LETTER

Ancient genomes reveal early Andean farmers selected common beans while preserving diversity

Trucchi Emiliano(1,2)*#, Benazzo Andrea(2)*, Lari Martina(3)*, Iob Alice(2), Vai Stefania(3), Nanni Laura(4), Bellucci Elisa(4), Bitocchi Elena(4), Francesca Raffini(2), Xu Chunming(5), Jackson A Scott(5), Lema Verónica (6,7), Babot Pilar(7,8), Oliszewski Nurit(7), Gil Adolfo(9,10), Neme Gustavo(9,10), Michieli Catalina Teresa(11), De Lorenzi Monica(12), Calcagnile Lucio(13), Caramelli David(3), Star Bastiaan(14), de Boer Hugo(15), Boessenkool Sanne(14)$, Papa Roberto(4)$, Bertorelle Giorgio(2)$#.

(1) Department of Life and Environmental Sciences, Marche Polytechnic University, Ancona, Italy
(2) Department of Life Sciences and Biotechnology, University of Ferrara, Ferrara, Italy
(3) Department of Biology, University of Firenze, Firenze, Italy
(4) Department of Agricultural, Food, and Environmental Sciences, Marche Polytechnic University, Ancona, Italy
(5) Center for Applied Genetic Technologies, University of Georgia, Athens, Georgia, USA
(6) Universidad Nacional de Córdoba, Córdoba, Argentina
(7) Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina
(8) Instituto de Arqueología y Museo, Universidad Nacional de Tucumán, Tucumán, Argentina
(9) Instituto de Evolución, Ecología Histórica y Ambiente (CONICET & UTN FRSR), San Rafael, Argentina
(10) Museo de Historia Natural de San Rafael, Argentina
(11) Instituto de Investigaciones Arqueológicas y Museo "Prof. Mariano Gambier" - UNSJ, San Juan, Argentina
(12) Museo Arqueológico de Cachi, Cachi, Argentina
(13) Department of Mathematics and Physics "Ennio De Giorgi", University of Salento, Italy
(14) Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Oslo, Norway
(15) Natural History Museum, University of Oslo, Norway

* Contribution equally
$ Contribution equally
# Corresponding authors: ggb@unife.it; e.trucchi@staff.univpm.it

Introductory paragraph

All crops are the product of a domestication process that started less than 12,000 years ago from one or more wild populations [1, 2]. Farmers selected desirable phenotypic traits, such as improved energy accumulation, palatability of seeds and reduced natural shattering [3], while leading domesticated populations through several more or less gradual demographic contractions [2, 4]. As a consequence, erosion of wild genetic variation [5] is typical of modern cultivars making them highly susceptible to pathogens, pests and environmental change [6,7]. The loss of genetic diversity hampers further crop improvement programs to increase food production in a changing world, posing serious threats to food security [8,9]. Using both ancient and modern seeds, we analyzed the temporal dynamic of genetic variation and selection during the domestication process of the common bean (*Phaseolus vulgaris*) that occurred in the southern Andes. Here we show that most domestic traits were selected for prior to 2,500 years ago, with no or only minor loss of whole-genome heterozygosity. In fact, i) the majority of changes at coding genes and linked regions that differentiate wild and domestic genomes are already present in the ancient genomes analyzed here; ii) all ancient domestic genomes dated between 600 and 2,500 years ago are highly variable - at least as variable as modern genomes from the wild, and single seeds from modern cultivars show reduced variation when compared to ancient seeds, indicating that intensive selection within cultivars in the last centuries likely partitioned ancestral variation within different genetically homogenous cultivars. When cultivars from different Andean regions are pooled, genomic variation of the pool is higher than that observed in the pool of ancient seeds from north and central western Argentina. Considering that most desirable phenotypic traits are likely controlled by multiple polymorphic genes [10], a plausible explanation of this decoupling of selection and genetic erosion is that early farmers applied a relatively weak selection pressure [2] by using many phenotypically similar but genetically diverse individuals as parents. Our results imply that selection strategies during the last few centuries, as compared to earlier times, more intensively reduced genetic variation within cultivars, and produced further improvements focusing on few plants carrying the traits of interest at the cost of marked genetic erosion within Andean landraces.

Main text

The onset of domestication by early farmers has been suggested as the period during which the most intense genetic bottleneck affecting genome-wide diversity occurred [4,11]. Yet, artificial selection, possibly at different rates in different times, was likely a continuous process that produced both landraces and, more recently under modern breeding programs, high-yielding and more resistant elite cultivars [2,12,13]. Understanding when the majority of genetic diversity was lost, and how such loss is related with the intensity of the selection process, is not only relevant from an evolutionary perspective but it will also help planning more sustainable breeding options to revert or mitigate crop genetic erosion [14,15]. We directly address these questions by analyzing both modern and ancient genomes of common bean from South America.

Common bean constitutes one of the major sources of vegetable proteins worldwide. The wild progenitor of common bean originated in Mesoamerica and colonized South America along the Andes [16,17]. This natural colonization process, dated by whole genome analysis between 146,000 and 184,000 years ago [17], was accompanied by an intense and long bottleneck, as supported by the higher genetic diversity observed in the Mesoamerican as compared to the Andean gene pool [18]. Domestication of common bean occurred independently more or less simultaneously in both Mesoamerica and the Andes ca. 8,000 years ago, followed by divergence into distinct landraces due to drift, local selection, and adaptation [16].

We here obtained the first whole genome *shotgun* data from 30 ancient common bean seeds representing nine archaeological sites in north and central western Argentina (Supplementary Information S1-2,4, Supplementary Figure S1, Supplementary Table S1). All 30 seeds were dated between 2500 and 600 years before present (yrs BP) by AMS radiocarbon dating (Supplementary Information S3). Initial screening showed remarkable DNA preservation, with on average 44% (SD = 12%) endogenous DNA mapped to the common bean reference genome [17] (Supplementary Table S1). Unimodal length distributions with short fragment lengths (average = 65, SD = 22 bp) and purines enrichment upstream of break points were consistent with those expected for degraded DNA (Supplementary Table S1, Supplementary Figures S3 an S4). The presence of deamination patterns at the end of the reads varied from 0.76% to 32.75% C to T and G to A misincorporations, with half of the seeds showing percentages below 10% (Supplementary Table S1, Supplementary Figure S5). Damage patterns were partially explained by the location of the archaeological sites: all seeds from sites at >2,500 meters above sea level showed less than 10% base misincorporations (with many lower than 5%), whereas the majority of seeds from lower altitude exhibited levels above 15% (Supplementary Figure S6). The observed pattern shows that both favorable environmental conditions and the structure of the seed (*e.g.,* presence of an external cuticle) likely resulted in good DNA preservation, as previously suggested for quinoa seeds from the same sites [19]. A subset of the specimens (Figure 1a,b) was selected for further sequencing based on DNA preservation and representation of age and locations, resulting in a final dataset of 15 ancient bean genomes with coverage from 4.2X to 23.2X in the non-repeated regions (Supplementary Table S1). Whole-genome data from modern common bean accessions (domesticated landraces/cultivars: 9 Andean, 3 Mesoamerican; wild: 1 Andean, 1 Mesoamerican) and a closely related species (*P. hintonii,* 1 accession) were also included in the analyses (Supplementary Information S6; Supplementary Table S3). All statistical analyses (excluding the selection test where only mutations observed in all ancient seeds are considered) were performed on transversions to avoid the bias introduced by post-mortem deamination [e.g., 20].

Individual heterozygosity (probabilistically estimated as the variation parameter θ every thousand nucleotides; Supplementary Information S7) is higher in all ancient seeds than in each of nine modern landraces/cultivars from the Andean gene pool (Figure 1c, Supplementary Table S4), and the average heterozygosity is more than ten times higher in ancient than in modern seeds (0.144 vs. 0.013 θ/kbp, respectively, Mann-Whitney U, $P<0.001$). This difference remains high (0.110 vs. 0.013 θ/kbp) and highly significant when the ancient seeds with the lowest coverage are excluded (see Supplementary Information 7). The effect of decreased heterozygosity due to the selfing procedure applied to modern accessions in seed banks was tested in two ways: by pooling pairs of modern seeds and estimating θ/kbp in different seeds pairs, and by increasing the heterozygosity values in modern seeds as a function of the number of selfing cycles to which they were subjected (Supplementary Information S8). As expected, variation in modern pairs increased compared to single seeds, but it is on average still significantly smaller (at least 2.5 times) than that observed in ancient seeds (Mann-Whitney U test, $P<0.005$, Supplementary Figure S9; Supplementary Table S5). It is possible that our pairing strategy did not entirely recover the heterozygosity just before the implementation of the seed bank protocol. A more exact estimate cannot be obtained since the detailed protocols applied to each modern Andean cultivar in seed banks are not known, and some level of outcrossing during this procedure cannot be excluded. Furthermore, the average θ/kbp is likely increased in modern seed pairs because our paired seeds belong to different landraces (i.e., the increase in the seed pairs is inflated by the genetic structure among landraces). When single seed heterozygosity was conservatively doubled or quadrupled in modern seeds subjected to one or two cycles of selfing, respectively, average variation remained significantly higher in the ancient

139 seeds (see Supplementary Information S8). We therefore conclude that single modern seeds
140 representative of modern landraces have lower genetic variation than single ancient seeds.
141
142 The modern genome available from a wild Andean seed [21], and two publicly released but
143 unpublished genomes from wild Argentinean seeds (Supplementary Information S7), have a lower
144 heterozygosity than the wild Mesoamerican seed in our dataset but comparable to that observed
145 in the ancient Argentinean seeds (Figure 1c, Supplementary Figure S7). This result is a) consistent
146 with the bottleneck that occurred during the natural colonization of the Andes from Mesoamerica
147 [18], b) it indirectly supports the finding that modern Andean domestic seeds have lower diversity
148 than ancient domestic seeds, because otherwise we should exclude any decrease of heterozygosity
149 throughout the whole domestication process up to the modern cultivars, and c) it appears
150 compatible with a minor effect of the initial Andean domestication on bean genetic diversity. With
151 the current dataset, it cannot be excluded that heterozygosity was initially reduced following
152 domestication, but subsequently restored by mutation. Such an effect would, however, require a
153 ten times higher mutation rate than that estimated on the basis of the divergence time between *P.*
154 *vulgaris* and *P. hintonii* (Supplementary Information S10. These findings challenge the conventional
155 view of the domestication bottleneck resulting in severe genetic erosion [3] and is consistent with
156 recent observations in maize, sorghum and barley [22].
157
158 The lack of a temporal trend across ancient seeds of different age (Figure 1c) indicates that
159 genome-wide diversity was largely maintained between 2,500 and 600 years ago by agricultural
160 practices of Andean societies. Therefore, loss of diversity within each cultivar occurred more
161 recently, certainly less than 600 years ago and likely in the last century [13]. A similar loss of
162 diversity characterizes the recent domestication history of horses [20]. The absence of data on
163 seeds from the last 600 years currently hinders a direct reconstruction of the temporal dynamic of
164 genomic diversity during the last six centuries.
165
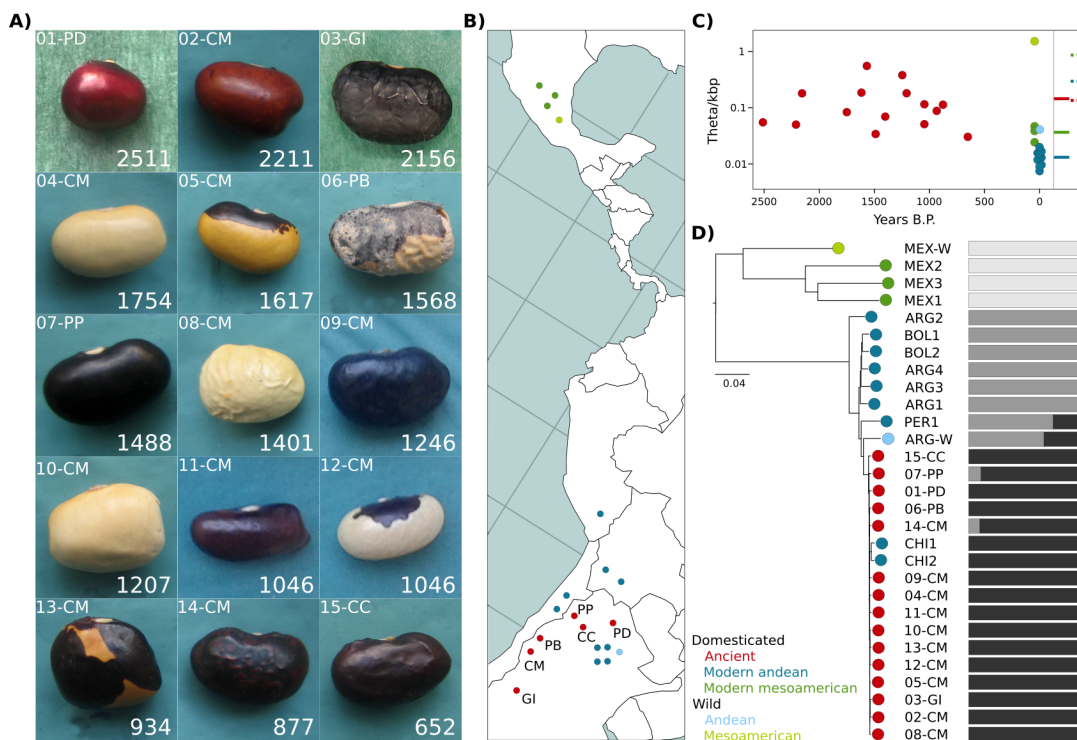
166
167
168 **Figure 1. Diversity in ancient and modern common bean**. **A)** Pictures of the 15 common bean seeds sequenced for whole-genome
169 analyses: sample ID with two letters referring to the archaeological site of origin as shown in panel B is reported in the top-left corner
170 of the pictures (see Supplementary Information S1); AMS radiocarbon age in yrs BP is reported in the bottom-right corner. **B)** Map

179 When estimating genetic variation in groups of modern Andean (0.294 θ/kbp) or Mesoamerican
180 seeds (0.860 θ/kbp), we observe a substantial increase compared to the average individual
181 estimates (Figure 1c), indicating that modern genomic diversity is highly structured in different,
182 highly homogeneous, cultivars. The abandonment and extinction of any modern cultivar would
183 therefore result in the significant loss of its private fraction of the whole crop diversity [8,23]. On
184 the contrary, the estimate of genetic diversity obtained by pooling all ancient individuals is very
185 similar to the average of the 15 individual estimates (0.134 θ/kbp; Figure 1c), implying that these
186 ancient genomes represent a single genomic pool. Ancient seeds, as a group, do not reach the
187 level of variation observed in the pooled modern cultivars. This is not unexpected considering that
188 a) the latter have a long, complex and partially independent history of recent landrace
189 establishment, and b) the former (the ancient domestic seeds) were used during the Holocene by
190 human populations with shared subsistence practices, which included the use of wild beans [24]. A
191 large fraction of the ancient seeds could individually recover most of the ancient crop diversity in
192 that area, while modern seeds never reach individual heterozygosity values close to the pooled
193 group from Chile, Bolivia, Argentina and Peru, or the groups obtained by separately pooling the
194 Chilean (0.045 θ/kbp), the Bolivian (0.102 θ/kbp), or the Argentinean (0.198 θ/kbp) landraces. A
195 more widespread sample of ancient beans might also reveal that ancient populations harbored
196 more genetic diversity than pooled cultivars.

