned back to the reference chloroplast genome using bwa *mem* and samtools. Subsequently, samtools mpileup was used for reference-guided scaffolding of assembled contigs.

Ploidy estimation. Mapped reads from enriched UDG-treated libraries were used for direct investigation of the ploidy of historical and modern samples sequenced for this study (Supplementary Methods 2.7). Likelihoods were maximized under the assumptions of di-, tri- and tetraploidy using nQuire[24]. These maximized log-likelihoods were normalized by the best fit, and used to cluster all samples in three dimensions using multivariate normal mixtures[59].

SNP calling and dataset filtering. Variant calling was carried out following Genome Analyses ToolKit best practices[60,61]. For each individual, separately, gatk v.3.8 was used to realign mapped reads around indels. The programmes HaplotypeCaller and GenotypeGVCFs were used to discover and genotype variable sites. Filtering criteria were established separately for different datasets (Supplementary Methods 2.8).

The programme VariantRecalibrator was used to train six Gaussian mixture models based on a true positive set of SNPs previously ascertained. Parameters that were taken into account during training were: (1) distributions of qualities normalized by depth; (2) mapping of quality distribution compared by Wilcoxon test; (3) read position bias measured by Wilcoxon test; (4) strand bias measured by Fisher' test and symmetric odds ratio test; (5) number of hard-clipped distributions compared by Wilcoxon test; and (7) base quality distribution compared by Wilcoxon test. After training, the programme ApplyRecalibration was used for filtering with a tranche that recovers 90% of true-positives and a very small number of false-negatives to generate the filtered SNP list. This dataset contained 1,135,966 polymorphic sites. SNPs with average coverage of 20× or greater in SMS, EMS and EHS were kept for random allele sampling.

Phylogenetic and population genetics analyses. *Chloroplast trees.* Chloroplast scaffolds previously generated using Spades were aligned to the reference chloroplast genome sequence using bwa *mem* v.0.7.10 (ref. [51]) (Supplementary Methods 2.9). A multiple sequence alignment was then reconstructed using mafft v.7.310 (ref. [62]). Regions with indels in at least five different samples were deleted together with flanking regions containing a SNP within 20 bp of the gap. Aligned fasta files for chloroplast genomes were used to search for a maximum likelihood phylogenetic tree using RAxML v.8.1.20 (ref. [63]) with the GTRCATI model of molecular evolution. To test the robustness of clades in this tree, 100 bootstrap resampling trees were compared.

Multi-dimensional scaling. Genetic distances were calculated between all pairs of samples using Plink1.9 (refs. [64,65]) with the formulation 1– IBS, where IBS represents identity by state (Supplementary Methods 2.10). The distance matrix was imported into R[55] where the *cmdscale* function was used to calculate eigenvectors[66], which were then plotted in three dimensions. All-data PCA and reduced-data PCA with projection were conducted with *SmartPCA*, part of the Eigensoft package[67] (Supplementary Methods 2.11).

Population F-statistics. For these analyses, populations were defined as outlined in Supplementary Methods 2.12 and Supplementary Tables 1 and 2. Population *F*-statistics were calculated for all unique combinations of four populations (f_4) and three populations (f_3) with AdmixTools v.4.1 (ref. [28]) *qp3Pop* and *qpDstat* functions (Supplementary Methods 2.13). In the case of f_4 statistics, we always used *S. habrochaites* (OUT) as an outgroup. For visualization we used the admixture Graph package[68] in R v.3.4.1 (ref. [55]).

We carried out an f_4 ratio test using the *qpDstat* method in AdmixTool v.4.1 (ref. [28]) (Supplementary Methods 2.14). For each individual X from the group of earliest introduced EHS, we calculated the numerator as a *D*-statistic of the form of $D_{(OUT,AN1950;X,CH1850)}$, where OUT is *S. habrochaites* as an outgroup (ERR418099), AN1950 is each of three selected individuals of Andean tetraploid landraces (SA08, SA26, SA31) and CH1850 is each of the three historical Chilean samples (HB22, HB32, HB33). As the denominator we calculated a *D*-statistic of the form $D_{(OUT,AN1950;EU1650,CH1850)}$, where OUT, AN1950 and CH1850 are as described above and EU1650 are the five earliest EHS (HB27, HB28, HB29, HB30 and HB31) serving as partial ancestors.