198 Three main genetic groups were revealed by different clustering methods (Supplementary
199 Information S9) and a whole-genome NJ tree (Figure 1b; Supplementary Figures S10-S11). The
200 largest divergence is observed between Mesoamerican and Andean genomes, while the Andean
201 seeds are further partitioned into two less differentiated groups. All ancient seeds belong to the
202 same genomic clade, indicating that the same or similar ancient landraces were in use in Argentina
203 for about 2,000 years. - This clade also includes modern landraces used today on the other side of
204 the Andes, in Chile. The so-called Chilean race [25] therefore appears to be the direct descendants
205 of the seeds used in Argentina before the Incaic conquest of the area occurred at the end of the
206 15th century [26]. We interpret this result as a consequence of the trade of goods, including crop,
207 between the eastern (in Argentina) and western (in Chile) slopes of the Andes, with early farmers
208 relying on Llama caravans for regular-long distance exchanges [27]. Modern Argentinian cultivars,
209 instead, do not trace their ancestry back to the local ancient cultivar but were likely introduced in
210 the area some time during the last 600 years.

212 A gene-by-gene selection scan (Figure 2a) was performed to discriminate early (>2,500 years ago)
213 from late (<600 years ago) selection targets, by taking into account the different levels of genetic
214 drift expected in the two time periods of different length (Supplementary Information S11). We
215 found that: *i*) a selection signature is present in 443 out of 27,000 genes tested (FDR<0.001); *ii*)
216 selection affected many more genes (ca. 4.2 times more) in the early compared to the late phase
217 (Fig. 2b), suggesting that recent improvement of cultivars strongly affected few genes, but the
218 main genomic turnover occurred several thousand years ago; if adjacent genes are merged in
219 "sweep blocks", thus considering that genes under selection may produce selection signals also in
220 adjacent genes, the ratio between ancient and recent sweeps remains larger than 3 (see
221 Supplementary Information S11); *iii*) genes selected in the early phase belong to functional clusters

222 mainly related to glycerol metabolism, carbohydrate and sugar transport and metabolism,
223 intracellular transport (endosome membrane), regulatory elements (nucleotide binding),
224 modification of proteins and glycosylation, whereas recent selection primarily targeted traits
225 involved in immunity and defense, regulatory elements and transmembrane transport (Figure 2c;
226 Supplementary Table S7). Evidence for selection on sugar biosynthesis genes at the early stages of
227 domestication has also been found in ancient maize from 2,000 and 750 years ago [28]. One of the
228 functional clusters identified in the early selected genes includes the MYB-domain genes that are
229 important regulatory elements of development, metabolism and responses to biotic and abiotic
230 stresses [29], and have been suggested as causative of different phenotypic changes associated to
231 the shattering trait in common bean [30], maize and rice [4,31].
232
233 Genomic studies of ancient DNA are much less numerous in plants compared to animals remains
234 [32-33], but have already provided fundamental insights into the process of domestication [28,34-
235 38]. Our results show that common bean seeds, and likely legume seeds in general, can be an
236 excellent source of high quality ancient DNA, especially if seeds are preserved in favorable climatic
237 conditions (*i.e.*, cold and dry environments). The dynamic of domestication is still debated [22],
238 but, as recently summarized [2], the consensus is that domestication was slow and gradual, with
239 selected traits emerging in association with population decline and loss of genetic variation. Here,
240 we contribute to this debate with the first study on ancient bean genomics, and provide an
241 alternative example in which selection and loss of variation are decoupled during the process of
242 domestication. Our data reveals that in common bean early agriculture in the Andes genomic
243 diversity was preserved within single seeds, whereas more recent breeding practices produced
244 structured landraces that are internally more homogenous. At the same time, however, assuming
245 that the genetic changes we infer correspond at least in part to phenotypic changes, early
246 agriculture was also very efficient in selecting most of the desirable traits now typical of this crop.
247 We hypothesize that these patterns are the result of the larger number of seeds used by early
248 farmers as founders in each generation, all likely displaying the selected trait(s) but heterogeneous
249 in the rest of the genome. Initial improvement of the common bean was therefore based on serial
250 soft sweeps, as has also been suggested for the domestication of maize [10]. Encompassing several
251 thousand years, and likely assisted by cultivar exchanges and hybridization with wild plants
252 [2,25,39-40], such breeding practice allowed many traits to be selected for, without the significant
253 loss of genomic variation that has likely occurred in more recent times within landraces and
254 cultivars.
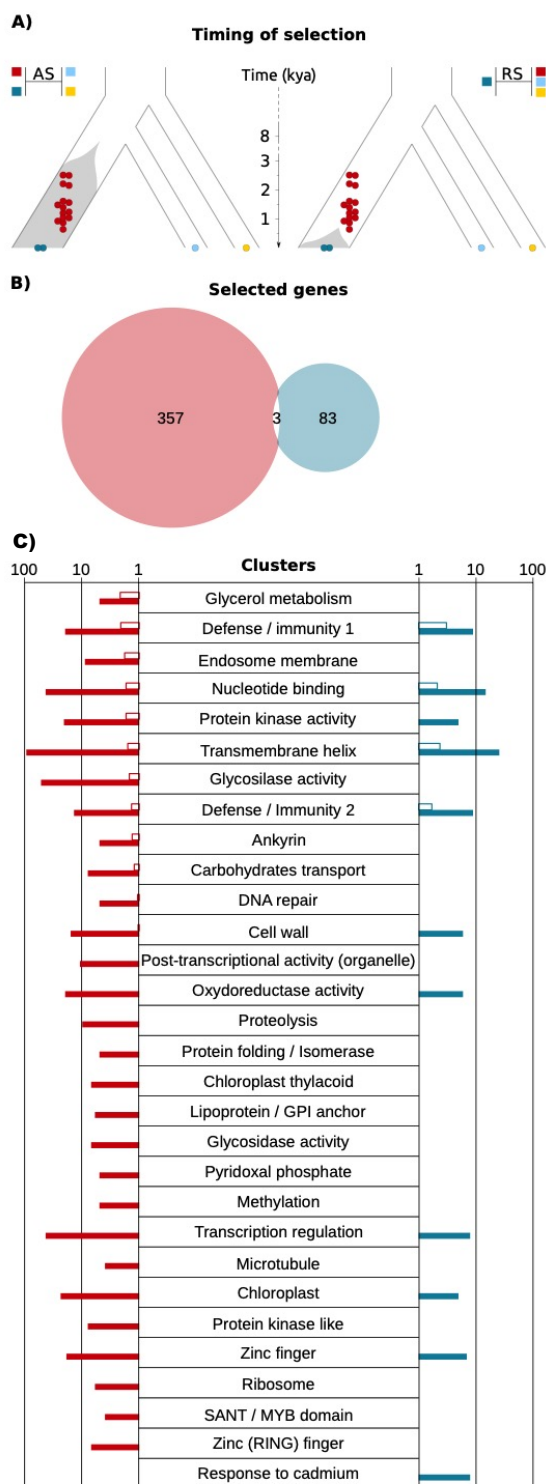
Figure 2. Timing of selection in common bean from South America. All genes (27,000) in the common bean genome were tested for a significant enrichment of fixed alternative alleles according to the topologies described below, taking into account the different level of drift expected in periods of time of different length (see Supplementary Information S11 for details). **A)** Alternative topologies for a SNP: on the left (AS: ancient selection) the SNP is fixed for an alternative variant in both the ancient seeds from Argentina (red circles) and the modern cultivars from Chile (blue circles) as compared with the wild Andean *P. vulgaris* seed (pale blue circle) and the outgroup (*Phaseolus hintonii*, yellow circle); on the right (RS: recent selection) the SNP is fixed for an alternative variant in the modern cultivars from Chile only as compared with the ancient seeds, the wild Andean and the outgroup; the gray shade represents the temporal fixation of an alternative allele; ancient seeds age as well as the putative divergence time between Andean domesticated and wild *P. vulgaris* is reported. Using the genome of the wild Mesoamerican *P. vulgaris* as outgroup instead of *P. hintonii* produced largely similar results (see Supplementary Table S7). **B)** Number of genes significantly enriched with SNPs in RS (blue) and/or AS (red) topology (FDR<0.001, top 1% for the proportion of SNPs within gene ± 1kb, and with an *Arabidopsis thaliana* ortholog). **C)** Functional gene clusters (see Supplementary Information S11) in RS (blue) or AS (red) enriched groups (the complete list of genes can be found in Supplementary Table S7). Filled bars: max gene count in the cluster; empty bars: enrichment score of the cluster (only scores >1 are reported).

264 **Data accessibility**
265 Sequenced raw reads from ancient samples are publicly available as NCBI Bioproject (ID:
266 PRJNA574560). Unpublished modern sample genomic data are available with the following NCBI
267 Accession Codes: SRR10161640; SRR10161629; SRR10161767; SRR10161647; SRR10161646;
268 SRR10161645; SRR10161601; SRR10161592; SRR10161584; SRR10161684; SRR10161683;
269 SRR10161697; SRR10161690; SRR10161688.
270

271 **Competing Interests statement**
272 The authors declare no competing interests.
273

References

1. Diamond, J. Evolution, consequences and future of plant and animal domestication. *Nature* **418,** 700-707 (2002).

2. Purugganan, M. D. Evolutionary Insights into the Nature of Plant Domestication. *Curr. Biol*. **29,** 705-714 (2019).

3 Meyer, R. S., & Purugganan, M. D. Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet*. **14,** 840-852 (2013).

4. Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127,** 1309-1321 (2006).

5. Van de Wouw, M., Kik, C., van Hintum, T., van Treuren, R. & Visser, B. Genetic erosion in crops: concept, research results and challenges. *Plant. Genet. Resour*. **8,** 1-15 (2010).

6 Babiker E. M. et al. Mapping resistance to the Ug99 race group of the stem rust pathogen in a spring wheat landrace. *Theor. Appl. Genet.* **128,** 605-612 (2015).

7. Dale, J. et al. Transgenic Cavendish bananas with resistance to Fusarium wilt tropical race 4. *Nat. Comm*. **8,** 1496 (2017).

8. Esquinas-Alcázar, J. Protecting crop genetic diversity for food security: political, ethical and technical challenges. *Nat. Rev. Genet.* **6,** 946 (2005).

9. Gepts, P. Plant genetic resources conservation and utilization. *Crop. Sci*. **46,** 2278-2292 (2006).

10. Beissinger, T. M. et al. Recent demography drives changes in linked selection across the maize genome. *Nat. Plants* **2,** 16084 (2016).

11. Hyten, D. L. et al. Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl Acad. Sci. USA* **103**, 16666-16671 (2006).

12. Fuller, D. Q. et al. Convergent evolution and parallelism in plant domestication revealed by an expanding archaeological record. *Proc. Natl Acad. Sci. USA* **111**, 6147-6152 (2014).

13. Khush, G. S. Green revolution: the way forward. *Nat. Rev. Genet*. **2**, 815-822 (2001).

14. Fu, Y. B. Understanding crop genetic diversity under modern plant breeding. *Theor. Appl. Genet*. **128**, 2131-2142 (2015).

15. Bevan, M. W. et al. Genomic innovation for crop improvement. *Nature* **543**, 346-354 (2017).

16. Bitocchi, E. et al. Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. *Proc. Natl Acad. Sci. USA*, **109**, E788-E796 (2012).

17. Schmutz, J. et al. A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet*. **46**, 707-713 (2014).

18. Bitocchi, E. et al. Beans (*Phaseolus* ssp.) as a model for understanding crop evolution. *Front. Plant Sci*, **8**, 722 (2017).

19. Winkel, T. et al. Discontinuities in quinoa biodiversity in the dry Andes: An 18-century perspective based on allelic genotyping. *PLoS One* **13**, e0207519 (2018).

20. Fages, A. et al. Tracking five millennia of horse management with extensive ancient genome time series. *Cell*, **177**, 1419-1435 (2019).

21. Rendón-Anaya, M. et al. Genomic history of the origin and domestication of common bean unveils its closest sister species. *Genome Biol*. **18**, 60 (2017).

22. Allaby, R. G., Ware, R. L. & Kistler, L. A re-evaluation of the domestication bottleneck from archaeogenomic evidence. *Evol appl*, **12,** 29-37 (2019).

316 23. Castañeda-Álvarez, N. P. et al. Global conservation priorities for crop wild relatives. *Nat. Plants* **2**, 16022
317 (2016).

318 24. Pochettino, M. L. & Scattolin M. C. Identificación y significado de frutos y semillas carbonizados de sitios
319 arqueológicos de la ladera occidental del Aconquija, Prov. Catamarca, Rca. Argentina. *Rev. Mus. La Plata,*
320 *Antropol.* **9**: 169-181 (1991).

321 25. Singh, S. P., Gepts P. & Debouck D. G. Races of common bean (*Phaseolus vulgaris*, Fabaceae). *Econ. Bot.*
322 **45**, 379-396 (1991).

323 26. Williams, V.I. Formaciones sociales en el noroeste argentino: variabilidad prehispánica en el surandino
324 durante el Periodo de Desarrollos Regionales y el estado Inca. Rev. Haucaypata **9**, 62-76 (2015).

325 27. Núñez, L. & Nielsen A.E. *En ruta: Arquelogía, Historia y Etnografía del Tráfico Surandino* (Encuentro
326 Grupo Editor, Córdoba, 2011).

327 28. Da Fonseca, R. R. et al. The origin and evolution of maize in the Southwestern United States. *Nat. Plants*
328 **1**, 14003 (2015).

329 29. Dubos, C. et al. MYB transcription factors in Arabidopsis. *Trends. Plant Sci*. **15**, 573-581 (2010).

330 30. Rau, D. et al. Genomic dissection of pod shattering in common bean: mutations at non-orthologous loci
331 at the basis of convergent phenotypic evolution under domestication of leguminous species. *Plant J.* **97**,
332 693-714 (2019).

333 31. Saitoh, K., Onishi, K., Mikami, I., Thidar, K. & Sano, Y. Allelic diversification at the C (OsC1) locus of wild
334 and cultivated rice: nucleotide changes associated with phenotypes. *Genetics* **168**, 997-1007 (2004).

335 32. Estrada, O., Breen, J., Richards, S. M. & Cooper, A. Ancient plant DNA in the genomic era. *Nat. Plants* **4**,
336 394 (2018).

337 33. Brunson, K. & Reich, D. The promise of paleogenomics beyond our own species. *Trends Genet.* **35**, 319-
338 329 (2019).

339 34. Mascher, M. et al. Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication
340 history of barley. *Nature Genet*. **48**, 1089-1093(2016).

341 35. Ramos-Madrigal, J. et al. Genome sequence of a 5,310-year-old maize cob provides insights into the
342 early stages of maize domestication. *Curr. Biol*. **26**, 3195-3201 (2016).

343 36. Wagner, S. et al. High Throughput DNA sequencing of ancient wood. *Mol.Ecol*. **27**, 1138-1154 (2018).

344 37. Kistler, L. et al. Multiproxy evidence highlights a complex evolutionary legacy of maize in South America.
345 *Science*, **362**, 1309-1313 (2018).

346 38. Smith, O. et al. A domestication history of dynamic adaptation and genomic deterioration in sorghum.
347 *Nat. Plants* **5**, 369-379 (2019).

348 39. Lema, V. Non domestication cultivation in the Andes: plant management and nurturing in the
349 Argentinean Northwest. *Veg hist archaeobot*, **24**, 143-150 (2015).

350 40. Oliszewski, N. & Babot P. Procesos de selección del poroto común en los valles altos del noroeste argentino
351 en tiempos prehispánicos. Análisis micro y macroscópico de especímenes arqueobotánicos. In: *Avances y*
352 *desafíos metodológicos en arqueobotánica: miradas consensuadas y diálogos compartidos desde Sudamérica*
353 (eds. Belmar C. & Lema, V.) 301-324 (Facultad de Patrimonio Cultural y Educaciòn Universidad SEK Chile,
354 2015).