We applied an unbiased, 'brute force' approach to evaluate all possible admixture graphs for potato populations with the admixtureGraph package[68] in R v.3.4.1 (ref. [55]) (Supplementary Methods 2.15). We evaluated models with zero, one or two admixture events for six populations: outgroup, AN1950n4, AN1950n2, CH1850 (as the potential source populations for European potato), EU1650 and EU1850 (as our test populations for evolution of potato in Europe). In the first search we evaluated models using the complete unique set of f_3 and f_4 statistics (Supplementary Table 7) and validated this with AdmixTools v.4.1 *qpGraph* method[28], keeping only those models characterized with *z*-scores <3 (Supplementary Fig. 15a). Subsequently, we extended the admixture graph models by including the historical potato population EU1850 as a sister group of the genetically similar EU1750 (Supplementary Fig. 15b). Following that, we added the modern potato population EU1950 (Supplementary Fig. 15c).

Analyses of candidate genes. Allelic depth in *StCDF1.* All merged reads for our sequenced samples were mapped against a reference constructed by concatenating the four *StCDF1* structural variants (*StCDF1.1*, *StCDF1.2*, *StCDF1.3* and *StCDF1.4*; Supplementary Fig. 13) with bwa *mem* v.0.7.10 (ref. [51]), and then sorted utilizing samtools v.1.2 (ref. [52]) (Supplementary Methods 2.16). Finally, PCR duplicates were removed using DeDup[56]. To remove spuriously mapped reads in the *StCDF1* local context, we converted all successfully mapped reads to fastq format and mapped these against the potato reference genome (v.4.03)[14], repeating the above-mentioned procedure and filtering out reads with multiple mapping possibilities (mapping quality 1). Subsequently we quantified reads that overlapped the insertion site in variants *StCDF1.1*, *StCDF1.2* and *StCDF1.4* with at least 20 bp on each flanking side. Since no single short read could span the entire insertion of variant *StCDF1.3* (865 bp), we quantified reads that included at least 20 bp of transposon and 20 bp of flanking sequence on each side of the insertion site (Supplementary Fig. 13).

Loss of heterozygosity. We calculated allele frequencies in populations AN1950 4n, EU1650, EU1750, EU1850 and EU1950, the difference in this analysis being that we included four distinct individuals of the AN1950 4n population, which were filtered out for population structure and *F*-statistic analyses (Supplementary Methods 2.17). We then calculated heterozygosity for each SNP as Het = 1 − SSF, where SSF is the sum of squares of each allele's frequency. For each gene we then calculated the sum of Het for each segregating SNP and divided this by the length of genic sequence. To calculate loss of heterozygosity for each gene, we divided Het for each temporal European population by Het for AN1950 4n and expressed this on a negative logarithmic scale. We visualized the empirical distribution for loss of heterozygosity using the yarrr package in R v.3.4.1 (ref. [55]) and highlighted genes in the 99th percentile of this empirical distribution.

Tajima's D. We calculated nucleotide diversity (mean of pairwise sequence differences within population) and number of segregating sites for each population in each photoperiod response gene (Supplementary Methods 2.18). We then calculated Tajima's *D*[53] by applying equations implemented in *tajima.test* function of the pegas[69] package to quad-allelic individuals. Finally, we visualized the empirical distribution for Tajima's *D* using the yarrr package in R v.3.4.1 (ref. [55]) and highlighted genes in the 99th percentile of this empirical distribution.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sequencing data generated in this study are available in the European Nucleotide Archive (ENA) under project Nos. PRJEB31013 and PRJEB31171 for UDG-treated and non-UDG-treated libraries, respectively.

Received: 29 January 2019; Accepted: 10 May 2019;
Published online: 24 June 2019

References

1. Spooner, D. M., McLean, K., Ramsay, G., Waugh, R. & Bryan, G. J. A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. *Proc. Natl Acad. Sci. USA* **102**, 14694–14699 (2005).
2. Hawkes, J. G. & Francisco-Ortega, J. The early history of the potato in Europe. *Euphytica* **70**, 1–7 (1993).
3. Ames, M. & Spooner, D. M. DNA from herbarium specimens settles a controversy about origins of the European potato. *Am. J. Bot.* **95**, 252–257 (2008).
4. Mcneill, W. H. How the potato changed the world's history. *Soc. Res.* **66**, 67–83 (1999).
5. Diamond, J. Evolution, consequences and future of plant and animal domestication. *Nature* **418**, 700–707 (2002).

6. Allaby, R. G. et al. Using archaeogenomic and computational approaches to unravel the history of local adaptation in crops. *Philos. Trans. R. Soc. Lond. B* **370**, 20130377 (2015).
7. Shennan, S. et al. Regional population collapse followed initial agriculture booms in mid-Holocene Europe. *Nat. Commun.* **4**, 2486 (2013).
8. Colledge, S., Conolly, J. & Shennan, S. The evolution of neolithic farming from SW asian origins to NW european limits. *Eur. J. Archaeol.* **8**, 137–156 (2016).
9. Fuller, D. Q. & Allaby, R. G. in *Fruit Development and Seed Dispersal* (ed. Ostergaard, L.) 238–295 (Blackwell, 2009).
10. Jackson, S. D. Multiple signaling pathways control tuber induction in potato. *Plant Physiol.* **119**, 1–8 (1999).
11. Kloosterman, B. et al. Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nature* **495**, 246–250 (2013).
12. Spooner, D., Jansky, S., Clausen, A., del Rosario Herrera, M. & Ghislain, M. The enigma of *Solanum maglia* in the origin of the chilean cultivated potato, *Solanum tuberosum* Chilotanum group. *Econ. Bot.* **66**, 12–21 (2012).
13. Ghislain, M., Núñez, J., Herrera, M., del, R. & Spooner, D. M. The single Andigenum origin of Neo-Tuberosum potato materials is not supported by microsatellite and plastid marker analyses. *Theor. Appl. Genet.* **118**, 963–969 (2009).
14. The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189–195 (2011).
15. Spooner, D. M., Ghislain, M., Simon, R., Jansky, S. H. & Gavrilenko, T. Systematics, diversity, genetics, and evolution of wild and cultivated potatoes. *Bot. Rev.* **80**, 283–383 (2014).
16. Hamilton, J. P. et al. Single nucleotide polymorphism discovery in elite North American potato germplasm. *BMC Genom.* **12**, 302 (2011).
17. Hirsch, C. N. et al. Retrospective view of North American potato (*Solanum tuberosum* L.) breeding in the 20th and 21st centuries.*G3* **3**, 1003–1013 (2013).
18. Felcher, K. J. et al. Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS ONE* **7**, e36347 (2012).
19. Shan, J. et al. Transcriptome analysis reveals novel genes potentially involved in photoperiodic tuberization in potato. *Genomics* **102**, 388–396 (2013).
20. Uitdewilligen, J. G. A. M. L. et al. A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS ONE* **8**, e62355 (2013).
21. Briggs, A. W. et al. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl Acad. Sci. USA* **104**, 14616–14621 (2007).
22. Briggs, A. W. et al. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* **38**, e87 (2010).
23. Bukasov, S. M. Principles of the systematics of potatoes. *Trudy po prikladnoi botanike, genetike i selektsii* **62**, 3–35 (1978).
24. Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S. & Burbano, H. A. nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics* **19**, 122 (2018).
25. Salaman, R. N. The early European potato: its character and place of origin. *Bot. J. Linn. Soc.* **53**, 1–27 (1946).
26. Austin Bourke, P. M. Emergence of potato blight, 1843–46. *Nature* **203**, 805 (1964).
27. Yoshida, K. et al. The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. *Elife* **2**, e00731 (2013).
28. Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
29. Bradshaw, J. E., Bryan, G. J. & Ramsay, G. Genetic resources (including wild and cultivated *Solanum* species) and progress in their utilisation in potato breeding. *Potato Res.* **49**, 49–65 (2006).