SUPPLEMENTARY METHODS AND RESULTS

399 **S1. Ancient beans collections and archaeological context**
400

401 Ancient bean seeds were selected from five museum collections in Argentina: Museo de La Plata (La Plata,
402 Buenos Aires, Argentina); Museo de Historia Natural de San Rafael (Parque Mariano Moreno, San Rafael,
403 Argentina); Instituto de Investigaciones y Museo Arqueológico "Prof. Mariano Gambier", Universidad
404 Nacional de San Juan (San Juan, Argentina); Instituto de Arqueología y Museo, Universidad Nacional de
405 Tucumán (Tucumán, Argentina); Museo Arqueológico Pio Pablo Díaz (Cachi, Salta, Argentina).
406

407 Seeds were originally collected from nine archaeological sites located in different geographical regions of
408 north and central-western Argentina (Figure 1a, main text). A brief description of each site is reported
409 below. Codes for each seed, referring to Figure 1b (main text) and Figure S1, are reported in Table S1.
410

411 *Puente del Diablo*
412 Puente del Diablo (SSalLap20) archaeological site is a cave situated in the northern sector of the Calchaqui
413 Valleys in a pre-puna landscape (Salta province). It was excavated in the 1970s and includes both domestic
414 and funerary contexts. This is a multicomponent site with burials dating back to 10,000 yrs BP, remains
415 corresponding to burials and temporal occupations of hunter gatherers of ca. 3,000 yrs BP and also from ca.
416 2,000 yrs BP corresponding to the first agro-pastoralist societies in the area. From these last occupations,
417 remains of *Cucurbita maxima*, *Prosopis* spp. and cactaceae were recovered together with remains of
418 rodents (*Lagidium* sp.), *Cervidae* sp. and camelids (*Lama* sp.) (Lema, 2009, 2015).
419

420 *Gruta del Indio*
421 Gruta del Indio is an archaeological site located in Central Western Argentina, close to San Rafael city, in
422 Monte phytogeographic province, at 700 m asl. Gruta del Indio was occupied by hunter-gatherers since the
423 Pleistocene-Holocene transition ca. 10,500 years ago (Semper and Lagiglia 1968, Long et al 1998). Hunter-
424 gatherers occupied the cave until at least ca. 1,900 yrs BP, when a strong domestic plant record appears in
425 the archaeological record. This domestic archaeobotanical record include *Zea mays*, *Chenopodium quinoa*,
426 *Cucurbita*, and *Phaseolus vulgaris* seeds, and *Zea mays* remains, such as canes and starches (Semper and
427 Lagilia 1968, Lagiglia 1999). This domestic plant context (the oldest in the region) is isolated in time and
428 space, and is associated to human bone remains, indicating the use of the cave as a cemetery. This
429 archaeological context was called "Atuel II culture" (Semper and Lagiglia 1968) and all radiocarbon dates are
430 between 2,100 and 1,900 yrs BP (Gil et al 2014). Stable isotopes analyses from human bone suggest that
431 the *Zea mays* was probably part of the diet but not as major component (Gil et al 2010). The *Phaseolus*
432 *vulgaris* beans were found in a grass container holding ca. 500 grs. of seeds. Another grass container was
433 full of *Chenopodium quinoa* seeds in the same archaeological mortuary context. After Atuel II, there is little
434 evidence of human occupation and temporally close to the arrival of the Spanish (ca. 500 yrs BP).
435

436 *Los Morrillos, Río Salado, Cerro Calvario and Punta del Barro (San Juan sites)*
437 The San Juan sites are in the two western Departments of the Province of San Juan, Iglesia (to the north) and
438 Calingasta (to the south). They are in a southern Andean area that extends between 28º25'S and 32º33'S on
439 the arid eastern slope of the Andes. This territory includes (from west to east) the "Cordillera de los Andes",
440 two wide valleys between 1,900 and 1,600 m asl and another system of high altitude but of older age than
441 the mountain range called "Precordillera of La Rioja, San Juan and Mendoza"; its maximum summits
442 constitute the eastern limit of these Departments.
443

444 Los Morrillos (caves 2 and 3) and Río Salado cave
445 The first evidence of agricultural and breeding (*Lama glama*) activity in the eastern slope of the Andes in San
446 Juan province did not originate in the area, but were introduced together with the domesticated species. The
447 human groups that introduced and developed agriculture were located in favorable small sites corresponding
448 to the exit of the cordilleran streams in the great plain of the high foothills (approximately 2,900 m asl). Such
449 human groups cultivated in summer, gathered wild fruits and ñandu eggs, and entered the high Andean
450 valleys for the hunt of the guanaco (*Lama guanicoe*). They mostly inhabited natural caves in the area.
451 Examples of such caves are Los Morrillos (caves 2 and 3). In the Los Morrillos caves, the thick layer with
452 agriculture evidence overlapped with an older layer of hunter gatherers. By contrast, in the Río Salado cave

453 there is only a thick level with agricultural remains (Gambier 1977, 1988, 2000, Roig 1977). Los Morrillos caves
454 were excavated in seven long seasons between 1969 and 1977, while the Río Salado cave was excavated in
455 1972. The three caves correspond to habitation sites without specific areas of activity. The botanical remains
456 were found among the anthropic sediments with diverse remains: lithic, ceramic, artifacts in wood and bone,
457 textiles, etc.).
458
459 <u>Cerro Calvario sites (Calingasta Valley) and Punta del Barro site (Iglesia Valley)</u>
460 The same agricultural and breeder groups described previously, with some foreign influences, later
461 descended from the "Cordillera de Los Andes" towards the valleys at lower altitude and settled in open areas
462 with the possibility of irrigation from small springs or streams. This is the case for the Cerro Calvario sites in
463 the Calingasta Valley and Punta del Barro in the north of the Iglesia Valley (Basurero Norte, Basurero 3)
464 (Gambier 1988, 2000; Michieli 2016).
465
466 <u>Punta del Barro "1º canal" site</u>
467 Around 1200 AD the local groups settled in the valleys of the Eastern slope of Los Andes in the current
468 province of San Juan developed a large-scale agricultural system based on irrigation by means of large
469 channels connected to the rivers. The agricultural sites of Punta del Barro "1º canal" represent a residential
470 area, excavated in 2011, which included an occupational floor almost at surface level with storage and
471 combustion wells. Various cultivated botanical remains were found in one of the storage wells: corn, squash,
472 quinoa and beans. It was related to a large hydraulic channel and three tombs with excellent preservation of
473 both the bodies and the woven garments that enveloped them and other items (ceramics, pumpkin
474 containers, wooden objects, etc.) (Michieli 2015).
475
476 ***Punta de la Peña***
477 Punta de la Peña 9 is a residential site with stone-walled structures, blocky shelters, bedrock mortars,
478 caches, open air activity areas, and rock art that was continuously occupied by agro-pastoralist farmers
479 between ca. 2,000 to 400 yrs BP -post-Hispanic time (Babot et al 2006, López Campeny et al 2017, Somonte
480 and Cohen 2006). It comprises processing, discard, workshop, funerary and ritual areas spatially segregated.
481 The site is located on the southern margin of the Las Pitas River at an elevation of 3,665 m asl in the Salty
482 Puna desert (Catamarca province). Subsistence was based on llama (*Lama glama*) herding, vicuña (*Vicugna*
483 *vicugna*) hunting, small-scale cultivation, and plant gathering and exchange. Wild, domesticated and weed
484 plant macro and micro-remains belong to a number of native and foreign taxa as *Chenopodium quinoa*,
485 *Solanum tuberosum, Oxalis tuberosa, Phaseolus vulgaris, Prosopis* sp., *Geoffroea decorticans, Zea mays*,
486 *Amaranthus caudatus/A. mantegazzianus, Lagenaria siceraria, Cucurbita* sp*., Canna edulis* and Cyperace
487 and Cactaceae species (Babot 2009, Rodríguez 2013). Alero 1 is a blocky shelter rich in well preserved plant
488 remains from the first millennium DC, placed in the sector I of the Punta de la Peña 9 site (Babot et al 2013).
489
490 ***Pampa Grande***
491 Pampa Grande is an archaeological locality including seven caves explored in the 1970s. It is located in the
492 Las Pirguas mountain range (Salta province), between 2,500 and 3,000 m asl in a montane grassland and
493 forest landscape corresponding to the highest zone of the *Yunga* biogeographical province. Ancient remains
494 are mainly related to funerary contexts and few occupational areas, corresponding to Candelaria groups.
495 The age of these occupations range between 400 – 3,000 yrs BP and corresponds to the first agro-
496 pastoralist societies in the area. Remains related with subsistence include domestic, wild and hybrid plant
497 species of *Cucurbita maxima* and *Phaseolus vulgaris. Phaseolus lunatus*, *Lagenaria siceraria*, *Zea mays*,
498 *Arachis hypogaea*, *Prosopis* spp., *Geoffroeae decorticans*, tubers, *Capsicum* spp. and other wild and weed
499 forms were recovered. Remains of wild and domestic camelids were also recovered together with fish,
500 rodent and bird bones (Baldini et al 2003, Lema 2010).
501
502 ***Cueva de los Corrales***
503 Cueva de Los Corrales 1 is an archaeological site located in the Quebrada de Los Corrales (El Infiernillo,
504 Tucumán) at 3,000 m asl in a dry shrub / grassland landscape, corresponding to the highest zone of the
505 Yunga biogeographical province (Oliszewski et al 2008). It is a cave that was occupied at different times
506 throughout 2,400 years (ca. 3,000-600 yrs BP), from transitional time between hunter-gatherer groups to

507 agro-pastoralist groups until purely agro-pastoralist time (Oliszewski et al. 2018). Remains related to
508 subsistence include domestic camelids and dasiphodids; wild plants: *Prosopis nigra*, *Geoffroeae decorticans*,
509 *Celtis tala* and *Trichocereus* sp.; cultivated plants: *Phaseolus vulgaris*, *Zea mays* and *Chenopodium quinoa*
510 and wild *Cucurbita maxima* (Oliszewski and Arreguez 2015, Oliszewski and Babot 2015, Lema 2017).

511 ## S2. Seeds selection

512

513 Thirty ancient bean seeds (Figure S1) were selected for preliminary screening based on the site locations
514 and archaeological context. All seeds had a clear domestic phenotype. After visual inspection, we selected
515 seeds with different morphological features (e.g. coat color and dimension) and good preservation (e.g.,
516 less visible cracks or damages in the outer coat).

517

518
519
520 **Figure S1.** Pictures of the 30 seeds sequenced at low coverage for preliminary screening of aDNA content and preservation.
521 Archaeological site information for each seed is reported in Supplementary Table S1.
522

S3. AMS radiocarbon dating

A fragment of ca. ¼ of the cotyledon was removed from each seed and used for AMS radiocarbon dating at CEDAD, Center of Applied Physics, Dating and Diagnostics, Department of Mathematics and Physics "Ennio De Giorgi", University of Salento. Conventional radiocarbon ages and calibrated ranges for 30 seeds are reported in Table S1. Calibration was performed using OxCal Ver. 3.10 based on atmospheric data (Reimer et al 2009). Distribution curves of the calibrated age for each seed, and additional details on the dating procedure are available on request.


S4. Ancient DNA extraction and sequencing

DNA extraction and library preparation were performed at the Molecular Anthropology and Paleogenetic Laboratory of the Department of Biology, University of Florence, using facilities that are exclusively dedicated to processing ancient DNA and following all the necessary precautions to minimize contamination from exogenous DNA. One sample (09-CM) was extracted and a library built at the ancient DNA laboratory of the University of Oslo. Negative controls were included in each step of the molecular analysis. All seeds were photographed prior to sampling (Figure S1). After removing the cuticle, each seed was UV irradiated in a crosslinker for 45 minutes. The embryo and a portion of cotyledon were then pulverized by hand with mortar and pestle that had been decontaminated with bleach and UV. Half of the powder, was used for DNA extraction following a high-salt CTAB extraction protocol according to the Small Fragment Protocol of Qiagen DNeasy *mericon* Food Handbook. The amount of powder used in the DNA extraction ranged between 48 and 240 mg, depending on the size of the seed. Food Lysis Buffer was substituted with a homemade CTAB high salt extraction buffer according to the following recipe: 100mM Tris-HCl pH 8, 25mM EDTA pH 8, 4M NaCl, 2% CTAB, 0.3% β-mercaptoethanol. An aliquot of each extract was converted into a sequencing library with 7 bp P7 indexes according to Meyer & Kircher 2010. No UDG treatment was performed. The fill-in reaction was split in 4 aliquots that were amplified separately in 100µl reaction mixes with 0.25mM dNTPs mix (New England Biolabs), 0.30 mg/ml BSA (New England Biolabs), 2.5U Agilent Pfu Turbo Polymerase (Agilent Technologies) and 0.4µM of each primer. The following indexing PCR profile was performed by default on each library: 95°C 2', [95°C 30'', 58°c 30'', 72°C 1'] x 14, 72°C 10'. DNA quantification and quality control were performed on a Agilent 2100 Bioanalyzer using the Agilent DNA 1000 kit, following manufacturer's instructions. All libraries and extraction blanks showed flat BioAnalyzer profiles (see Figure S2), but were included in the pool of the first sequencing round (see below). The libraries were sequenced at the Norwegian Sequencing Centre, University of Oslo, Norway, where all samples were pooled in equimolar ratios and paired-end sequenced on the Illumina HiSeq 2500. First, samples were screened for DNA preservation by low-coverage sequencing. Based on the results of this first sequencing run and on sample characteristics (age, location, phenotype, mapping quality, age-related damage patterns), a subset of 18 samples was re-pooled and sequenced further on four Hiseq2500 lanes. Additionally, one sample (O1-PD) was sequenced on one lane each on the HiSeq 4000 platform. One sample (09-CM) was sequenced on the Hiseq 2500 in a different pool from all other beans. Based on the final results we further selected 15 samples with coverage higher than 4X for downstream population genomic analyses (Table S1).
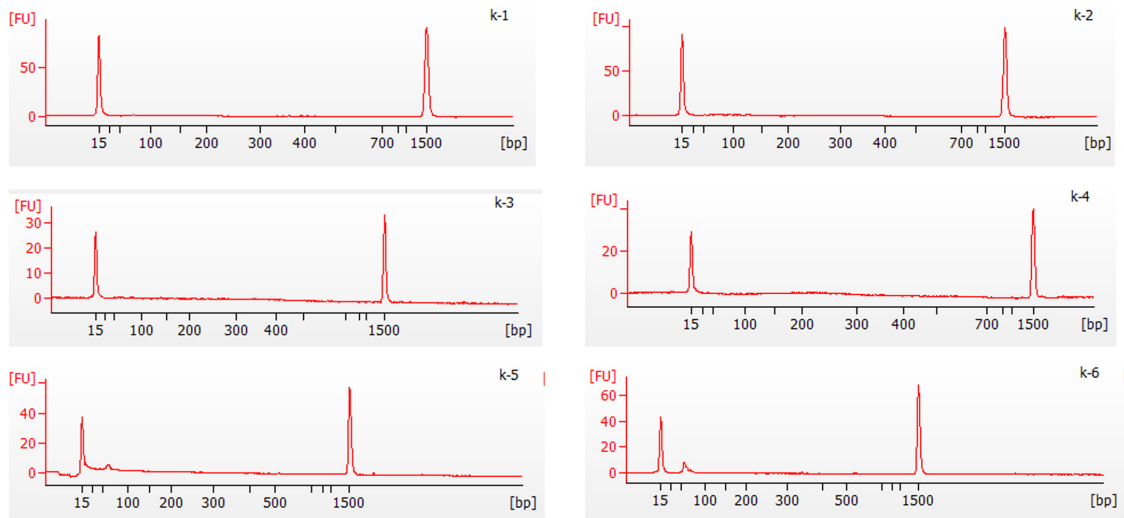
566
567 **Figure S2.** BioAnalyzer profiles of library (k-1, k-3, k-5) and extraction (k-2, k-4. k-6) blanks.
568
569
570 **S5. DNA damage assessment and mapping to the reference genome**
571
572 Paleomix v.1.2.12 (Schubert et al 2014), a pipeline specifically designed for aDNA, was employed for the
573 filtering and mapping of aDNA data. Adapter trimming, collapsing of overlapping mate-pairs, trimming of
574 low-quality bases and removal of reads shorter than 25 bp was performed with AdapterRemoval (Lindgreen
575 et al 2012). Filtered reads were aligned to the *Phaseolus vulgaris* reference genome v2.1 (Schmutz et al
576 2014), downloaded from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html) using bwa aln (Li and
577 Durbin 2009) and retaining reads with mapping quality >= 25. PCR duplicates were filtered using the
578 Paleomix rmdup_duplicates option. Fragment length distributions, break point and misincorporation
579 patterns were explored in each sample (Figures S3, S4, S5) using MapDamage2.0 (Jónsson et al 2013)..
580 Indel-realigned bams were generated using the GATK Indel realigner (McKenna et al 2010). Realigned and
581 filtered bam files were then indexed and sorted with Samtools-0.1.19 (Li et al 2009). The percentage
582 endogenous DNA and coverage were determined from these realigned and filtered bam files.
583
584 An evaluation of the DNA preservation (percentage of endogenous DNA and deamination pattern) in the
585 ancient bean seeds as a function of sample age and site location is reported in Figure S6. Altitude, and not
586 age, seems to be a good predictor of deamination patterns.
587
588 Almost all of the following analyses including both ancient and modern (see section "Modern accessions
589 genomic data) samples were based on transversions, since transitions are more affected by post-mortem
590 damages than transversion and could reflect DNA damage rather than true DNA variation (e.g., Hofreiter et
591 al 2012). Only the selection test (Supplementary Information S11) considered both transitions and
592 transversions because in this test only changes observed in all the 15 ancient seeds are used. We estimate
593 that these simultaneous changes are very unlikely related to post-mortem damages. The following regions
594 were analyzed separately for the Admixture and Neighbor-Joining analyses (see below): *i*) *callable* regions,
595 defined as the whole genome excluding repeated regions; *ii*) *neutral* regions defined as callable regions
596 excluding genes (both exons and introns) and 10kb upstream and downstream each gene and *iii*) *exons*. The
597 estimation of Wattersons Theta (θ_w) and the FineStructure analysis were performed on *callable* regions only
598 (see below).
599
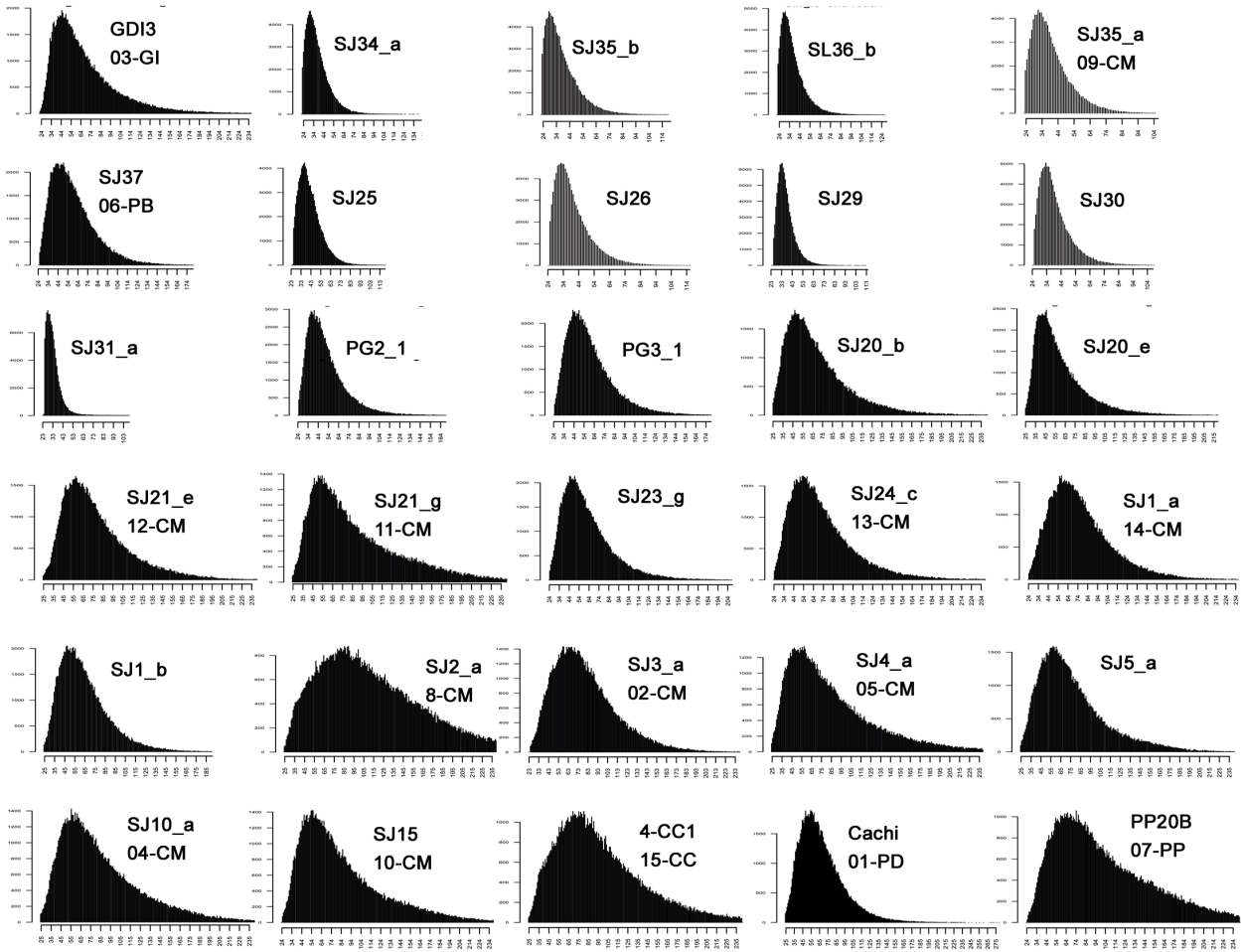
Figure S3. Fragment length (bp) distributions for the 30 ancient seeds. Labels present the original sample ID and the manuscript ID (see Table S1). Plots were generated using MapDamage 2.0.
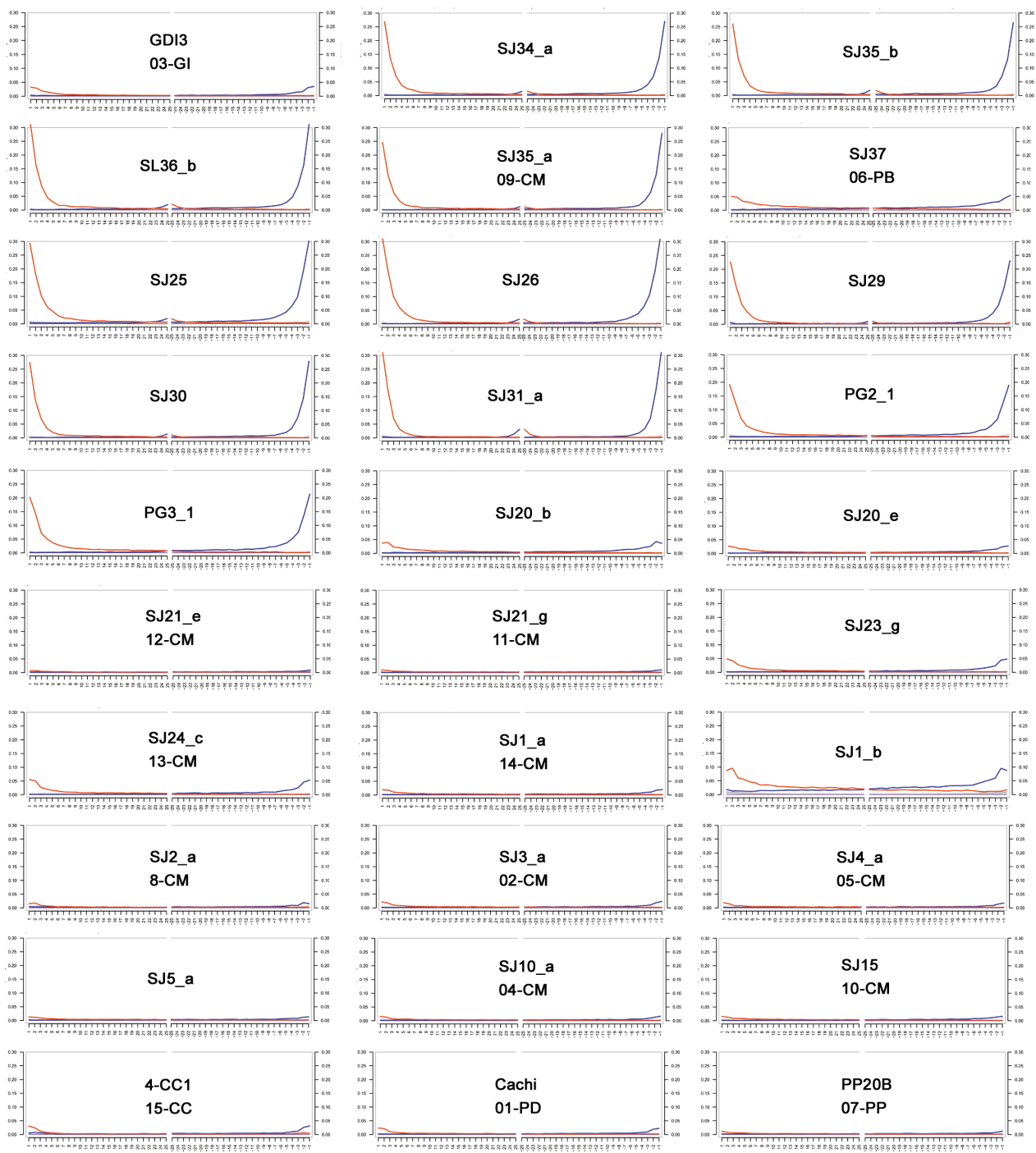
607
608
609 **Figure S5.** Frequency of C to T (red) and G to A (blu) misincorporation at 5' and 3' ends for the 30 ancient seeds. Plots were generated
610 using MapDamage 2.0. Labels present the original sample ID and the manuscript ID (see Table S1).
611
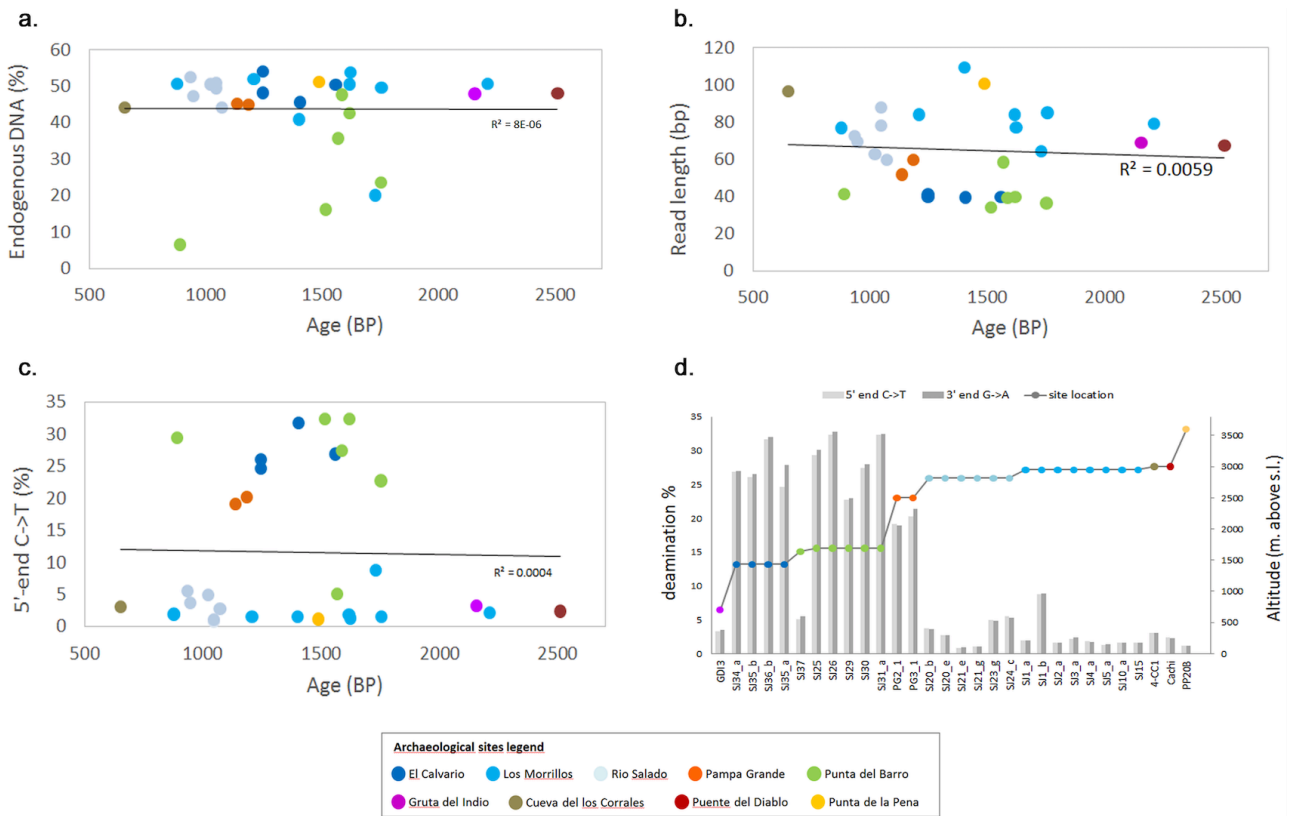
Figure S6. Endogenous DNA preservation in the 30 ancient bean seeds of the preliminary screening. a) Proportion of reads mapping to the common bean reference genome vs. age of the samples; b) Average read length vs. age of the samples; c) Misincorporation pattern at 5' end vs. age of the samples; d) Misincorporation pattern at 5' and 3' ends vs. site locations.

The fragments lengths and the fragmentation and deamination patterns of our data concur with what could be expected from these kinds of samples and we are confident about the authenticity of our ancient genomes. To further support the absence of contamination in our ancient DNA data, we estimated the heterozygosity level in the chloroplast calls. Contamination can introduce additional chloroplast haplotypes in the original library resulting in the presence of heterozygous sites. We performed a SNP calling analysis in chloroplast reads extracted from 14 ancient seeds (one seed was excluded because of low chloroplast coverage. The complete chloroplast sequence from the Negro Jamapa cultivar (accession: NC_009259, Guo et al, 2007) was added to the nuclear reference, and the bioinformatic approach described above was used to align the sequencing data from ancient samples to this new reference genome. Each resulting bam file was processed using ANGSD v0.916 (Korneliussen et al 2014) and the chloroplast consensus sequences were called for each seed setting the "-doFasta 2" option. The analysis was restricted to reads having a mapping quality score higher than 20 and to nucleotides showing a minimum base quality score of 20. The mean coverage ranges from 69X to 633X, and approximately 40% of the chloroplast genome was recovered in each sample (Table S2). This fraction is expected considering that several short chloroplast fragments (as those obtained from our samples), have multiple mapping position on the nuclear genome and were therefore excluded from further analyses. Each bam file was independently analysed with the mpileup/call commands from BCFtools, calling both monomorphic and polymorphic sites of the chloroplast, setting the ploidy of the region to 2. The following flags were added: a minimum base quality score of 20 (-Q 20), a minimum mapping quality score of 30 (-q 30) and the mapping quality of reads with two mismatches compared to the reference was adjusted (-C 50). Indels were not called. The resulting vcf files were filtered and only sites having a QUAL field > 30 were retained. The number of called heterozygous seeds was very small: 0 in 11 seeds, 1 in two seeds (13-CM and 03-GI), and 2 in one seed (02-CM). The total number of high quality/mapping base pairs used varied in different seeds between 45448 and 61699. The three ancient seeds with one or two heterozygous sites in the chloroplast genome do not show any significant pattern of increased nuclear heterozygosity when compared to other seeds.

Table S2. Chloroplast genomes coverage.

MANUSCRIPT ID	MEAN COVERAGE	FRACTION OF CALLED (NON-MISSING) BASE PAIRS
15-CC	167.7	0.42
01-PD	633.1	0.40
03-GI	151.4	0.37
07-PP	475.1	0.43
04_CM	290.6	0.38
10-CM	153.7	0.37
14-CM	102.4	0.35
12-CM	511.9	0.41
11-CM	752.4	0.43
13-CM	369.3	0.42
08-CM	450.3	0.41
06-PB	79.5	0.38
02-CM	627.7	0.42
05-CM	336.8	0.43

S6. Modern accessions genomic data

Genomic data from modern domesticated and wild common bean samples were included in our analyses (Table S3). In particular, we included: *i*) 12 modern domestic common bean genomes from both Andean and Mesoamerican domestication gene pools, sequenced by the ERA-CAPS funded project "BeanAdapt: The Genomics of Adaptation during Crop Expansion in Bean" (unpublished; provided by the coordinator of the project, Prof. Roberto Papa, Marche Polytechnic University); *ii*) two published genomes from wild *P. vulgaris* (G19901 from Argentina and G24594 from Chiapas, Mexico; Rendòn-Anaya et al 2017); *iii*) one published genome of the wild relative *P. hintonii* to be used as outgroup in our selection scan (Rendòn-Anaya et al 2017). Trimmomatic v0.36 (Bolger et al 2014) was used to remove poor-quality regions from downloaded Illumina reads, discarding nucleotides having an average base quality score less than 20 across a 5bp window. Reads longer than 35bp were mapped to the *P. vulgaris* reference genome v2.1 in the paired-end mode using bwa mem (Li and Durbin 2009) with the parameter "-M". The quality of bam alignments was improved removing PCR duplicates with Picard (software.broadinstitute.org) and performing the Indel realignment procedure, as described above for ancient samples. Depth of coverage was estimated in each sample using Samtools depth. Two publicly available but unpublished genomes from Argentinean wild seeds, G16798 and G18705 in the SRA archive and obtained from 2x150 bp libraries sequenced on an Illumina Hiseq 3000 platform, were used to confirm that the global pattern of variation of ancient seeds is compatible with the pattern of variation observed in modern seeds form the same geographic area. The mean coverage of these genomes was estimated at 9.9X and 8.8X, respectively.

Table S3. Modern genomes included in our analyses. Gene pool indicates to which gene pool the sample belongs; AD: Andean domesticated, MD: Mesoamerican domesticated, AW: Andean wild, MW: Mesoamerican wild. Sequencing mode was 2x125 bp for all Bean Adapt samples and 2x100 for the other samples.

Sample ID	Manuscript ID	Species	Wild/ Domestic	Gene pool	Country/Sampling site	Mean coverage (genome / callable regions)	Source
ECa005	CHI1	*P. vulgaris*	D	AD	Chile	11.6 / 13.0	Bean Adapt Project
ECa006	BOL1	*P. vulgaris*	D	AD	Bolivia	24.9 / 27.7	Bean Adapt Project
ECa016	CHI2	*P. vulgaris*	D	AD	Chile	9.1 / 9.8	Bean Adapt Project
ECa038	PER1	*P. vulgaris*	D	AD	Perú	8.9 / 9.9	Bean Adapt Project
ECa040	ARG1	*P. vulgaris*	D	AD	Argentina	16.6 / 19.5	Bean Adapt Project
ECa041	ARG2	*P. vulgaris*	D	AD	Argentina	6.1 / 6.7	Bean Adapt Project

ECa086	MEX1	*P. vulgaris*	D	MD	México	10.5 / 11.6	Bean Adapt Project
ECa095	MEX2	*P. vulgaris*	D	MD	México	6.2 / 6.7	Bean Adapt Project
ECa104	MEX3	*P. vulgaris*	D	MD	México	10.1 / 11.2	Bean Adapt Project
ECa332	ARG3	*P. vulgaris*	D	AD	Argentina	5.3 / 5.8	Bean Adapt Project
ECa330	ARG4	*P. vulgaris*	D	AD	Argentina	13.9 / 13.9	Bean Adapt Project
ECa333	BOL2	*P. vulgaris*	D	AD	Bolivia	7.9 / 8.9	Bean Adapt Project
G24594	MEX-W	*P. vulgaris*	W	MW	México	19.6 / 22.5	Rendòn-Anaya et al 2017
G19901	ARG-W	*P. vulgaris*	W	AW	Argentina	16.6 / 15.5	Rendòn-Anaya et al 2017
P. hintonii	-	*P. hintonii*	W	-	México	9.5 / 10.6	Rendòn-Anaya et al 2017

S7. Analysis of genomic diversity

To test whether there has been a decrease in genetic diversity due to the history of domestication, we compared levels of diversity between ancient and modern common bean seeds both as individual- and as population-level diversity. Genetic diversity was measured (using only transversions) estimating Watterson's Theta (θ_w) in non-overlapping windows of 10kb, taking into account per-individual inbreeding coefficients and hence accounting for possible deviations from Hardy-Weinberg equilibrium, as proposed by Vieira et al 2013. This is a two-step procedure: first, ANGSD v0.916 (Korneliussen et al 2014) was used to call a high confidence SNP set based on GATK genotype likelihoods (-GL 2) including only alignment positions having an associated base and mapping quality score equal or higher than 20 and 30, respectively. To increase the reliability of the called set, we set the baq computation (-baq 1), adjusted the mapping quality in case of excessive mismatches (-C 50), and only retained positions having a high probability to be polymorphic (-SNP_pvalue 1e-6), without any missing data. From this set, we randomly sampled 2,000 SNPs without replacement and we used ANGSD to recompute the genotype likelihoods at these positions (-doGlf 3) applying the same quality filters. The program ngsF (Vieira et al 2013) was then used to estimate the individual genome-wide inbreeding coefficient for each accession using default parameters, except for the –min_epsilon option that was decreased to 10^{-9}. The SNP sampling and the estimation of inbreeding processes were repeated five times in order to obtain five independent estimates of the inbreeding coefficient for each individual. The average across replicates was then used as point estimate of each individual inbreeding coefficient.

We then estimated a site frequency spectrum, using the command -doSaf 2 in ANGSD, to calculate the per-site posterior probabilities of the site allele frequencies based on individual genotype likelihoods accounting for the previously estimated inbreeding coefficients. Genotype likelihoods were calculated with the GATK method (–GL 2 option), and a folded spectrum was used (*i.e.*, unknown ancestral state, option –fold 1). Options -minQ 20 -minMapQ 30 -setMaxDepth 50 were applied to filter low quality bases and positions showing excessive coverage, while the option -sites was set to filter for *callable* regions. The option -noTrans 1 was applied to all the above steps to systematically remove transitions from the analysis. The output of this command line was then used as input for the calculation of per-site thetas, using the command -doThetas 1. Window-based theta estimates were calculated using the command thetaStat do_stat. Windows having a proportion of usable sites less than 50% were discarded.

Table S4. Estimates of diversity in ancient and modern accession of common bean from Central (Meso) and South (Andean) America. The number of 10kb-windows with at least 50% of the region covered is reported.

Manuscript ID	Gene pool	Country	10kb-windows	Θ_w mean	Θ_w median
CHI1	Andean	Chile	24486	1.29E-05	9.62E-08
BOL1	Andean	Bolivia	24686	1.56E-05	0
CHI2	Andean	Chile	24493	1.18E-05	1.73E-07
PER1	Andean	Perú	24597	9.62E-06	8.87E-08
ARG1	Andean	Argentina	24636	1.51E-05	1.18E-08

ARG2	Andean	Argentina	23827	7.44E-06	2.98E-07
MEX1	Meso	México	21749	4.66E-05	9.05E-07
MEX2	Meso	México	20071	2.42E-05	1.06E-06
MEX3	Meso	México	20965	3.86E-05	5.76E-07
ARG3	Andean	Argentina	23674	8.96E-06	1.40E-06
ARG4	Andean	Argentina	24272	1.67E-05	2.35E-06
BOL2	Andean	Bolivia	24491	1.99E-05	4.43E-07
15-CC	Andean	Argentina	15865	3.02E-05	1.72E-05
09-CM	Andean	Argentina	13284	3.78E-04	3.44E-04
01-PD	Andean	Argentina	20897	5.48E-05	5.41E-07
03-GI	Andean	Argentina	15241	1.79E-04	1.50E-04
07-PP	Andean	Argentina	15303	3.40E-05	1.03E-05
14-CM	Andean	Argentina	9372	1.12E-04	9.15E-05
04-CM	Andean	Argentina	9892	8.31E-05	6.10E-05
10-CM	Andean	Argentina	8422	1.80E-04	1.50E-04
08-CM	Andean	Argentina	13130	6.91E-05	4.27E-05
12-CM	Andean	Argentina	16859	1.15E-04	9.23E-05
11-CM	Andean	Argentina	18346	5.08E-05	1.25E-05
13-CM	Andean	Argentina	22600	8.76E-05	3.10E-05
02-CM	Andean	Argentina	21224	4.99E-05	8.47E-06
06-PB	Andean	Argentina	19426	5.51E-04	5.17E-04
05-CM	Andean	Argentina	17519	1.85E-04	1.58E-04
ARG-W	Andean wild	Argentina	24255	4.08E-05	1.93E-06
MEX-W	Meso wild	México	22583	1.51E-03	9.25E-04
G16798 (unpublished)	Andean wild	Argentina	23982	5,38E-05	1,91E-05
G18705 (unpublished)	Andean wild	Argentina	23960	5,33E-05	2,00E-05

712

713 The genomic variation in ancient beans is compatible with the variation observed in modern wild Argentinean
714 seeds (see also Figure 1b in the main text), larger than that observed in modern domestic seeds with Andean
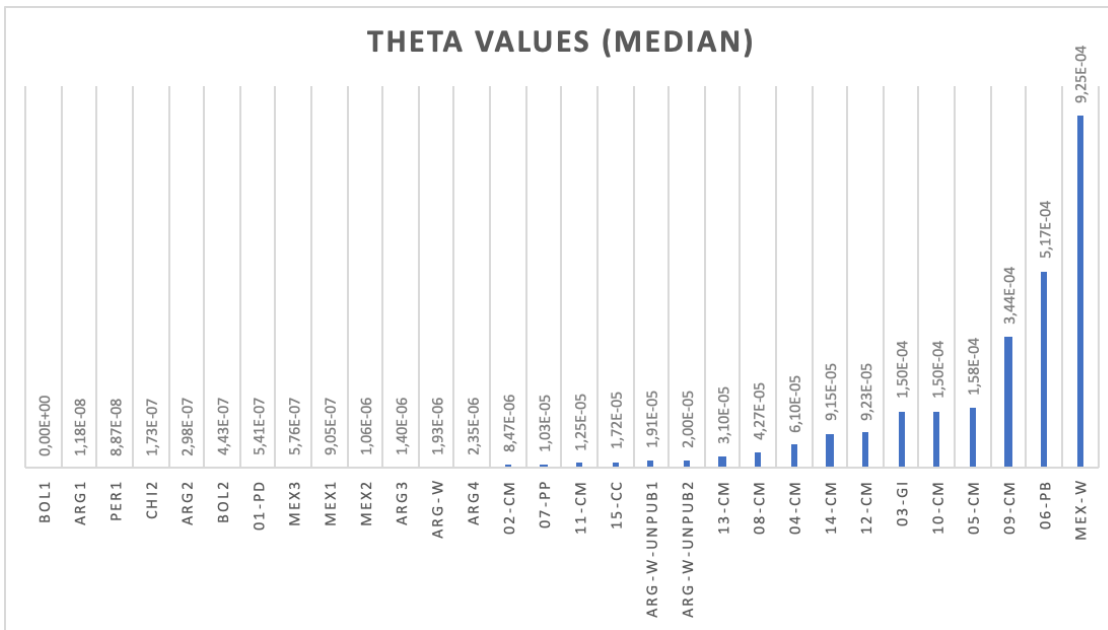715 gene pool, and smaller than that found in the Meso-American wild seed (Figure S7).

716

717
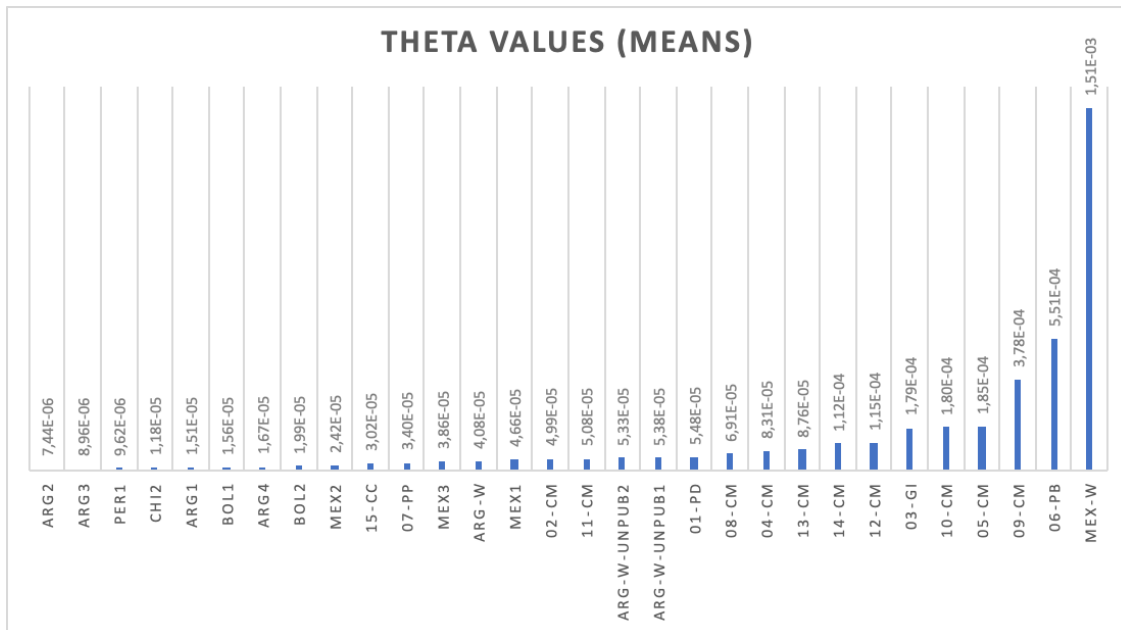
718
719

Figure S7. Mean (upper panel) and median (lower panel) values of Theta estimates for modern and ancient seeds, in ascending order. Seed codes as in Table S4.

The impact of coverage (in callable regions) on heterozygosity estimates cannot be excluded. However, we note that a) the range of coverage in these regions is very similar in the ancient (4.2X – 23.2X; average: 9,0X) as compared to modern (5.8X – 27.7X, average: 12.7X) genomes; b) in the modern genomes data set, excluding the wild Mexican sample (typed at high coverage, and where genomic variation is expected to be much higher than the Andean samples for biological reasons), the increase of coverage from the smaller value (which is >5) does not seems to affect genetic variation in different cultivars (Figure S8, left panel); b) in the ancient genomes data set, 6 seeds with coverage>10 have, on average, a heterozygosity that is one third smaller than the remaining seed at coverage <10 (Figure S8, right panel). There might be therefore a small bias, probably due to the overestimation of theta by the probabilistic ANGSD approach when the coverage is low. Nevertheless, using only the 6 ancient seeds at coverage >10, or only the 10 ancient seeds at coverage >5, the average theta results equal to 1.10E-4 and 1.09E-4, respectively, which is only 23% smaller than the global (15 seeds) average reported in the main text. Most importantly, this value of theta does not change the main pattern identified in our study: 1.1E-4 (or 1.09E-4) is still 8.5 times larger than the average variation in the modern Andean cultivars (1.3E-5), and these differences are highly significant (Mann-Whitney, U<0.001).
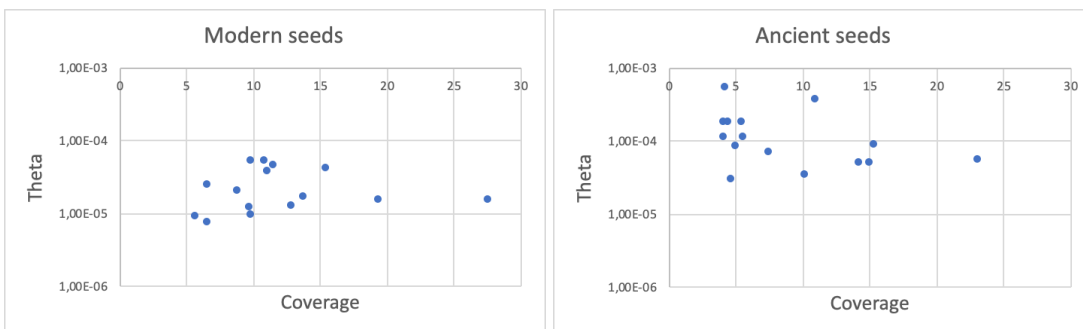
Figure S8. Plots of Theta estimated in modern (left panel) and ancient (right panel) seeds, as a function of the coverage in callable regions.

S8. Re-calibrating estimates of modern genomic diversity

748 Genomic diversity of each modern accession was likely reduced by the protocols commonly used before
749 whole-genome sequencing, which included a selfing protocol after collecting the seeds from the seed bank.
750 In addition, previous maintenance practices applied at each seed bank could have also affected genome-
751 wide diversity. To test whether such very recent events could be the cause of the observed pattern of lower
752 diversity in modern seeds as compared to ancient seeds, we re-estimated whole-genome diversity in pairs
753 of seeds from closely related cultivars. The rationale behind this test is the following. Diversity loss caused
754 by recent seed bank or sequencing protocols occurred at random along the genome in each cultivar, so that
755 each cultivar should have lost a different fraction of its diversity. Taking together two (or more) closely
756 related cultivars should restore most of the variation prior to the stocking into the seed banks. Of course,
757 this estimate is also affected by the possible genetic divergence between the cultivars used in each pair.
758
759 Estimates of diversity (θ_w) in pairs of cultivars should better represent the single cultivar variation before
760 seed bank and sequencing protocols (or even over-estimate, due to cultivar genetic divergence). We
761 selected the two modern Chilean cultivars, already included in all our analyses (see Table S3) and added
762 three cultivars from the same country sequenced within the ERA-CAPS funded project "BeanAdapt: The
763 Genomics of Adaptation during Crop Expansion in Bean" (unpublished; provided by the coordinator of the
764 project, Prof. Roberto Papa, Marche Polytechnic University). We estimated ϑ_w in ten random pairs of seeds.
765 Specifically, we compared average θ_w estimated in three groups of accessions: 10 pairs of Chilean
766 accessions, 12 single modern accessions from South America, and in 15 ancient seeds. Alongside individual
767 variation, we also compared the distribution of θ_w across all 10kb-windows from each group.
768
769 As expected, genome wide diversity in pairs of seeds is higher than in single seeds (Table S5). However,
770 average modern seed pairs diversity is still more than three times lower (Mann-Whitney U test, $P = 0.003$)
771 than that estimated in ancient seeds (Table S4). The same pattern appears when considering the
772 summarized diversity across all 10kb-windows in each group (Figure S5).
773
774
775
776 **Table S5.** Estimates of diversity in five accessions from Chile (two of them were also used in the main analyses) and in the ten pairs of
777 Chilean samples. The number of 10kb-windows with at least 50% of the region covered is reported.
778

Manuscript ID	10kb-windows	Θ_w mean	Θ_w median
CHI1 (ECa005)	24486	1.29E-05	0.96E-07
CHI2 (ECa016)	24493	1.18E-05	1.73E-07
CHI3 (ECa004)	24524	1.37E-05	0.62E-07
CHI4 (ECa018)	24609	1.57E-05	0.33E-07
CHI5 (ECa020)	24541	1.47E-05	0.94-07
CHI3.CHI1	27870	4.33E-05	2.89E-07
CHI3.CHI2	27875	4.24E-05	4.43E-07
CHI3.CHI4	27912	3.96E-05	1.95E-07
CHI3.CHI5	27891	4.63E-05	3.46E-07
CHI1.CHI2	27856	4.47E-05	4.89E-07
CHI1.CHI4	27903	5.32E-05	3.14E-07
CHI1.CHI5	27889	4.32E-05	3.69E-07
CHI2.CHI4	27916	4.57E-05	4.07E-07
CHI2.CHI5	27888	3.59E-05	4.49E-07
CHI4.CHI5	27930	5.61E-05	4.56E-07

779
780
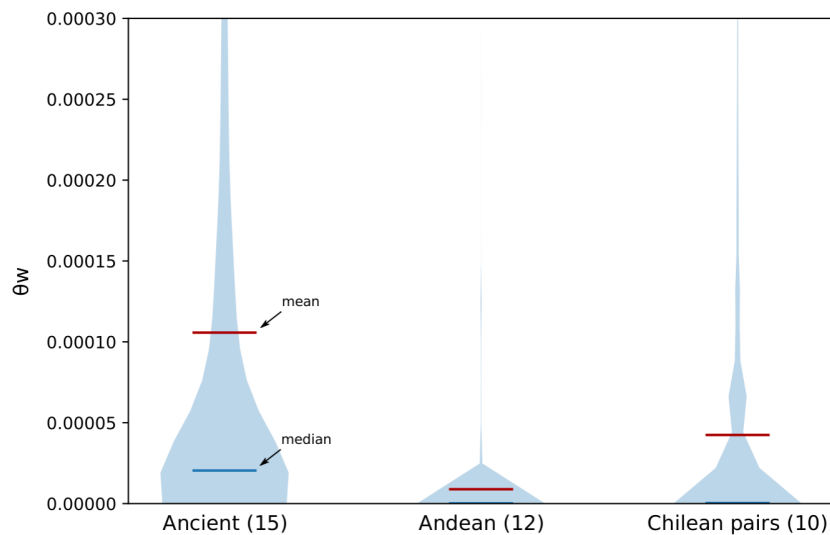
In order to exclude major effects of selfing before sequencing, we also applied a conservative procedure and increased the heterozygosity values of modern seeds as a function of the number of selfing cycles. We note that our modern cultivar samples are a subsample of the cultivars sequenced during the ERA-CAPS project "BeanAdapt" (see Data accessibility in the main text), used here for comparison. Notably, not all the modern seeds followed in that project the same selfing protocol: some of them were sequenced directly after the collection at the CIAT seed bank, where one cycle of selfing was performed, whereas the other seeds were subjected to two cycles of selfing. More specifically, some modern Andean seeds in our sample (ARG3, ARG4, and BOL2) were subjected to one cycle of selfing, and others (BOL1, CHI1, CHI2, PER1, ARG1, and ARG2) were subjected to two cycles of selfing. Two additional Andean seeds sequenced for the BeanAdapt project (two cultivars from Argentina) were selfed only once, and we included them in this analysis (ARG5 and ARG6). Considering only modern seeds with Andean ancestry, we have therefore 15 ancient seeds (no selfing before sequencing) and 11 modern seeds (5 and 6 with one and two cycles of selfing, respectively).

Four out of five modern seeds subjected to one cycle of selfing have a mean θ_w higher than all the modern seeds subjected to two cycles of selfing, and five out five modern seeds subjected to one cycle of selfing have a median θ_w higher than all the modern seeds subjected to two cycles of selfing. This is consistent with the expectation that selfing is reducing heterozygosity as a function of selfing cycles.

The original θ_w values for the modern seeds were then doubled or quadrupled (depending on the number of selfing cycles) to conservatively (some level of outcrossing during the procedure cannot be excluded) account for the reduction of variation before sequencing, allowing for a more robust comparison between ancient and modern seeds. Considering the means, 10 out of 15 ancient seeds have a θ_w value larger than the 11 modern seeds. This produces a significant difference between θ_w values (p<0.01, Mann-Whitney U test). Considering the medians, 14 out of 15 ancient seeds have a θ_w value larger than the 11 modern seeds. This produces a significant difference between θ_w values (p<0.01, Mann-Whitney U test).

S9. Structure of the genomic diversity in ancient beans

A haploid calling on all samples (15 ancients, 14 modern) was performed on each chromosome, separately, using ANGSD 0.916 (Korneliussen et al 2014). We set -minQ 20 and -minMapQ 30 to exclude low quality bases and low mapping quality bases; -only_proper_pairs 0 to use all reads (not only reads with both mapping pairs); -dohaplocall 2 to select the most frequent base in the haploid calling (*i.e.*, not a random one); -doCounts 1 to count the bases at each sites after filters; -minMinor 2 to exclude singletons and -

820 noTrans 1 to exclude transitions. The output of the haploid calling was then masked according to the three
821 genomic regions as defined above (*callable*, *neutral* and *exons*) using *Bedtools intersect* (Quinlan and Hall
822 2010). The fraction of missing calls (after haploid calling) and the total number on variable sites is reported
823 in Table S6. Higher values in ancient compared to modern genomes can be attributed to the shorted read
824 length.
825
826 TableS6. Proportion of missing calls after haploid calling in neutral, exons and callable regions. The total number of variable
827 positions across samples is indicated within brackets.
828

| Manuscript ID	Neutral (500,580)	Exons (335,124)	Callable (3,517,565)
15-CC	0.29	0.09	0.29
09-CM	0.41	0.13	0.33
03-GI	0.30	0.10	0.30
10-CM	0.45	0.13	0.45
02-CM	0.17	0.07	0.15
05-CM	0.27	0.10	0.26
04-CM	0.42	0.11	0.41
14-CM	0.41	0.14	0.41
08-CM	0.33	0.09	0.34
01-PD	0.14	0.06	0.12
07-PP	0.29	0.07	0.30
06-PB	0.25	0.11	0.21
12-CM	0.29	0.09	0.27
11-CM	0.24	0.06	0.23
13-CM	0.13	0.07	0.11
CHI1	0.04	0.03	0.03
BOL1	0.03	0.02	0.02
CHI2	0.05	0.04	0.03
PER1	0.04	0.03	0.02
ARG1	0.04	0.03	0.02
ARG2	0.06	0.04	0.04
MEX1	0.09	0.04	0.05
MEX2	0.11	0.06	0.07
MEX3	0.09	0.03	0.05
ARG4	0.04	0.02	0.02
ARG3	0.07	0.06	0.05
BOL2	0.05	0.04	0.03
ARG-W	0.04	0.03	0.03
MEX-W	0.07	0.03	0.04

829
830
831 The main components of the genetic diversity and the eventual admixture proportion in all samples were
832 inferred by *Admixture* analyses (Alexander and Novembre 2009; Supplementary Figure S10). This analysis
833 was performed using the haploid data after combining all chromosomes in one file. As *Admixture* input is
834 binary *Plink* (.bed) or ordinary *Plink* (.ped), the haploid data was converted in tped format (*Plink* transposed
835 text genotype table), using the haploToPlink function implemented in *ANGSD*. Non biallelic sites were
836 excluded. This file was then converted to bed format, using –make-bed option in plink-1.90 (Chang et al
837 2015). The option –bp-space 50000 was added in order to select one SNP every 50kb (thinning) and
838 minimize the effect of linkage disequilibrium in the *Admixture* analysis. We tested *k* values from 1 to 15, and
839 cross validation error (Alexander and Lange 2011) was calculated for each *k* value, adding --cv flag to the
840 command line. Cross validation plots and admixture plots (for each K value) were created in *R*
841 *(*Supplementary Figure S10).
842
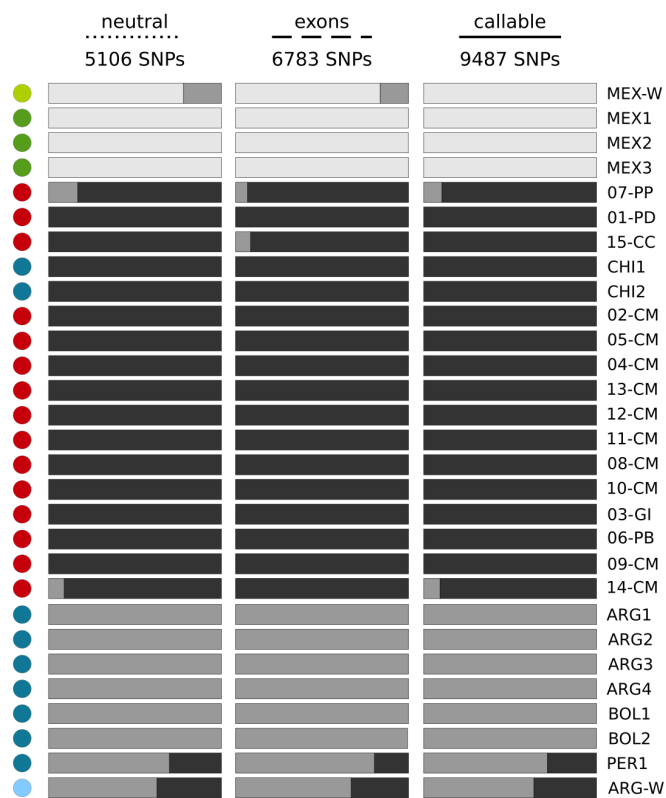
843
844 **Figure S10.** Structure of the genetic diversity and cross-validation error inferred using three different genomic regions (*callable:* solid
845 line*, exons:* large dashed line*, neutral:* small dashed line). For all genomic regions *k* = 3 represented the nest fitting model. Ancient
846 domesticated Andean (red), modern domesticated Andean (blue) and Mesoamerican (green), wild Andean (pale blue) and
847 Mesoamerican (pale green).

848
849 The output of the haploid calling masked according to the three genomic regions was also used to calculate
850 per chromosome pairwise genetic distances across all modern and ancient individuals using a custom *awk*
851 script (available at https://github.com/anbena/ancient-beans). Per-chromosome distances were then
852 combined into a single genome-wide distance matrix, which was used to reconstruct a Neighbor-Joining
853 tree using the *ape* package in *R* (Paradis et al 2004)*.* Trees were visualized with FigTree
854 (http://tree.bio.ed.ac.uk/software/figtree). As in the Admixture analysis, the three genomic regions
855 produced almost identical results (the Neighbor-Joining tree based on callable region only is presented in
856 Figure 1d in the main text).

857
858 Population structure was also inferred using the method implemented in FineStructure (Lawson et al 2012;
859 Figure S11). The pattern of similarity and dissimilarity between haploidized individuals can be visualized by
860 a distance matrix, and discrete populations can be separated by a cladogram. This method is able to capture
861 at least as much information as that captured by other exploratory analyses such as PCA and Structure, but
862 it also has the advantage to help identifying the pattern of population structure and to combine the
863 information at linked markers. The latter advantage is of course reduced, or absent, in our haploidized data
864 set. We used Plink to filter the dataset in order to exclude any variable position with missing data (not
865 allowed in FineStructure) and the plink2chromopainter.pl conversion script to prepare the input file to run
866 the FineStructure pipeline. In the first step of the pipeline, haplotype blocks are "painted" (Li and Stephens
867 2003) according to the nearest (in terms of genetic distance) individual in the sample (chromosome painting
868 step). Painting results are then employed to build a distance matrix summarizing the blocks similarity
869 between all pairs of individuals. The distance matrix is, in turn, the basis of a Bayesian clustering approach
870 aiming at partitioning the sample into *K* homogeneous groups (Supplementary Figure S11).
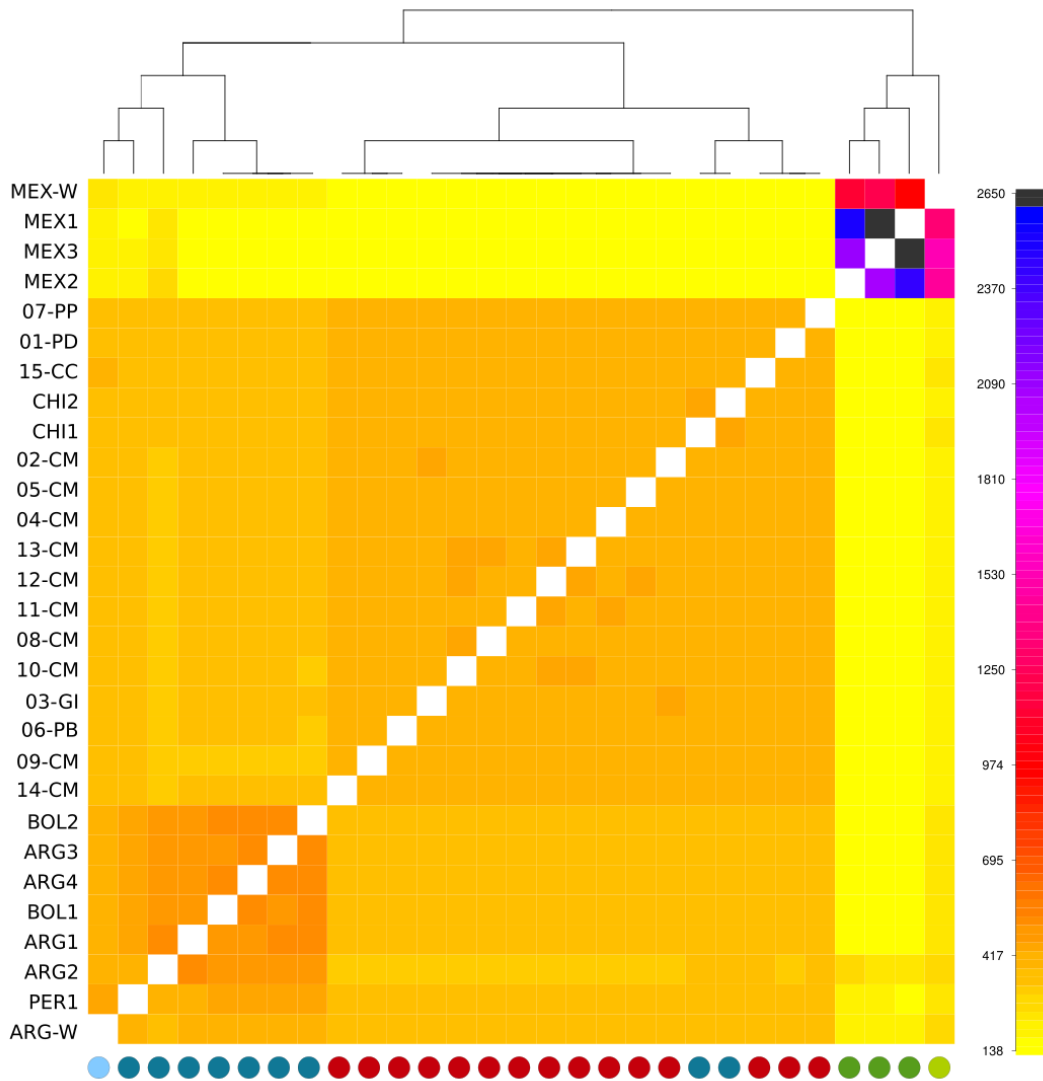
S10. Loss and recovery of variation after domestication?

Here we briefly support the view that the heterozygosity in the ancient seeds is comparable to that observed in the wild seeds because of a smaller or negligible drift effect during the initial phases of domestication, and not as a consequence of a strong bottleneck during domestication and subsequent recovery by mutation. Let's assume an individual heterozygosity of 1/10000 in an ancient seed at 2500 BP (the order of magnitude we observe). If a severe bottleneck occurred around the initial phases of domestication (say 8000 BP), followed by expansion, the coalescence tree of two homologous bases in the ancient seed would likely coalesce at 8000 BP, i.e, 5500 year earlier. Then, if we equate individual heterozygosity with the expected difference between the two bases (10^{-4}), the transversion rate per base pair per year should take a value of almost 10^{-8}/base/year ($10^{-4}/(2 \times 5500) = 0.91 \times 10^{-8}$). This rate is more than one order of magnitude larger than what can be estimated by, for example, comparing *P. vulgaris* with *P. hintonii* (separated 5MYA, Delgado-Salinas et al., 2006), which we estimated at 0.72×10^{-9}/base/year using only transversions in a pairwise comparison between *P. hintonii* and the Mexican reference genome for *P. vulgaris*). In other words, it seems very unlikely that if a bottleneck occurred, a significant fraction of the variation was recovered after 5500 years, unless a much higher transversion rate than currently estimated is assumed.

893
894
895
896
897 **S11. Gene-focused scan for signature of ancient and/or recent selection**
898
899 A gene-by-gene selection scan was performed to identify genes targeted by artificial selection during early
900 and/or late stages of domestication. A gene underlying a desirable trait is expected to have undergone one
901 or more selective sweeps targeting specific alleles coupled with recurrent hitchhiking at linked neutral sites.
902 Selected genes will then accumulate variants which are either directly advantageous or got fixed because of
903 locally low effective population size due to linked selection (Slotte 2014, Renaut and Rieseberg 2015,
904 Beissinger et al 2016). We then expect to find an excess of fixed differences at genes under selection when
905 samples of pre (*i.e.*, wild) and post domestication populations are compared.
906
907 Leveraging our ancient data, we tested each of the 27,000 genes for a significant enrichment of fixed
908 alternative alleles in two possible configurations. The first configuration is AS (Ancient Selection), indicated
909 by the topology in the left panel in Figure 2A (main text), where all the ancient + modern Chilean samples
910 have the same allele, and the wild samples (one wild Andean *P. vulgaris* and one *P. hintonii* as wild
911 outgroup) have the alternative allele. We expect AS to be frequent within genes that experienced changes
912 before 2,500 years ago. The second configuration is RS (Recent Selection), indicated by the topology in the
913 right panel in Figure 2B (main text), where all modern Chilean samples have the same allele, and all the
914 other samples (ancient + wild samples) have the alternative alleles. We expect RS to be frequent within
915 genes that experienced changes in last 600 years, i.e., more recently than the estimated age interval for the
916 ancient seeds.
917
918 Clearly, the enrichment of AS or RS allelic configurations (corresponding to the different topologies based
919 on fixed differences) depends on many factors beside the selection dynamic, such as the sample sizes and
920 the drift intensity in different time intervals. We therefore used a randomization procedure that keeps
921 constant the groups, their sample sizes, and the total number of SNPs, but randomly assigns the SNPs to the
922 genes. The relative frequency of SNPs with fixed differences in AS and RS topologies, computed after this
923 randomization, becomes our null hypothesis, which is used to test if the observed frequencies of such
924 topologies in each gene are higher than expected by chance.
925
926 In particular, we applied the following protocol: 1) for 27,000 genomic region each identified by a gene and
927 10 kb before and after it, we counted the proportion of fixed differences in the AS and the RS configuration;
928 2) we randomly re-assigned each SNP found in the real dataset (including the modern seeds from Chile, the
929 ancient seeds from Argentina, the wild seed and the outgroup) to the 27,000 regions; 3) we estimated from
930 different random assignment runs the proportion of AS and RS configurations expected by chance in a
931 "average" genomic region; 4) we used the proportion estimated in point 3 as the null hypothesis, and
932 applied a binomial test to each proportion observed in the real genes (point 1) to compute a *p*-value of
933 significant enrichment of that particular configuration; 5) we converted the *p*-values in *q*-values (False
934 Discovery Rate, FDR) using the *python* function qvalue.py
935 (https://github.com/nfusi/qvalue/blob/master/qvalue/qvalue.py) based on (Storey and Tibshirani, 2003); 6)
936 only genomic regions with a FDR<0.001 were initially considered as genes significantly enriched in a
937 particular topology (see Supplementary Figure S12); 7) finally, we applied an additional conservative
938 criterion to filter each list of AS and RS enriched genes, and retained only those genes with numbers of fixed
939 differences within the coding region and a 1kb region upstream and downstream the gene higher than a
940 threshold defined by the 99-percentile of all the 27,000 genes.
941
942 The procedure we applied is highly conservative, and we likely missed genes where the selection process
943 may have modified the allelic compositions. It allows however to significantly reduce the rate of false
944 positives, and to correctly estimate the relative proportion of genes with relevant modifications before and
945 after the time interval defined by the ages of the seeds we analyzed, respectively.
946

947 We found 360 and 86 genes with a significant enrichment of SNPs in the AS and RS topology, respectively,
948 with a AS/RS ratio = 4.2. We also estimated the ratio of ancient vs. modern sweeps considering the
949 hypothesis that different genes on the same linkage block may produce a signal of selection without being
950 effectively selected. Partially following the approach used by Schmutz et al (2014), we merged all AS or RS
951 genes in our study if they were separated by less than 40kb, and we considered these extended regions as
952 anciently selected regions (ASR, i.e., ancient sweeps) or recently selected regions (RSR, i.e., recent sweeps).
953 The number of ASR and RSR is of course smaller than AS and RS, but the ratio between ancient and recent
954 sweeps is very similar:161 vs. 46 for ASR and RSR, respectively, with a ratio of 3.5. If the maximum distance
955 allowed to merge genes was increased from 40kb to up to 150kp, the ratio ASR/RSR remains larger than 3.1.
956
957 Considering the ancient seeds as a single population, we estimated Tajima's D (Tajima, 1989) along the
958 genome and compared it with the position of ancient selection signals (AS) identified by our previous test.
959 Using the same pipeline applied in section S7 to estimate genomic diversity in groups of seeds (i.e,
960 command thetaStat do_stat in ANGSD 0.916), we computed Tajima's D in non-overlapping 10kb genomic
961 windows. This is a one-sample test (i.e., it is not based on the divergence among groups), and negative
962 values from this test can be therefore considered as additional evidence of selection which is independent
963 from the results of the ad hoc test we developed. Using a threshold of -1.25 to identify Tajima's D values
964 supporting a sweep (this threshold was chosen so to have a comparable numbers of candidate
965 windows/genes in both tests), we found that 66 out of 360 (18.3%) genes (including 50Kb before and after
966 the gene) show both a signature of ancient selection in our test and a Tajima's D score smaller than -1.25.
967 As a comparison, 73 out of 678 (10.8%) Andean domestication genes identified in Schmutz et al (2014)
968 show also a Tajima's D score smaller than -1.25. When blocks around AS genes are defined including 50 kb
969 upstream and downstream of the gene, and the same is done for the genes identified in Schmutz et al
970 (2014), 25 blocks are shared among studies and also overlap with the windows with reduced Tajima's D
971 (<1.25).
972 The chromosomal position of the selective sweeps that may result from domestication found in our data
973 was visually compared in Supplementary Figure S13 with those inferred in Schmutz et al, (2014) and with
974 the Tajima's D candidate windows Several chromosomal regions (e.g, beginning of chromosome 1; between
975 46 and 48 Mb in chromosome 2; between 36 and 38 Mb in chromosome 7; end and beginning of
976 chromosome 8) show co-localization of the selection signals, pointing to regions that deserve specific
977 attention in further studies. The ASR/RSR ratio considering only the blocks that overlap with the sweeps
978 inferred in Schmutz et al, (2014) is 12.5 (25/2) and 5.9 (41/7) applying a strict overlap criterion or extending
979 the blocks to 50kb upstream and downstream, respectively.
980 The two final lists of genes showing signature of AS and/or RS were grouped into functional clusters using
981 the web-based software David (Huang et al 2009). For this analysis, we could only use genes with a known
982 ortholog in *A. thaliana* (317 and 72 for the ancient and the recent selection targets, respectively, excluding
983 multiple matches with the same *A. thaliana* orthologs). We used the functional enrichment score estimated
984 by David only to rank the functional clusters found in the two lists of genes without excluding any cluster.
985 Analyses were also replicated using the genome of a wild Mesoamerican *P. vulgaris* as outgroup instead of
986 *P. hintonii* (Table S7), with 65% and 68% of the genes in the first set (with *P. hintonii* as outgroup) found as
987 AS and RS in the second set (using Mesoamerican *P. vulgaris* as outgroup). In this case, the ratios AS/RS and
988 ASR/RSR reduced to 2.3 and 2.1 respectively, still indicating that at least twice as many selection events
989 occurred before the time period covered by the ancient seeds. The reduction of this ratio is almost entirely
990 due to an increase of genes classified as RS (147) or RSR (75), whereas AS genes and ASR regions remain
991 approximately constant (333 and 160, respectively). If RS (and RSR) topologies with the *P. hintonii* outgroup
992 are likely related to real changes occurring in genes in recent centuries (because modern seeds are
993 genetically differentiated from genes conserved in wild and ancient beans, which are still similar to the
994 distantly related outgroup *P. hintonii*), RS (and RSR) topologies may increase when a wild Mesoamerican *P.*
995 *vulgaris* is used as outgroup because of incomplete lineage sorting within the *P. vulgaris* species. When the
996 wild Mesoamerican *P.vulgaris* is used as outgroup, the ASR/RSR ratio for the blocks that overlap with
997 sweeps inferred in Schmutz et al, (2014) becomes 6.5 (26/4) and 4.1 (45/11) applying a strict overlap
998 criterion or extending the blocks to 50kb upstream and downstream, respectively.
999
1000

1001
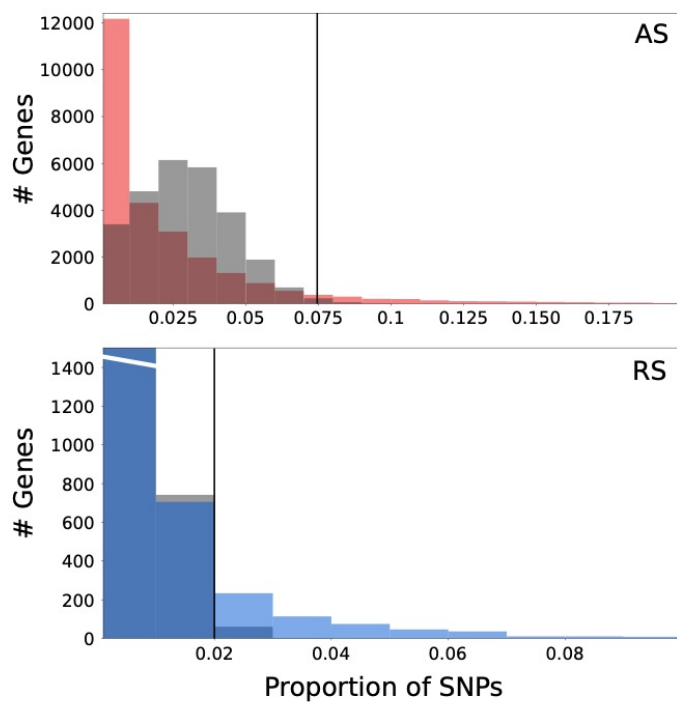
1002
1003
1004 **Figure S12.** Observed (in color) and null (gray) distribution of proportion of SNPs per gene in AS (red, top panel) or RS (blue, bottom
1005 panel) topology used as summary statistics in the selection scan. The lowest fraction of AS or RS topologies in the selected genes
1006 (after using the FDR threshold of q<0.001, and before the implementation of the additional conservative filter based on the
1007 absolute number of SNPs, see above) is indicated as a solid line in each panel.

1008
1009
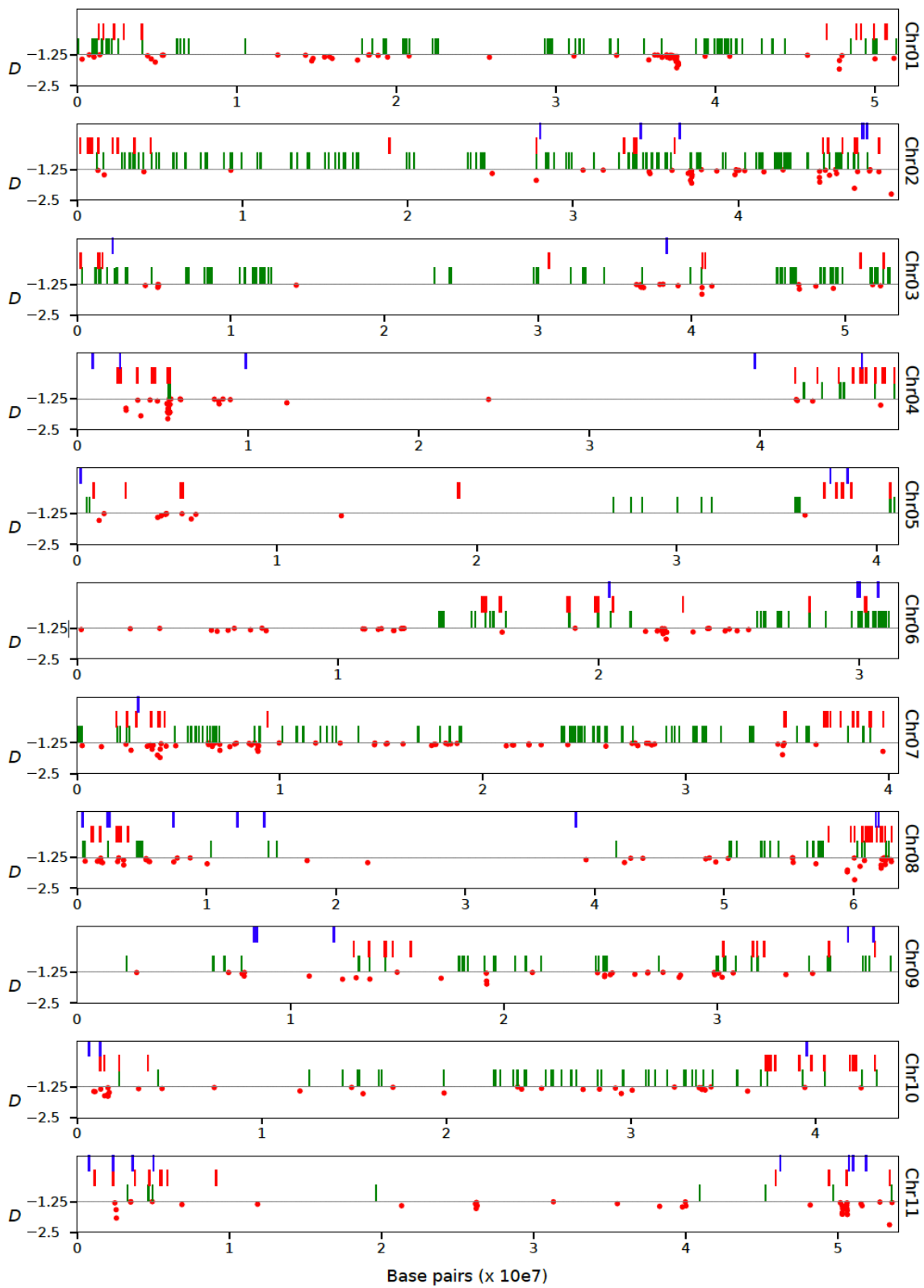
Figure S13. Candidate genes as selected during ancient (AS, red) and recent (RS, blue) domestication stages identified by our test, and genes suggested as candidates for selection during domestication in the Andes (Schmutz et al 2014, dark green). Genomic windows with a Tajima's D smaller than -1.25 are also shown (red dots)

S12. References

Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinform.* **12**, 246 (2011).

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

Babot, P., Aguirre, M.G. & Hocsman, S. Aportes del sitio Punta de la Peña 9 (Puna de Catamarca) acerca de la producción y usos prehispánicos de quinua. In: *Libro de Resùmenes del Simposio Internacional de la Quinua*. Universidad Nacional de Jujuy, Instituto Nacional de Tecnología Agropecuaria, Centro de Investigaciones y Transferencia de Jujuy , 65-66, (CONICET, 2013).

Babot, P., et al.Ocupaciones agropastoriles en los sectores intermedios de Antofagasta de la Sierra (Catamarca): un análisis desde Punta de la Peña 9.I. *Comechingonia* **9**,57-78 (2006).

Babot, P. La cocina, el taller y el ritual: explorando las trayectorias del procesamiento vegetal en el Noroeste argentino. Darwiniana 47, 7-30 (2009).

Baldini, M., Baffi, E., Salaberry, M.&Torres, M. Candelaria: una aproximación desde un conjunto de sitios localizados entre los cerros de Las Pirguas y El Alto del Rodeo (Dto. Guachitas, Salta, Argentina). In: La mitad verde del mundo andino. Investigaciones arqueológicas en la vertiente oriental de los Andes y las Tierras Bajas de Bolivia y Argentina (eds. Ortiz, G. & Ventura, B.)131-151 (Universidad Nacional de Jujuy,, 2003)

Beissinger, T. M., et al. Recent demography drives changes in linked selection across the maize genome. *Nat. Plants* 2, 16084 (2016).

Bolger, A. M., Lohse, M., & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

Chang, C. C., et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

Delgado-Salinas, A., Bibler, R., & Lavin, M. Phylogeny of the Genus Phaseolus (Leguminosae): A Recent Diversification in an Ancient Landscape. *Systematic Botany* **31**, 779-791 (2006).

Gambier, M. *La cultura de Ansilta* (San Juan, Instituto de Investigaciones Arqueológicas y Museo UNSJ, 1977).

Gambier, M. *La fase cultural Punta del Barro* (Instituto de Investigaciones Arqueológicas y Museo, Facultad de Filosofía, Humanidades y Artes, Universidad Nacional de San Juan, 1988)

Gambier, M. *Prehistoria de San Juan*. (Ansilta, 2000).

Gil, A. F., Neme, G.A & Tykot, R. H. Isótopos estables y consumo de maíz en el Centro Occidente Argentino: Tendencias temporales y espaciales. *Chungara* **42**, 497–513 (2020).

Gil, A. F., Giardina, M. A., Neme G. A. & and Ugan A. Demografía humana e incorporación de cultígenos en el centro occidente argentino: Explorando tendencias en las fechas radiocarbónicas. *Rev. Esp. de Antropol. Am*. **44**,523–553 (2014).

Guo, X., et al. Rapid evolutionary change of common bean (*Phaseolus vulgaris L*) plastome, and the genomic diversification of legume chloroplasts. *BMC Genomics* **8**, 228 (2007)

Hofreiter, M., Collins, M., & Stewart, J. R. Ancient biomolecules in Quaternary palaeoecology. *Quat. Sci. Rev.* **33**, 1-13 (2012).

Huang, D.W., Sherman, B.T., Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat. Protoc*. **4**, 44-57 (2009).

1059 Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L., & Orlando, L. mapDamage2. 0: fast approximate
1060 Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682-1684 (2013).

1061 Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC*
1062 *Bioinform.***15**, 356 (2014).

1063 Lagiglia, H. Nuevos fechados radiocarbónicos para los agricultores incipientes del Atuel. In: *Actas del XII*
1064 *Congreso Nacional de Arqueología Argentina*(ed.Marín, C. D.), **3**,239-250 (Universidad Nacional de La
1065 Plata,1999).

1066 Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. Inference of population structure using dense haplotype
1067 data. *PLoS Genet*. 8, e1002453 (2012).

1068 Lema, V. Domesticación vegetal y grados de dependencia ser humano-planta en el desarrollo cultural pre
1069 hispánico del Noroeste argentino. PhD Thesis at Universidad Nacional de La Plata, Argentina, 2009.

1070 Lema, V. Confluencia y emergencia: domesticación y prácticas de manejo del entorno vegetal en la
1071 frontera. In: *Actas del XVII Congreso Nacional de Arqueología Argentina –Arqueología Argentina en el*
1072 *Centenario de la Revolución de Mayo* (eds. Bárcena, R. & Chiavazza, H.) , **3**, 1043-1048 (2010).

1073 Lema, V. Non domestication cultivation in the Andes: plant management and nurturing in the Argentinean
1074 Northwest. *Veg. Hist. Archaeobot.* **24**, 143-150 (2015).

1075 Lema, V. Geografías y prácticas: plantas que circulan, que se quedan y que se van para no volver. In
1076 *Arqueología de la vertiente oriental surandina: interacción macro-regional, materialidades, economía y*
1077 *ritualidad* (eds.Ventura, B.,Ortiz G. & Cremonte, B.) 267-276 (Sociedad Argentina de Antropología, Buenos
1078 Aires, 2017).

1079 Li, H., & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*
1080 **25**, 1754-1760 (2009).

1081 Li, H., et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009)..

1082 Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-
1083 nucleotide polymorphism data. *Genetics* **165**, 2213-2233 (2003).

1084 Lindgreen, S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res. Notes* **5**, 337
1085 (2012).

1086 Long, A., Martin, P. & Lagiglia, H. Ground sloth extintion and human occupation at Gruta del Indio,
1087 Argentina. *Radiocarbon* **40**, 693-700 (1988).

1088 López Campeny, S.M.L., Romano, A.S. & Aschero, C.A.   Remodelando el Formativo. Aportes para una
1089 discusión de los procesos locales en las comunidades agropastoriles tempranas de Antofagasta de la Sierra
1090 (Catamarca, Argentina). In: *Crónicas materiales precolombinas. Arqueología de los primeros pobladores del*
1091 *Noroeste argentino*, 313-353 (Sociedad Argentina de Antropología, Buenos Aires, 2017)

1092 McKenna, A., Hanna, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-
1093 generation DNA sequencing data. *Genome Res.*, **20**, 1297-1303 (2010).

1094 Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and
1095 sequencing. *Cold Spring Harb. Protoc.*, 2010, pdb-prot5448 (2020).

1096 Michieli, C. T.. *Arqueología de Angualasto: historia, ruinas y cóndores* (Facultad de Filosofía, Humanidades y
1097 Artes Universidad Nacional deSan Juan, 2015).

1098 Michieli, C. T. *Aprovechamiento del agua en las instalaciones "Aguada" de la Provincia de San Juan: nuevas*
1099 *evidencias. Anti, perspectivas y proyectos culturales de América*. 85-98 (Inter-American Committee on Ports
1100 Buenos Aires, 2016).

1101 Oliszewski, N., Martínez, J.G.,Arreguez, G., Gramajo Bühler, M. & Naharro, E. "La transición" vista desde los
1102 valles intermontanos del noroeste argentino: nuevos datos de la Quebrada de Los Corrales (El Infiernillo,
1103 Tucumán, Argentina). *Chungara Revista de antropología Chilena* **50**,71-86 (2018).

1104 Oliszewski, N., Martínez, J. & Caria M. Ocupaciones prehispánicas de altura: el caso de Cueva de los
1105 Corrales 1 (El Infiernillo, Tafí del Valle, Tucumán). *Relac. Soc. Argent. Antropol.* **33**, 209-221 (2008).

1106 Oliszewski, N. & Arreguez G. Los recursos vegetales alimenticios de la Quebrada de Los Corrales en El
1107 Infiernillo, Tucumán, durante el 1° milenio d.C. *Comechingonia* **19**111-140 (2015).

1108 Oliszewski, N. & Babot P. Procesos de selección del poroto común en los valles altos del noroeste argentino
1109 en tiempos prehispánicos. Análisis micro y macroscópico de especímenes arqueobotánicos. In: *Avances y*
1110 *desafíos metodológicos en arqueobotánica: miradas consensuadas y diálogos compartidos desde*
1111 *Sudamérica* (eds.Belmar C. & Lema, V.) 301-324 (Facultad de Patrimonio Cultural y Educaciòn Universidad
1112 SEK Chile, 2015).

1113 Paradis, E., Claude, J., & Strimmer, K. APE: analyses of phylogenetics and evolution in R language.
1114 *Bioinformatics* **20**, 289-290 (2004).

1115 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features.
1116 *Bioinformatics* **26**, 841-842 (2010).

1117 Reimer, P.J., et al. IntCal09 and Marine09 radiocarbon age calibration curves, 0–50,000 years cal BP.
1118 *Radiocarbon* **51**,1111–50 (2009).

1119 Renaut, S. & Rieseberg, L. H. The accumulation of deleterious mutations as a consequence of domestication
1120 and improvement in sunflowers and other compositae crops. *Mol. Biol. Evo*l. **32**, 2273-2283 (2015).

1121 Rendón-Anaya, M., et al. Genomic history of the origin and domestication of common bean unveils its
1122 closest sister species. *Genome Biol* **18**, 60 (2017)..

1123 Rodríguez, M.F. Los grupos humanos y las plantas en la Puna meridional argentina: arqueobotánica de
1124 Antofagasta de la Sierra. *Intersecciones en Antropología* **14**, 315-339 (2013).

1125 Roig, Fidel A. Frutos y semillas arqueológicos de Calingasta, San Juan. In: *La cultura de Ansilta*.(ed. Gambier,
1126 M.) 216-250 (Instituto de Investigaciones Arqueológicas y Museo San Juan, 1977).

1127 Schmutz, J., et al. A reference genome for common bean and genome-wide analysis of dual domestications.
1128 *Nat. Genet.*, **46**, 707 (2014).

1129 Schubert, M., et al. Characterization of ancient and modern genomes by SNP detection and phylogenomic
1130 and metagenomic analysis using PALEOMIX.*Nat. Prot.* **9**, 1056 (2014).

1131 Semper, J. & Lagiglia, H. Excavaciones arqueológicas en el Rincón del Atuel. *Revista Científica de*
1132 *Investigaciones* **1**,89-158 (1968).

1133 Slotte, T. The impact of linked selection on plant genomic variation. *Brief. Funct. Genom.*, **13**, 268-275
1134 (2014).

1135 Somonte, C. & Cohen, M. L. Reocupación y Producción Lítica: un aporte a la historia ocupacional de los
1136 recintos 3 y 4 del sitio agropastoril Punta de la Peña 9 - Sector III (Antofagasta de la Sierra, Catamarca,
1137 Argentina). *Revista Werken* **9**,135- 158 (2006).

1138 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* **100**,
1139 9440-9445 (2003).

1140 Tajima, F. "Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism." *Genetics*
1141 **123**: 585–95 (1989).

1142 Vieira, F. G., Fumagalli, M., Albrechtsen, A., & Nielsen, R. Estimating inbreeding coefficients from NGS data:
1143 impact on genotype calling and allele frequency estimation. *Genome Res*. **23**, 1852-1861 (2013)