30. Hardigan, M. A. et al. Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proc. Natl Acad. Sci. USA* **114**, E9999–E10008 (2017).
31. Sawler, J. et al. Genomics assisted ancestry deconvolution in grape. *PLoS ONE* **8**, e80791 (2013).
32. Bryc, K. et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc. Natl Acad. Sci. USA* **107**, 786–791 (2010).
33. Tajima, F. Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
34. Gao, H. et al. Genome-wide survey of potato MADS-box genes reveals that *StMADS1* and *StMADS13* are putative downstream targets of tuberigen *StSP6A*. *BMC Genomics* **19**, 726 (2018).
35. Martínez-García, J. F., García-Martínez, J. L., Bou, J. & Prat, S. The interaction of gibberellins and photoperiod in the control of potato tuberization. *J. Plant Growth Regul.* **20**, 377–386 (2001).
36. Walia, H. et al. Dosage-dependent deregulation of an *AGAMOUS-LIKE* gene cluster contributes to interspecific incompatibility. *Curr. Biol.* **19**, 1128–1132 (2009).
37. Köhler, C., Mittelsten Scheid, O. & Erilova, A. The impact of the triploid block on the origin and evolution of polyploid plants. *Trends Genet.* **26**, 142–148 (2010).
38. Nielsen, R. et al. Tracing the peopling of the world through genomics. *Nature* **541**, 302 (2017).
39. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
40. Choi, J. Y. et al. The rice paradox: multiple origins but single domestication in asian rice. *Mol. Biol. Evol.* **34**, 11 (2017).
41. van Heerwaarden, J. et al. Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proc. Natl Acad. Sci. USA* **108**, 1088–92 (2011).
42. da Fonseca, R. R. et al. The origin and evolution of maize in the Southwestern United States. *Nat. Plants* **1**, 14003 (2015).
43. Gutaker, R. M., Reiter, E., Furtwängler, A., Schuenemann, V. J. & Burbano, H. A. Extraction of ultrashort DNA molecules from herbarium specimens. *Biotechniques* **62**, 76–79 (2017).
44. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, db.prot5448 (2010).
45. Meyer, M. et al. A high-coverage genome sequence from an archaic denisovan individual. *Science* **338**, 222–226 (2012).
46. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**, e3 (2012).
47. Hodges, E. et al. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat. Protoc.* **4**, 960–974 (2009).
48. Burbano, H. A. et al. Targeted investigation of the Neandertal genome by array-based sequence capture. *Science* **328**, 723–725 (2010).
49. Jiang, H., Lei, R., Ding, S.-W. & Zhu, S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15**, 1–12 (2014).
50. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
51. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997v2 (2013).
52. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
53. Hofreiter, M., Jaenicke, V., Serre, D., von Haeseler, A. & Pääbo, S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* **29**, 4793–4799 (2001).
54. 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).
55. R Core Team *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2014).
56. Peltzer, A. et al. EAGER: efficient ancient genome reconstruction. *Genome Biol.* **17**, 60 (2016).
57. 100 Tomato Genome Sequencing Consortium. et al. Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. *Plant J.* **80**, 136–148 (2014).
58. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
59. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *R J.* **8**, 289–317 (2016).
60. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–33 (2013).
61. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
62. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
63. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
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. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
66. Mardia, K. V. Some properties of clasical multi-dimesional scaling. *Commun. Stat. Theory Methods* **7**, 1233–1241 (1978).
67. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
68. Leppälä, K., Nielsen, S. V. & Mailund, T. admixturegraph: an R package for admixture graph manipulation and fitting. *Bioinformatics* **33**, 1738–1740 (2017).
69. Paradis, E. Pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419–420 (2010).

Acknowledgements

We thank D. Weigel for initial discussions during the conception of this project; B. Glover and C. Bartram (Cambridge University Herbarium), M. Graniszewska and H. Leśniewska (University of Warsaw) for granting access to historic herbarium samples; J. Krause (MPI for the Science of Human History), V. Schuenemann and E. Reiter (University of Tuebingen) for access to clean-room facilities and technical support; M. Neumann and S. Latorre for laboratory assistance; L. Shannon (University of Minnesota) for input on potato diversity; C. Bachem, H. van Eck and J. Willemsen (Wageningen University) for input on array design; N. Arciniegas (National University of Colombia) for collection assistance; T. Karasov, M. Zaidem and members of the Burbano laboratory for comments on the manuscript; and M. Purugganan (New York University) for supporting R. Gutaker during the final stages of the project. We thank the Spanish National Research Council (CSIC) and the Ministry of Economy and Competitiveness for financing the project No. CGL 2010–19747, which facilitated the visits to Colombia and access to the collections of the Real Jardin Botanico Herbarium (MA). This study was funded by the Max Planck Society and its Presidential Innovation Fund (H.A.B).

Author contributions

R.M.G and H.A.B. conceived and designed the study with input from S.P. D.E., N.L.A., S.K., J.L.F.-A. and S.P curated, identified, pre-selected and contributed contemporary and historic potato samples. R.M.G carried out DNA extractions, hybridization captures and library preparations. R.M.G. performed the bioinformatic processing of sequencing data with input from C.L.W and H.A.B. C.L.W. carried out ploidy estimation of samples. R.M.G. performed the phylogeographic and population genomics analyses with input from H.A.B. R.M.G., S.P. and H.A.B. contributed to the interpretation of the data. R.M.G and H.A.B wrote the manuscript with input from all authors. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-019-0921-3.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to H.A.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Hernán A. Burbano

Last updated by author(s): May 2, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Illumina cBot Control Illumina MiSeq Control/RTA2 Illumina HiSeq Control/RTA Illumina bcl2fastq
Data analysis	Skewer v0.1.120 Flash v1.2.11 bwa mem v0.7.10 samtools v1.2 picard tools v2.3.0 MapDamage2.0 DeDup R v3.4.1 SPAdes v3.5.0 nQuire gatk v.3.8 bedtools v2.25.0 AdmixTools v4.1 mafft v7.310 RAxML v8.1.20 Plink1.9 Eigensoft

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data generated in this study are available in the European Nucleotide Archive (ENA) under project numbers PRJEB31013 and PRJEB31171, for UDG-treated and non-UDG-treated libraries, respectively.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Statistical sample size calculation was not performed. Samples of modern potatoes from South America were selected by the genebank unit of the International Potato Center (CIP) by constructing a "mini-core" of potato that represents genetic diversity in South America. We cannot control for sampling during historical time and the availability of historical herbarium samples is limited for destructive sampling. Therefore, we relied on the availability of old specimens for our project and collected as many as possible dated to 17th, 18th and 19th century. Subsequently, we adjusted the number of 20th century, modern European potatoes. Samples are described in Methods 1.1.
Data exclusions	Initially, all high-throughput sequenced individuals were used in the analyses. Firstly, we assessed the population structure and based on this assessment we removed individuals from i) the admixture graph construction and ii) genomic scans for evolutionary outliers. In particular, we did not include i) bitter potatoes, which from our sampling were deemed unrelated to any European potatoes, ii) admixed South American landraces, which had a signature of genomic influx from European potato; these are not purely native South American and would obscure our inference of origins of European potato. Detailed description of samples included in each analyses can be found in Methods 2.10-2.18.
Replication	All findings in this manuscript are not technically replicated, due to scarcity and uniqueness of historical samples. However we note that testing across multiple individuals provides a degree of experimental replication.
Randomization	Randomization is not relevant in this study, because there are no experimental groups in the study design.
Blinding	Genetic analyses performed in this study are sufficiently objective that investigator blinding is not required.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging