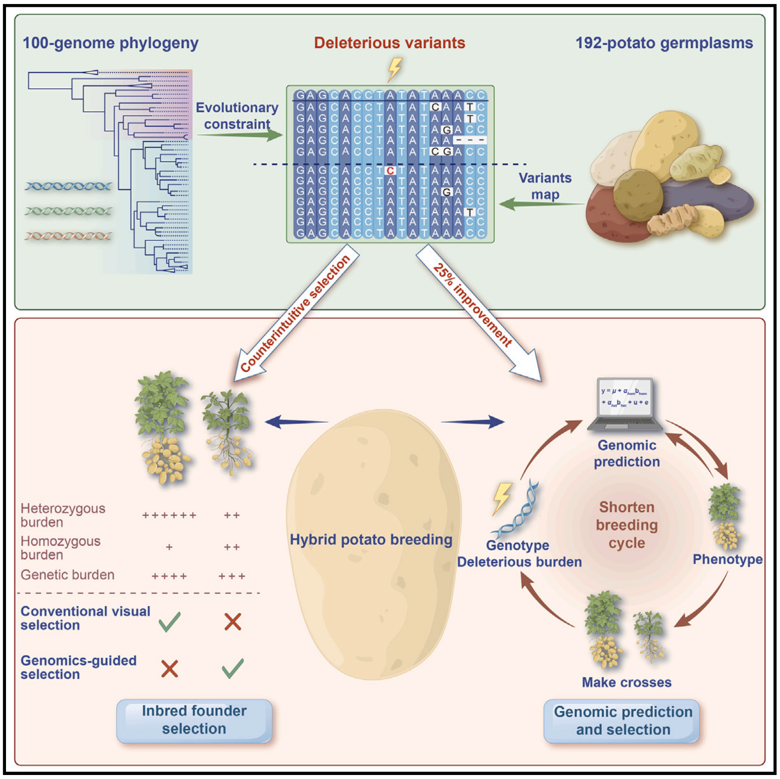
Cell

Article

Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding

Graphical abstract



Highlights

- Deep phylogenomics identifies evolutionary constraints and deleterious mutations
- In the potato genome, 15% of deleterious variants occur at synonymous sites
- Weaker diploids with higher homozygous deleterious burden are better-inbred founders
- Yield prediction accuracy increases by 24.7% by inclusion of deleterious burden

Authors

Yaoyao Wu, Dawei Li, Yong Hu, ...,
Thomas Städler, Edward S. Buckler,
Sanwen Huang

Correspondence

huangsanwen@caas.cn

In brief

Deep phylogenomic analyses of 92 species reveal evolutionary constraints in the nightshade family and deleterious mutations in potato genomes, increasing genome prediction accuracy and supporting the counterintuitive selection of inbred founders to facilitate hybrid potato breeding.

Wu et al., 2023, Cell *186*, 2313–2328
May 25, 2023 © 2023 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2023.04.008

CellPress
OPEN ACCESS

Article

Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding

Yaoyao Wu,[1,2,16] Dawei Li,[1,3,16] Yong Hu,[1,4,16] Hongbo Li,[1,17] Guillaume P. Ramstein,[5,17] Shaoqun Zhou,[1] Xinyan Zhang,[1] Zhigui Bao,[1,6] Yu Zhang,[1,7] Baoxing Song,[8] Yao Zhou,[1,9,10] Yongfeng Zhou,[1] Edeline Gagnon,[11] Tiina Särkinen,[12] Sandra Knapp,[13] Chunzhi Zhang,[1] Thomas Städler,[14] Edward S. Buckler,[2,15] and Sanwen Huang[1,3,18,*]

[1] State Key Laboratory of Tropical Crop Breeding, Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518120, China
[2] Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853, USA
[3] State Key Laboratory of Tropical Crop Breeding, Chinese Academy of Tropical Agricultural Sciences, Haikou, Hainan 571101, China
[4] The AGISCAAS-YNNU Joint Academy of Potato Sciences, Yunnan Normal University, Kunming, Yunnan 650500, China
[5] Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus 8000, Denmark
[6] Department of Molecular Biology, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany
[7] School of Agriculture, Sun Yat-sen University, Shenzhen, Guangdong 518107, China
[8] Peking University Institute of Advanced Agricultural Sciences, Weifang, Shandong 261000, China
[9] Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
[10] College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100094, China
[11] Technische Universität München, TUM School of Life Sciences, Emil-Ramann-Strasse 2, 85354 Freising, Germany
[12] Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh EH3 5LR, UK
[13] Natural History Museum, Cromwell Road, London SW7 5BD, UK
[14] Institute of Integrative Biology and Zurich-Basel Plant Science Center, ETH Zurich, 8092 Zurich, Switzerland
[15] USDA-ARS, Ithaca, NY 14853, USA
[16] These authors contributed equally
[17] These authors contributed equally
[18] Lead contact
*Correspondence: huangsanwen@caas.cn

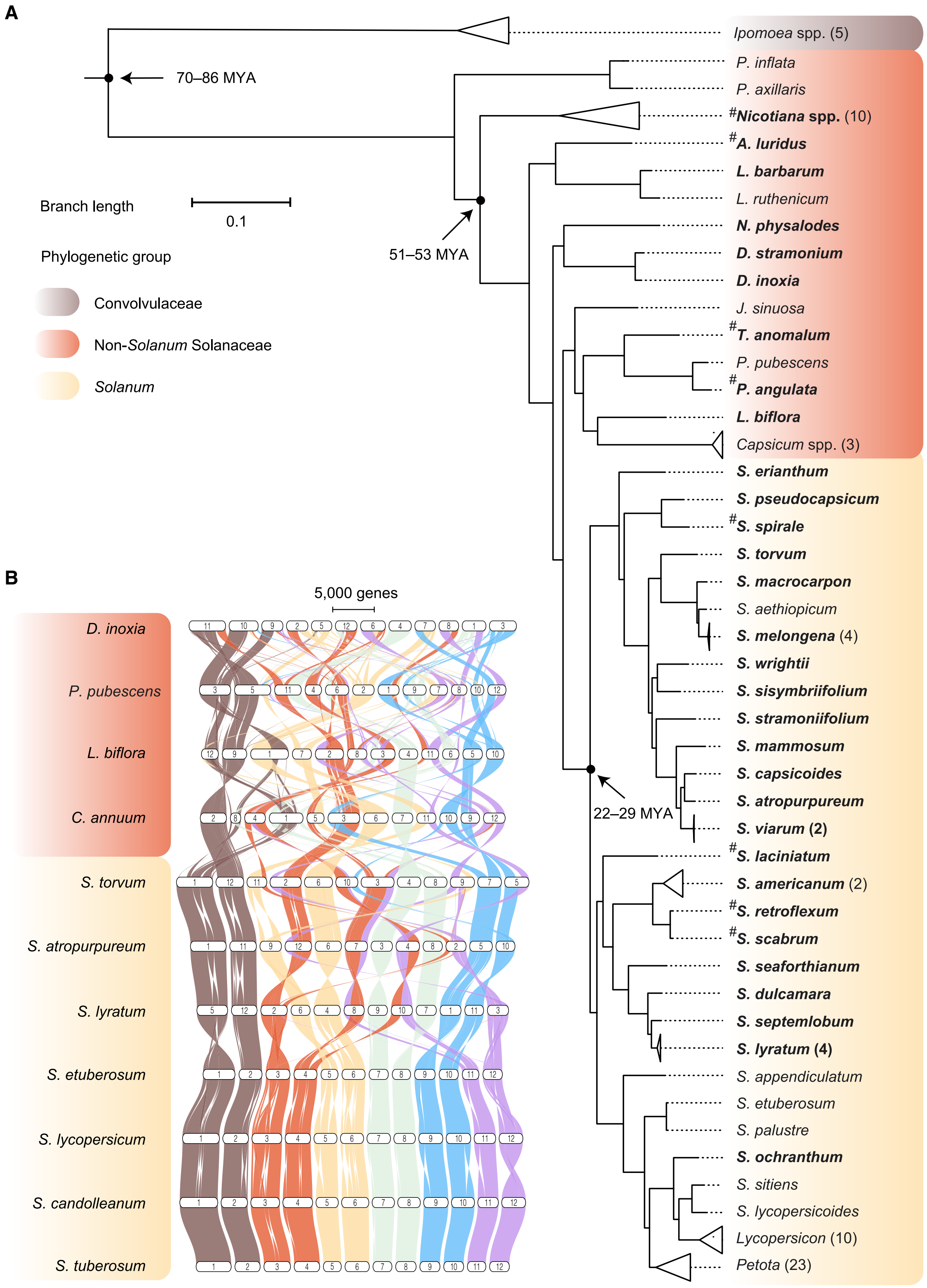
https://doi.org/10.1016/j.cell.2023.04.008

SUMMARY

Hybrid potato breeding will transform the crop from a clonally propagated tetraploid to a seed-reproducing diploid. Historical accumulation of deleterious mutations in potato genomes has hindered the development of elite inbred lines and hybrids. Utilizing a whole-genome phylogeny of 92 Solanaceae and its sister clade species, we employ an evolutionary strategy to identify deleterious mutations. The deep phylogeny reveals the genome-wide landscape of highly constrained sites, comprising ∼2.4% of the genome. Based on a diploid potato diversity panel, we infer 367,499 deleterious variants, of which 50% occur at non-coding and 15% at synonymous sites. Counterintuitively, diploid lines with relatively high homozygous deleterious burden can be better starting material for inbred-line development, despite showing less vigorous growth. Inclusion of inferred deleterious mutations increases genomic-prediction accuracy for yield by 24.7%. Our study generates insights into the genome-wide incidence and properties of deleterious mutations and their far-reaching consequences for breeding.

INTRODUCTION

The reinvention of potato from a clonally propagated autotetraploid to an inbred line-based diploid hybrid reproducing via seeds may transform the breeding of the most important tuber crop from a slow, non-accumulative mode to a fast iterative one.[1–5] The clonally propagated potato has accumulated a markedly large number of deleterious mutations.[6] Deleterious mutations are masked or partially masked in the heterozygous state, and their detrimental effects are exposed during the process of developing inbred lines by repeated rounds of self-fertilization (selfing),[6] which makes developing highly homozygous inbred lines a challenging enterprise. We previously developed two highly homozygous inbred lines by purging of large-effect deleterious mutations via phenotypic screening and genetic analyses; however, these two inbred lines still have large numbers of mild, moderate, and even some highly deleterious mutations in their genomes.[4] Jointly, these deleterious variants result in large negative fitness effects as evidenced by frail growth, reduced fertility, and low yield.[4] Recent genetic investigations



(legend on next page)

suggest that crop breeding could be facilitated by the inference and purging of deleterious mutations.[7–10] To accelerate hybrid potato breeding, it is imperative to identify and understand the properties of deleterious mutations with their effects on fitness at genome-wide scales. Many existing methodologies of predicting deleterious mutations, however, focus on protein-coding regions or face challenges to measure the effect of individual mutations.[11–13] Genomic Evolutionary Rate Profiling (GERP) methodology, a generally validated and effective method based on evolutionary constraints, can identify and quantify deleterious mutations at genome-wide scales including synonymous and non-coding sites.[14–17] A deep phylogeny and genome-wide alignment are crucial for robust prediction of evolutionary constraints and deleterious mutations by this approach.

Using the 100 assembled genomes of 92 species in the plant family Solanaceae and its sister clade Convolvulaceae to construct a deep phylogeny, we here identify constrained sites across the potato genome. Subsequently, our deep-phylogeny approach empowers the prediction and quantification of deleterious mutations, guiding the selection of starting materials for inbred-line development (inbred founders). Moreover, including deleterious burden significantly improves the genomic-prediction (GP) accuracy for agronomically important traits, facilitating genomic selection (GS) and thus hybrid potato breeding.

RESULTS

A deep phylogeny of Solanaceae

Potato (*Solanum tuberosum* L.) belongs to the nightshade family Solanaceae, which includes several economically important crop species such as tomato, chili pepper, eggplant, and tobacco.[18,19] A deep phylogeny encompassing sufficient numbers of substitutions at neutral sites and based on a genome-wide alignment is crucial for robust prediction of evolutionary constraint, estimated by the number of “rejected” substitutions.[14,15] The currently available Solanaceae reference genomes (details in STAR Methods), however, represent a limited number of unevenly distributed phylogenetic branches that omit many of the major lineages in the family. To infer a deep and densely sampled phylogeny of Solanaceae, we *de novo* sequenced 32 additional species (38 genomes) (Table S1). Jointly, the 87 Solanaceae species in our analysis cover most major clades (nine out of 13) in the family,[19] with an emphasis on the genus *Solanum* to which potato belongs.

We assembled the genomes of the 32 species (38 genomes) using Pacific Biosciences (PacBio) high-fidelity (HiFi) data with a mean sequencing depth of 25× (Table S1). The assembled monoploid genome sizes range from 0.8 Gb (*Lycianthes biflora*, a diploid) to 5.0 Gb (*Tubocapsicum anomalum*, an allopolyploid), consistent with those estimated by flow cytometry (Table S1). The mean contig N_{50} length of these 38 assemblies is 39 Mb, indicating their high continuity (Table S1). In addition, 20 of the 38 genomes were assembled at the chromosome level using high-throughput chromatin conformation capture (Hi-C) data (Table S1; Data S1). BUSCO evaluation[20] indicates an average score of 98%, suggesting near-completeness of our assemblies (Table S1). Combining *ab initio* gene prediction, transcript alignment, and evidence of protein homology, we predict 30,621–102,090 protein-coding genes in these genomes (Table S1). The number and length of genes, exons, and introns are comparable with other published Solanaceae genomes, and 58%–77% of putative genes could be assigned functional protein families using the Pfam database[21] (Table S1). These 38 assemblies and their annotation represent one of the highest-quality Solanaceae genomic datasets to date.

We performed whole-genome multiple alignments among the 95 Solanaceae genomes (87 species) and five published Convolvulaceae genomes, the sister clade of Solanaceae (see STAR Methods and Table S1). This generated between 32 Mb (5%) and 446 Mb (61%) of segments aligned to the potato genome, with their alignment length declining with increasing phylogenetic distance (Figures S1A and S1B). We next inferred a species tree for the 95 Solanaceae genomes using 4-fold degenerate sites, with five *Ipomoea* species (Convolvulaceae) as the outgroup (Figure 1A; Table S2). The phylogeny was time-calibrated by constraining the stem node age of the Berry clade as 52.2 million years ago (mya).[22] We estimate the Solanaceae stem node at 80 mya (95% highest posterior density interval, 70–86 mya) and the extant genus *Solanum* to have started diversifying 25 mya (95% highest posterior density interval, 22–29 mya; Figure 1A).

Previous studies documented incongruence between individual gene trees across several nodes of *Solanum* within wild potatoes (*Solanum* section *Petota*) and wild tomatoes (*Solanum* section *Lycopersicon*) and within the pepper tribe Capsiceae[23–26]; however, the extent of such discordance at the whole-genome scale has not been investigated along deeper nodes across Solanaceae. To assess the levels of discordance between the species tree and individual window trees, we split the whole-genome alignment into 3,627 windows (1-Mb length with 200-kb step size), followed by the construction of local phylogenies for each sliding window (Figure S1C). We found that the topology of the main branches is consistent with previous studies,[19,23] but shallower internal nodes exhibit broad incongruence (Figure S1C). These results are consistent with a scenario where different regions of Solanaceae genomes might have undergone diverse evolutionary trajectories, possibly due to variable levels of selective constraint and rapid diversification resulting in incomplete lineage sorting.

Figure 1. Inferred phylogeny and synteny of Solanaceae

(A) Phylogenetic relationships among the 95 Solanaceae and five Convolvulaceae outgroup accessions. Numbers next to nodes denote the estimated divergence time (million years ago [mya]). Species whose genomes were assembled *de novo* are in bold face, and polyploids are indicated by #. Numbers in parentheses indicate the number of species contained in the corresponding collapsed branch. *Petota*, species from *Solanum* section *Petota*. *Lycopersicon*, species from *Solanum* section *Lycopersicon*.

(B) Genome-wide synteny among 11 representative Solanaceae species. The numbered white ellipses denote chromosomes 1–12.

See also Figure S1 and Table S2.

The total branch length of the reconstructed deep phylogeny (indicating the number of substitutions for neutral sites) is 4.05, representing a 5.7-fold increase compared with a recently reported whole-genome phylogeny that mainly covers species of *Solanum* section *Petota* (total branch length of 0.71; Table S2).[27] Generating a deeper and more densely sampled phylogeny facilitates the quantification of the high-density level of evolutionary constraint and thus the characterization of deleterious mutations across the potato genome.

Evolutionary constraint in the potato genome

To investigate whether the whole-genome alignment of Solanaceae species allows robust prediction of evolutionarily constrained sites, we studied genome-wide synteny by assembling segments of syntenic genes among 93 Solanaceae genomes (Table S2; two species used in our phylogenetic reconstruction lack genome annotation). Species of *Solanum* exhibit the highest proportions of syntenic genes with potatoes, averaging 86% (Figure 1B; Table S2). Synteny remains high (72%) even between potato and *Petunia axillaris*, the species with the largest phylogenetic distance to potato among our sampled Solanaceae taxa (Table S2). This largely retained synteny hints that the whole-genome alignment can serve as a reliable dataset to predict levels of evolutionary constraint.

Previous research suggested that evolutionarily constrained sequences are likely to indicate biological functions with fitness consequences upon disruption, i.e., that mutations at these sites potentially reduce fitness.[14,15,17,28] Consequently, identification of such sites by GERP,[14,15] leveraging the deep phylogeny of Solanaceae, offers an avenue for discovering deleterious mutations in potato and other solanaceous crops. To characterize evolutionary constraints across the potato genome, we leveraged the whole-genome alignment and the inferred phylogeny of the 100 genomes by computing the GERP score for each assessable nucleotide site in the potato genome (Figures 2A, 2B, and S2A–S2D). Higher GERP scores indicate a larger evolutionary constraint level. A total of 267,915,549 genomic sites (36.64% of the potato genome) could be probed via high-confidence GERP scores, i.e., those based on a minimum of 15 aligned species (genomes). A total of 2.4% (17,362,955 bp) of the potato genome exhibits signals of evolutionary constraint (moderately and strongly constrained sites at GERP score ≥ 2.75; those including additional mildly constrained sites are reported in the supplemental information [Figures S2E and S2F; Table S2]), of which 36% is in non-coding regions (Figure 2C). A total of 28.8% of all coding sequences (CDSs) are predicted as evolutionarily constrained sites, followed by untranslated regions (UTRs) (7.6%) and introns (4.5%; Figure 2D) for which constrained sites tend to be enriched compared with intergenic regions (fold change, 383×, 101×, and 60×, respectively; χ^2 test, p values < 0.001; Figure S2F; Table S2). Moreover, constrained sites are significantly more enriched in 0-fold degenerate sites (0d) than in 4-fold degenerate sites (4d; 4.08×; χ^2 test, $p < 0.001$; Figure S2F), consistent with the well-established phenomenon that 0d sites are evolutionarily more conserved than 4d sites (Table S2). These patterns are even more pronounced for strongly constrained sites (GERP ≥ 3.5, Figure S2F). Moreover, the more strongly evolutionarily constrained sites encompass higher relative proportions of regions of 0d sites (Figure S2E; Table S2). The overall components and enrichment of evolutionarily constrained sites among these *bona fide* functional elements suggest the robustness of our inferences.

Gene ontology (GO) enrichment reveals that the top 1% constrained genes are significantly enriched in primary biological processes such as carbohydrate biosynthesis, protein glycosylation, and transport (Figure S2G). Among these genes, *Soltu.DM.01G028520.3*, encoding a citrate synthase, catalyzes the first step of the tricarboxylic acid cycle, a major energy-producing metabolic pathway.[29] A total of 83.5% of its CDS and 95.9% of 0-fold degenerate sites (773 out of 806) were identified as evolutionarily constrained by our pipeline (Figure S2H). The catalog of genome-wide constrained sites opens avenues to further characterize functional elements, especially in previously underrepresented non-coding regions in Solanaceae species.

A genome-wide atlas of deleterious variants

Robust identification of deleterious mutations is pivotal for developing elite inbred lines.[4,6,17,30] However, research in potatoes has thus far focused solely on large-effect deleterious mutations or nonsynonymous variants in gene-coding regions,[4,6] neglecting possible mild, moderate ones, as well as those in non-CDSs and synonymous variants. The above analyses of whole-genome constrained sites facilitate a broader understanding of the landscape of deleterious mutations across the potato genome. To identify deleterious variants in diploid potatoes, we used a representative diploid potato diversity panel, consisting of two inbred lines and 190 landrace potatoes, covering the four main subgroups of diploid potato landraces, *S. tuberosum* group Stenotomum, *S. tuberosum* group Phureja, *S. tuberosum* group Goniocalyx, and *S. tuberosum* group Ajanhuiri.[27] These diploid landraces, with 0.02 average heterozygosity, are valuable germplasm resources for diploid hybrid potato breeding, especially with the purpose of searching for potential inbred founders.

Single-nucleotide polymorphisms (SNPs) within the diploid potato diversity panel at evolutionarily constrained sites were considered deleterious variants (Figures 3A and 3B). The GERP scores for such sites also enable a rough quantification of the magnitude of deleterious variants' potential effects. A total of 367,499 SNPs (0.6% of 58,597,787 SNPs in this panel) were thus classified as putatively harboring moderately and highly deleterious variants at the threshold GERP $\geq$ 2.75. Those including the inferred mildly deleterious variants were reported in the supplemental information (Table S3). The majority of putatively deleterious variants are rare (274,003 of 367,499 [75%]; Figures 3C and S3A). Moreover, rare variants are more likely to be inferred as deleterious for both coding and non-coding regions (Figures 3D, S3B, and S3C). These results are in accordance with previous observations that purifying selection keeps deleterious variants at a low frequency in population.[31,32]

It is increasingly recognized that mutations in non-coding regions and synonymous variants may also contribute to reduced fitness.[31,33–35] In this diploid diversity panel, about 50.5% of inferred deleterious variants reside in non-coding regions (Figure 3E) and 15.1% of the inferred deleterious variants are synonymous (Figure S3D), which were previously inaccessible to

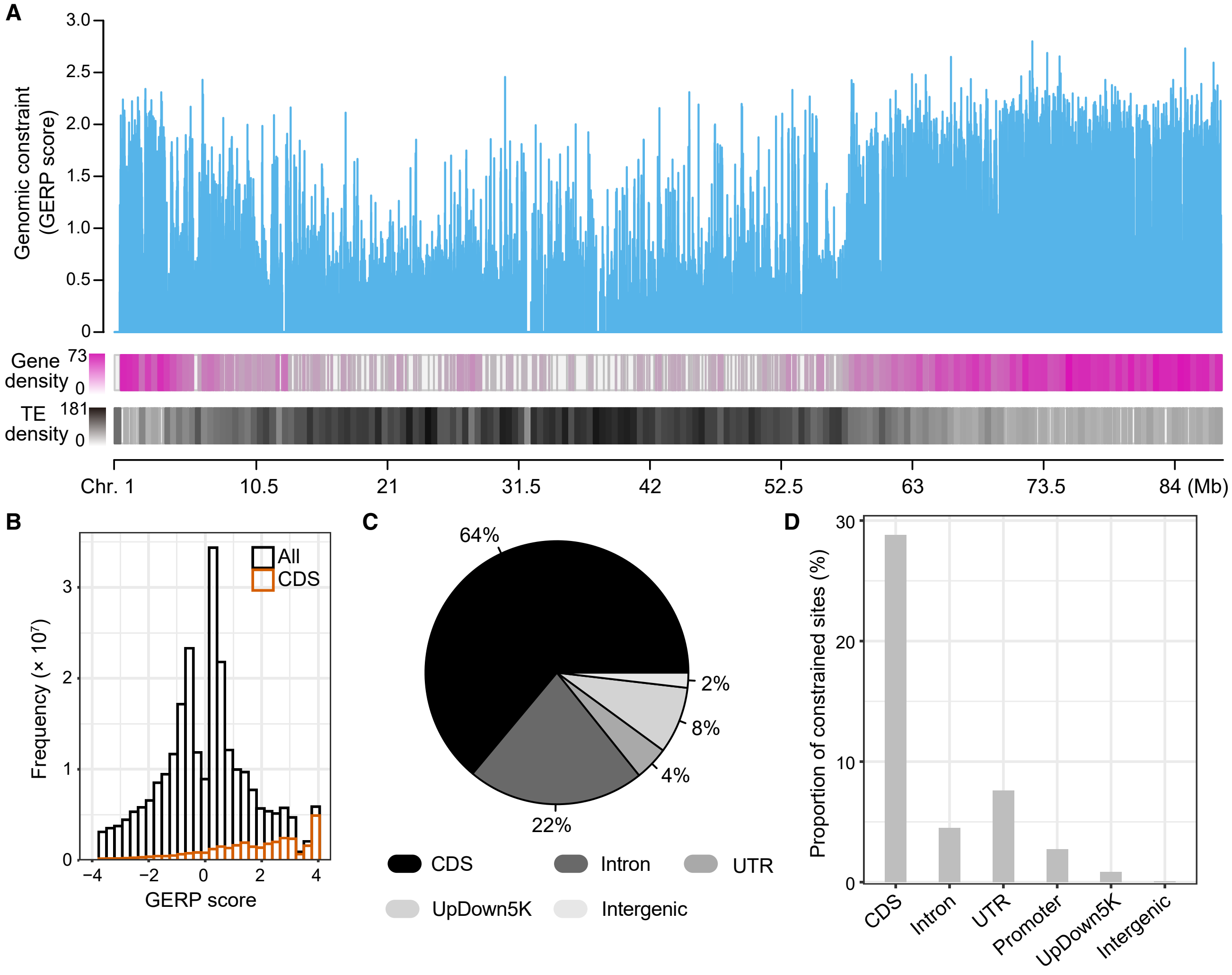


Figure 2. Evolutionary constraint across the potato genome

(A) Levels of genomic constraint of potato chromosome 1 in terms of the genomic evolutionary rate profiling (GERP) scores, visualized by splitting the chromosome into 8,860 non-overlapping 10-kb windows. Gene density is represented as the number of genes per 500 kb, and transposable element (TE) density in terms of number of TEs (size >1 kb) per 500 kb.

(B) Distribution of GERP scores in whole-genome sequences (ALL) and protein-coding sequence (CDS).

(C) Distribution of constrained sites across the potato genome.

(D) Proportion of constrained sites estimated as the number of constrained sites divided by all sites, shown separately for coding regions, introns, UTRs, promoters (1-kb upstream of CDS), UpDown5K (5-kb upstream and downstream of genes), and intergenic regions.

See also Figure S2 and Table S2.

scrutiny.[6] As the deleterious threshold increases, the number of inferred deleterious variants decreases, the proportion of nonsynonymous deleterious variants increases, and the fraction of non-coding deleterious variants decreases (Figure S3D; Table S3). Within the diploid diversity panel, our pipeline predicts only 0.33% of all SNPs in non-coding regions to be deleterious, in contrast with the much higher predicted fraction of deleterious variants in coding regions (8.15%; Figure 3F), a 25-fold enrichment in coding compared with non-coding regions. In addition, deleterious variants are more enriched in nonsynonymous sites, followed by synonymous sites, UTRs, and introns, compared with those in intergenic sites (Figure S3E). Moreover, the enrichment level increases for highly deleterious variants (Figure S3F). Our inferred proportions are broadly compatible with data in other species.[31,36]

Annotating all deleterious variants using snpEff,[37] we found 2,042 deleterious variants leading to premature stop codons (Figures S3G and S3H). The genes with deleterious variants caused by premature stop (stop gained) are enriched in important biological processes such as DNA replication and tRNA-related processes (Table S3). We identified deleterious mutations in all nine loci previously reported to be associated with

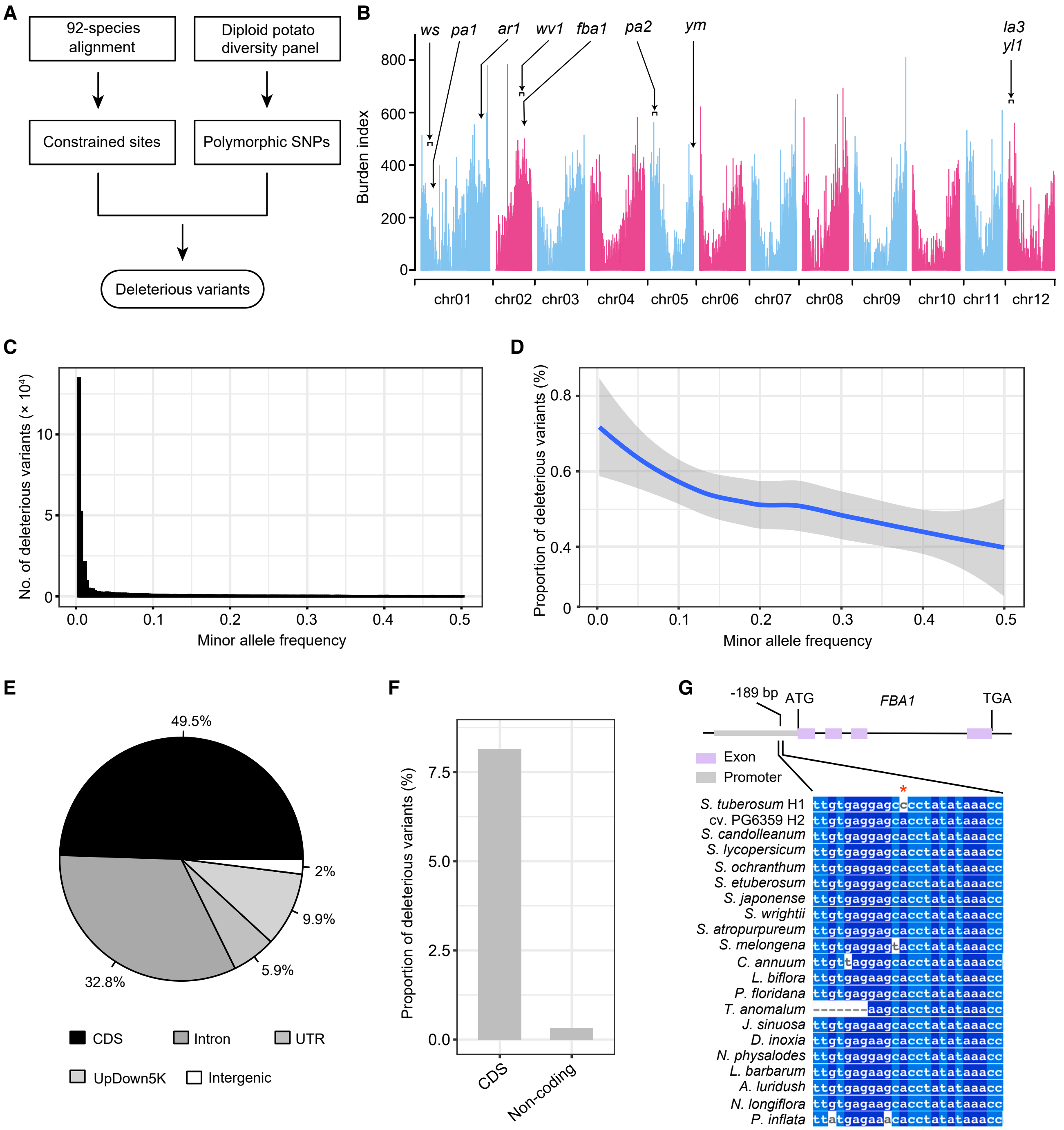


Figure 3. Deleterious variants in the diploid potato diversity panel

(A) Pipeline for identifying deleterious variants.

(B) Genome-wide distribution of deleterious mutation burden, calculated by summing GERP scores of all deleterious variant sites in 73,135 non-overlapping 10-kb windows. Previously reported fitness-related genes and quantitative trait loci (QTLs) underlying deleterious phenotypes are marked with black arrows.

(C) Allele-frequency spectrum of inferred deleterious variants in the diploid potato diversity panel.

(D) Decrease in the proportion of deleterious variants per SNP with increasing minor allele frequency in the diploid potato diversity panel.

(E) Distribution of all inferred deleterious variants among different genomic regions. The proportion was estimated by the rounding-off method.

(F) Proportion of deleterious variants per SNP among CDS and non-coding sites in the diploid potato diversity panel, calculated as the number of deleterious variants divided by the number of all SNPs within CDS and non-coding regions, respectively.

(legend continued on next page)

unfavorable phenotypes,[4,6,38] including three previously cloned genes: *abnormal rooting 1* (*ar1*), *floral bud abortion 1* (*fba1*), and *yellow margin* (*ym*) (Table S3). This represented a retrospective investigation of deleterious mutations affecting fitness-related phenotypic traits. We then focused on the gene *FBA1*, encoding a basic helix-loop-helix transcription factor that has previously been mapped using a self-fertilizing population of a diploid potato clone, PG6359.[4] It is orthologous to *Arabidopsis DYSFUNCTIONAL TAPETUM 1* (*DYT1*), which regulates stamen development.[39] Mutants in potatoes exhibit a significantly lower expression of *fba1* than the wild type, and knockout of this gene yields flower-bud abortion phenotype.[4] We identified 18 nucleotide variants between the two haplotypes of the diploid genome of PG6359 in FBA1's promoter region (1-kb upstream of CDS). Among them, we posit the A-to-C change at a site with the highest GERP score (GERP = 2.84) to represent a deleterious mutation that may be involved in regulating the expression of *FBA1* (Figure 3G). This is but one example of how the whole-genome map of deleterious alleles empowers further identification and characterization of functional sites or elements, which could guide targeted purging of deleterious mutations in potato breeding, especially those in non-coding regions.

Prediction of deleterious mutations guides hybrid potato breeding

To quantify the deleterious mutation burden of each accession in the diploid diversity panel, we first counted the number of deleterious mutations in the heterozygous and homozygous states, respectively (Figure S4A), and then enumerated the genome-wide deleterious burden in heterozygous and homozygous states, respectively (see STAR Methods and Table S4). Across the diploid landraces, the deleterious burden in the heterozygous state (heterozygous burden) is strongly negatively correlated with that in the homozygous state (homozygous burden, $R = -0.94$, $p < 2.2 \times 10^{-16}$; Figure 4A).

The diploid potato landraces constitute a valuable germplasm resource for inbred founders.[6] During inbred-line development, deleterious mutations masked or partially masked in the heterozygous state are exposed, making it difficult to develop highly homozygous inbred lines. This is exemplified by the line "Solyntus," which is still heterozygous for 20% of the genome despite nine generations of selfing.[40] The total burden includes both masked and exposed burdens, here referred to as the genetic burden (numerically equal to the additive burden), that represent the potential fitness burden that can be passed on to offspring and affect the fitness of its descendants (see STAR Methods for details).[41,42] Hence, robust prediction of genetic burden can guide and inform the selection of founders for inbred-line development (inbred founders).

To identify promising candidate material for inbred-line development, we quantified the genetic burden for each landrace in the diploid diversity panel. Intriguingly, we found that their genetic burden is strongly negatively correlated with homozygous burden (recessive burden; $R = -0.67$, $p = 3.8 \times 10^{-21}$) but strongly positively correlated with the heterozygous burden ($R = 0.89$, $p = 8.6 \times 10^{-57}$; Figures 4A and S4B). This pattern suggests that promising inbred founders (i.e., landraces with relatively low genetic burden) should be sought among lines with higher homozygous burden and lower heterozygous burden.

The expressed burden (homozygous burden + heterozygous burden weighted by the dominance coefficient *h*; here genome-wide average $h = 0.1$) contributes to decreasing fitness (see STAR Methods for details).[41–43] A corollary of these deductions is that those promising inbred founders likely exhibit more expressed burden and thus should be less vigorous (B_Genetic vs. B_Expressed, $R = -0.54$; B_Heter vs. B_Expressed, $R = -0.87$; B_Homo vs. B_Expressed, $R = 0.99$, $p < 1 \times 10^{-12}$; Figures S4C and S4D). Traditional selection criteria based on phenotypic performance tend to retain individuals with strong vigor. Our analyses, however, suggest that vigorous individuals (i.e., landraces with a low expressed and homozygous burden but high heterozygous burden) are likely to harbor higher genetic burden (Figures 4A and S4B–S4D), implying that they will transmit a higher deleterious burden to offspring, thus leading to eventual failure of inbred-line development. Therefore, we propose that individuals with a lower total genetic burden, despite being less vigorous owing to the relatively higher expressed burden and higher homozygous burden, should be considered the better founder material for inbred-line development.

RH89-039-16 (hereafter referred to as RH) is one of the heterozygous diploids in the diploid diversity panel with ideal performance for several agronomic traits (Figures 4A and S4B, red dot). Previous efforts, however, have failed to develop it into an inbred line by multi-generation selfing.[4] PG6359 and E86-69 (Figures 4A and S4B, blue dots) are two accessions from the diploid diversity panel that has been successfully developed into highly homozygous diploid inbred lines A6-26 and E4-63, respectively (Figures 4A and S4B, orange dots).[4] In accordance with our observations from the diploid potato landraces, RH carries 17% and 21% higher genetic burden than E86-69 and PG6359, respectively. A similar pattern can be observed for C10-20, another diploid landrace that could not be developed into an inbred line (11% and 15% higher genetic burden than E86-69 and PG6359, respectively; Figure 4A, red dot; Table S4). These observations on available inbred founders support our inference that a lower total genetic burden is an excellent predictor of successful inbred founders.

When a pair of neighboring heterozygous deleterious mutations resides on two different homologous chromosomes (so-called "repulsion-phase" deleterious mutations), the efficacy of purging deleterious mutations is reduced due to Hill-Robertson interference, and recombination is needed to purge both deleterious alleles.[9,44] Typically, however, there are well below 100 recombination events per generation per individual.[45–47] Hence, development of inbred lines may be hindered by abundant deleterious mutations in repulsion phase. Approximately 40% of

(G) A 24-bp nucleotide alignment of part of the *FBA1* promoter (*Soltu.DM.02G019340.1*, chr02: 33,595,387–33,596,386) among 20 representative Solanaceae species. *S. tuberosum* cv. PG6359 H1 and H2 represent the two haplotypes of the PG6359 genome assembly. The red asterisk marks the deleterious variants in PG6359.
See also Figure S3 and Table S3.

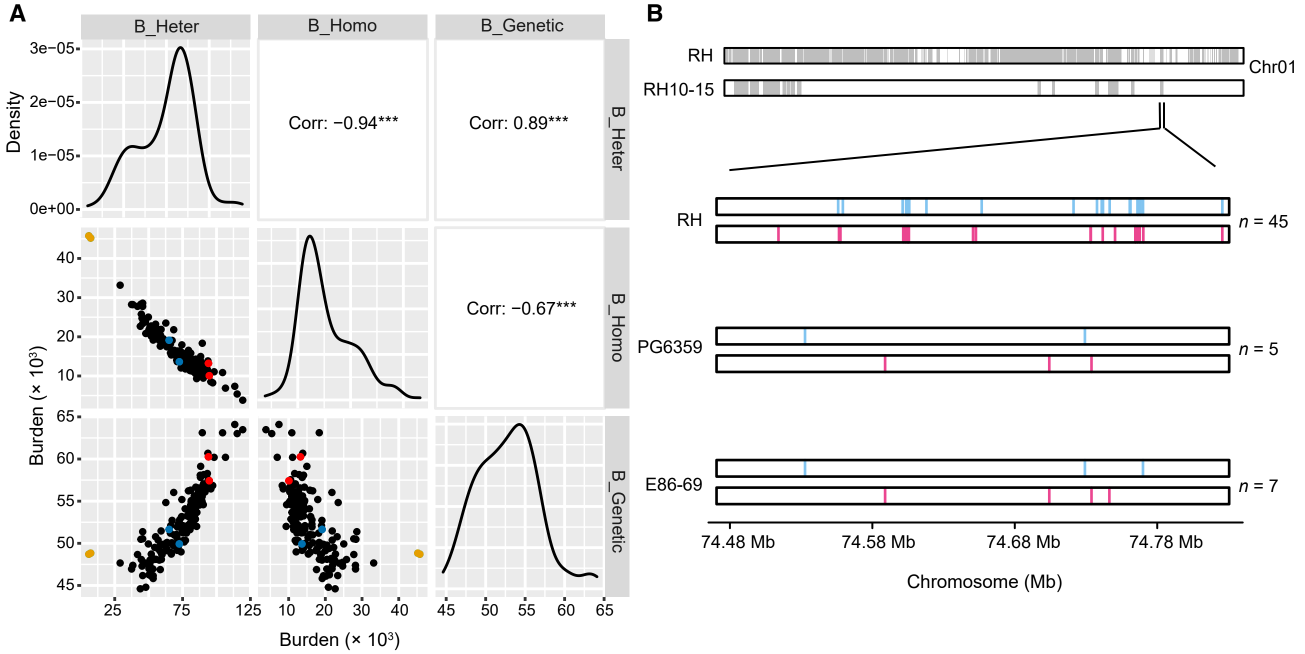


Figure 4. Prediction of deleterious mutations facilitates hybrid potato breeding

(A) Correlations among genetic burden (B_Genetic), homozygous burden (B_Homo), and heterozygous burden (B_Heter). Correlation coefficients (*R*) among different burdens are also shown. ***p < 0.001 in Pearson correlation tests. Failed inbred founders RH and C10-20 are highlighted as red dots. Successfully inbred founders PG6359 and E86-69 are denoted by blue dots, and the two inbred lines A6-26 and E4-63 are marked by orange dots.

(B) A zoom-in view of the distribution of heterozygous deleterious mutations in the repulsion phase within a 500-kb heterozygous genomic fragment of RH on chromosome 1 (chr01: 74,385,000–74,910,000). Heterozygous regions in RH and RH10-15 are shown in gray. Repulsion-phase deleterious mutations present in the two haplotypes of RH, PG6359 and E86-69, are illustrated in light blue and pink, respectively, and their numbers are listed on the right.

See also Figure S4 and Table S4.

deleterious heterozygous mutations in these three diploid lines are in the repulsion phase, with RH harboring 46% and 78% more deleterious mutations in the repulsion phase than PG6359 and E86-69, respectively (Figure S4E; Table S4). This suggests that additional recombination events will be required for purging these mutations in line RH.

Following four generations of recurrent selfing of line RH, we observed an unexpectedly high proportion of genomic regions that remain heterozygous. For example, 30.92 Mb of heterozygous genomic regions (heterozygous fraction $\geq$ 2%) characterize the RH10-15 genome, a progeny of the RH S_4 population (Figure S4F). Among them, 11.2 Mb are significantly enriched for higher deleterious mutations in the heterozygous state, compared with those being homozygous after four generations of selfing (Table S4). We zoomed in on a ~500-kb heterozygous fragment with a markedly higher number of repulsion-phase deleterious mutations in RH compared with PG6359 and E86-69 (Figure 4B; Table S4). Genomic regions that could not be made homozygous in RH usually carry higher numbers of heterozygous deleterious mutations in the repulsion phase. Homozygosity for these regions would possibly lead to an intolerably high deleterious burden, thus resulting in substantially reduced fitness.

Weighted deleterious mutation burden improves GP

GP has become a powerful tool to predict the genetic value among candidate individuals in animal and plant breeding, and incorporating deleterious mutations should be useful for GP and GS.[48–52] Unlike in other major crops with sexual propagation, deleterious mutations in cultivated potatoes have accumulated in an unmitigated fashion, plausibly associated with their long-term clonal propagation.[6,38,53] However, contributions of deleterious mutations to the accuracy of GP in potato have not yet been evaluated. To examine this issue, we used an F_2 population with genotype and phenotypes well measured, which was generated by selfing the diploid F_1 hybrid potato whose parents are the diploid inbred lines A6-26 and E4-63.[4,54] Across this F_2 panel, there were 5,527,427 SNPs with an average heterozygosity of 0.51. Across the 2,603 recombination bins among the 1,064 F_2 individuals with a mean bin length of 282.4 kb, we identified a total of 23,655 SNPs with deleterious variants, with 1,960 of the 2,603 bins containing at least one deleterious variant. The deleterious burden (weighted sum of GERP scores of all deleterious alleles) across these bins ranges from 0 to 844.8 (Figure S5A).

To assess how the magnitude of deleterious burden may affect agronomic traits, we calculated the whole-genome burden separately for each F_2 individual. Among these F_2 progeny, the heterozygous burden displays a strongly negative correlation with the homozygous burden, consistent with our results from the diploid diversity panel (Figure S5B). We analyzed the distribution of five agronomical phenotypes (yield, plant height, tuber number, tuber size, and flowering time) in this population (Figure 5A). Notably, the homozygous burden is significantly

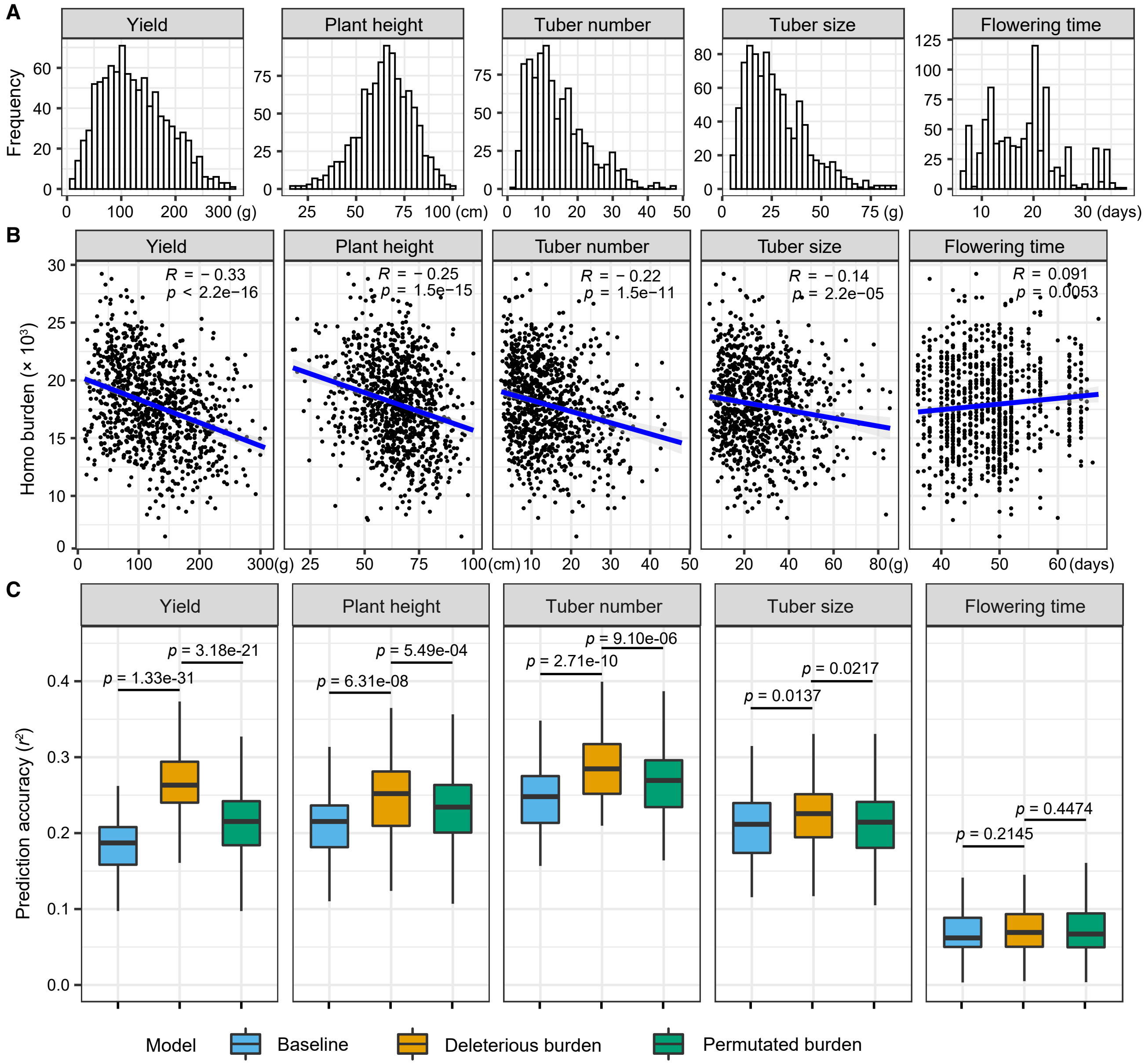


Figure 5. Weighted deleterious mutation burden improves genomic prediction in potato

(A) Phenotypic distribution of yield, plant height, tuber number, tuber size, and flowering time in the F_2 population.

(B) Correlations between the five phenotypes and the per-individual homozygous burden among individuals in the F_2 population. Pearson's correlation coefficients (*R*) between the phenotypic traits and homozygous burdens are indicated; p values were computed with Pearson correlation tests. Gray shadows represent the 95% confidence intervals, estimated by fitting a linear model using the "lm()" function in *R*.

(C) Genomic-prediction accuracy of five traits using the baseline model (blue boxes, without fitting deleterious burden), the deleterious burden model (orange boxes, fitting the deleterious burden; see STAR Methods), and the permutated burden model (green boxes, 100 times fitting the randomly shuffled deleterious mutation burden). Upper and lower edges of the boxes denote 75% and 25% quartiles, and the central line indicates the median. Whiskers extend to the lower hinge −1.5× interquartile range and upper hinge +1.5× interquartile range of the data. p values obtained by Student's t tests (one-tailed) are indicated.

See also Figure S5 and Table S5.

negatively correlated with agronomic traits, among which *bona fide* fitness-related agronomical traits reveal higher levels of negative correlation: *R* = −0.33, −0.25, and −0.22 for yield, plant height, and tuber number, respectively (Figure 5B). Given the strong negative correlation between homozygous burden and heterozygous burden among individuals (Figure S5B), we calculated partial correlation coefficients between homozygous burden, heterozygous burden, and phenotypic traits. Both homozygous burden and heterozygous burden are negatively correlated with fitness-related agronomical traits, with homozygous burden revealing higher partial correlation coefficients for yield and plant height (Figure S5C). These results indicate that deleterious mutation burden is significantly associated with fitness and thus should be more emphasized in potato breeding.

Given the significantly negative correlation between deleterious mutation burden and these fitness-related agronomical traits in this population, we performed GP for the five traits. Based on our baseline model, which accounts for the genomic relationship without considering the effects of deleterious mutations (see STAR Methods), the prediction accuracy (r^2) is 0.183, 0.212, and 0.248 for yield, plant height, and tuber number, respectively (Figure 5C; Table S5). After fitting the homozygous and heterozygous burden in a linear mixed model, the prediction accuracy rose to 0.264, 0.250, and 0.289 for yield, plant height, and tuber number, respectively, corresponding to 44.6%, 17.8%, and 16.4% improvements compared with the baseline model (Figure 5C; Table S5). We also observed significant relative increases (24.7% for yield, 7.2% for plant height, and 7.6% for tuber number; all p values < 0.05) compared with the prediction accuracy based on 100 permutations of deleterious mutation burden (see STAR Methods, Figure 5C, and Table S5). Furthermore, our GP accuracy reaches ∼50% of the broad-sense heritability for these three traits reported in previous studies[55–58] (Table S5). However, the prediction accuracy for tuber size and flowering time did not increase substantially (Figure 5C). Overall, our results suggest that the inclusion of deleterious mutations improves the power of GP for complex fitness-related agronomical traits controlled by multiple small-effect loci, revealing the potential to enhance the efficacy of GS. The breeding value predicted by this model (i.e., including deleterious mutation) can be applied to enhance the decision-making process, reducing the costs of phenotyping and time spent on early-generation decisions during the potato breeding process such as selection of parental lines and genomics-assisted purging of deleterious mutations.

DISCUSSION

Utilizing the genome-wide signature of the evolutionary constraint, we unveiled the landscape of deleterious mutations in the potato genome. The allele-frequency spectra of inferred genome-wide deleterious mutations and their enrichment in *bona fide* functional elements are consistent with expectations from evolutionary genetics and artificial selection, as well as the retrospective investigation of previously known candidate loci; all these features indicate the robustness of our genome-wide identification and quantification of deleterious mutations. Inclusion of inferred deleterious mutations increases the accuracy of GP for complex fitness-related agronomic traits, further suggesting that deleterious mutations are reliably identified and quantified at genome-wide scales by the GERP approach, here via leveraging the deep and densely sampled phylogeny of Solanaceae. Our analyses and dissection of the total deleterious burden for diploid landraces lead to a more robust identification of promising inbred founders based on the magnitude of their genetic burden, a metric balancing the current inbreeding rate and potential performance of the future inbred line by weighting heterozygous and homozygous burden into a single index accounting for segregation during selfing. We further uncovered the likely reasons why landraces previously selected based on their phenotype failed to be developed into inbred lines, which appears counterintuitive.

The first and critical step of hybrid potato breeding is the development of highly homozygous inbred lines, which eliminated the large-effect deleterious mutations.[1,4] Potato, as a clonally propagated crop, has accumulated a large number of deleterious mutations, a major cause of inbreeding depression, thus frustrating the successful development of inbred lines.[4,6,30] The selection of potential inbred founders that can strike a balance between the speed of inbreeding and vigorous performance of the homozygous inbred lines is thus important for hybrid potato breeding. Our seemingly counterintuitive proposal can identify lines that will yield progeny with relatively less burden on inbreeding, with the promise of shortening breeding time and increasing the overall success of developing inbred lines.

Maps of the genome-wide deleterious variants can also guide the next steps in hybrid breeding that aim to increase the frequency of favorable alleles and decrease the frequency of deleterious alleles in inbred lines and F_1 hybrids. Moreover, when large-scale genomic editing (with hundreds of edited sites per generation) becomes available in potatoes, genomics-assisted purging can be applied more efficiently,[9] and accessions with plenty of beneficial alleles such as RH and Solyntus can be used as founders. Once sufficient inbred lines have been developed, breeders can recombine favorable alleles and mask deleterious mutations in the heterozygous state to obtain vigorous F_1 hybrids. During all these processes, our inferred deleterious burden and genomic-prediction model can be directly applied in genomics-assisted purging and GS, assisting decision-making in terms of gains per unit of time, thereby accelerating hybrid potato breeding by reducing the breeding cycle, phenotyping cost, and the time spent on early-generation selections.

Although our study focused on diploid potatoes, we found that tetraploid potatoes harbor more heterozygous deleterious mutations and fewer homozygous deleterious mutations than diploid potatoes ($p < 0.001$, Figure S4A; Table S4). This makes intuitive sense as tetraploids have better field performance than diploid landraces and are more difficult to be developed into inbred lines. In addition to being useful for diploid potato breeding, our data on deleterious mutations would be useful for researchers and breeders to predict the performance of tetraploid cultivars.

The genome sequences, annotation, synteny, and alignments presented in this study offer valuable resources for genetic, genomic studies, and future breeding in this important plant family.[59] The genome-wide identification of evolutionarily constrained sites, especially for non-coding genomic regions among Solanaceae species, will facilitate further discovery and characterization of functional and regulatory elements, as well as guide the isolation of candidate genes underlying agronomic traits. Importantly, it is straightforward to apply our analytic pipeline comprising evolutionary and population genomics to closely related Solanaceae crop species such as tomato and eggplant, with a promise to enhance the efficacy of their GS and genomics-assisted purging of deleterious mutations by various strategies such as the application of genome editing and recurrent selection.

Limitations of the study

Due to the difficulty of obtaining seeds or live plants of rare lineages, the current genome-level phylogeny does not cover all genera of Solanaceae; the inclusion of species from additional genera will further improve the prediction of evolutionary constraints and localization of deleterious mutations. Importantly, our analyses were limited to SNPs; other variants such as inversions, large insertions/deletions, and copy-number variation may also affect the fitness as they are likely to behave as deleterious variants.[60] Further advances will also be made with better handling of paralogous sequences in whole-genome alignments that are ubiquitous in plant genomes due to prevalent, ancient whole-genome duplications.[61–63]

Unlike other crops like maize and sorghum, hybrid potato breeding is still in its infancy, and the available diploid potato samples and data are limited. Hence, we had to restrict our analyses to the two available populations (the diploid diversity panel and the F_2 panel) and four inbred-line founders. Our quantification of expressed burden based on *h* estimates obtained from an F_2 population may introduce biases relative to a broader population, but our *h* estimate is broadly congruent with previous experiments studies.[64,65] With the future release of increasing numbers of diploid potato genomes and potato panels, we may dissect the expressed burden more precisely, and more analyses can be performed to further validate the power of leveraging information on genome-wide deleterious mutations in accelerating hybrid potato breeding.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Plant growth conditions
- METHOD DETAILS
  - ○ Sample selection and sequencing
  - ○ Genome assembly of the 38 Solanaceae accessions
  - ○ Repetitive element annotation
  - ○ Prediction of protein-coding genes
  - ○ Whole-genome alignment
  - ○ Phylogenetic analyses
  - ○ Estimation of divergence times
  - ○ Detection of genomic synteny
  - ○ Identification of constrained sites
  - ○ Identification of deleterious variants in the diploid diversity panel
  - ○ Estimation of deleterious mutation burdens
  - ○ Identification of deleterious mutations in tetraploid potatoes
  - ○ Determination of heterozygous genomic regions of RH and RH10-15
  - ○ Genomic prediction
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2023.04.008.

ACKNOWLEDGMENTS

We thank Guangtao Zhu, Pei Wang, Yuxin Jia, and Dié Tang for plant grow and critical comments. We thank the U.S. National Plant Germplasm (Agricultural Research Service, U.S. Department of Agriculture) and the Germplasm Bank of Wild Species in China for providing the Solanaceae accessions. We thank the AGIS and CAAS-YNNU Joint Academy of Potato Sciences for greenhouse assistance. This work was supported by Guangdong Major Project of Basic and Applied Basic Research (2021B0301030004 to S.H.), the National Key Research and Development Program of China (2019YFA0906200), the National Natural Science Foundation of China (32102376), the Agricultural Science and Technology Innovation Program (CAAS-ZDRW202101), the Special Funds for Science Technology Innovation and Industrial Development of Shenzhen Dapeng New District (RC201901-05), and Shenzhen Outstanding Talents Training Fund.

AUTHOR CONTRIBUTIONS

S.H. and E.S.B. conceived the project. S.H. designed the study. Y.W., Y.H., and H.L. contributed to genome assembly, genome annotation, and synteny analyses. Y.W. performed phylogenetic, evolutionary-constraint, and deleterious-mutation analyses. Y.W. and G.P.R. contributed to genomic-prediction analyses. D.L. constructed the F_2 population and contributed to the sampling, genotypic, and phenotypic analysis of the F_2 population. Y.H. detected SNPs. H.L. performed the species dating analyses. Y.W., S.Z., Y.H., D.L., and X.Z. contributed to collecting, growing, sampling, and RNA extraction of Solanaceae material. Y.H., S.Z., and D.L. contributed to greenhouse work. T. Särkinen, S.K., and E.G. identified all samples, confirmed the phylogeny, and assisted in molecular dating analysis. S.H., E.S.B., T. Städler, Z.B., Y. Zhang, C.Z., Yao Zhou, and Yongfeng Zhou assisted in bioinformatics analyses. S.H. and Y.W. composed the outline of the manuscript. Y.W. wrote the draft manuscript. H.L., S.H., T. Städler, and Y.W. wrote and revised the manuscript. D.L., G.P.R., S.K., T. Särkinen, E.G., and B.S. revised the final manuscript.

DECLARATION OF INTERESTS

Agricultural Genomics Institute at Shenzhen has one pending patent application based on this study. S.H. is a member of the *Cell* advisory board.

Received: August 24, 2022
Revised: February 20, 2023
Accepted: April 5, 2023
Published: May 4, 2023

REFERENCES

1. Jansky, S.H., Charkowski, A.O., Douches, D.S., Gusmini, G., Richael, C., Bethke, P.C., Spooner, D.M., Novy, R.G., De Jong, H., De Jong, W.S., et al. (2016). Reinventing potato as a diploid inbred line-based crop. Crop Sci. *56*, 1412–1422. https://doi.org/10.2135/cropsci2015.12.0740.
2. Li, Y., Li, G., Li, C., Qu, D., and Huang, S. (2013). Prospects of diploid hybrid breeding in potato. Chin. Potato J. *27*, 96–99.

3. Lindhout, P., Meijer, D., Schotte, T., Hutten, R.C.B., Visser, R.G.F., and van Eck, H.J. (2011). Towards F_1 hybrid seed potato breeding. Potato Res. *54*, 301–312. https://doi.org/10.1007/s11540-011-9196-z.

4. Zhang, C., Yang, Z., Tang, D., Zhu, Y., Wang, P., Li, D., Zhu, G., Xiong, X., Shang, Y., Li, C., et al. (2021). Genome design of hybrid potato. Cell *184*, 3873–3883.e12. https://doi.org/10.1016/j.cell.2021.06.006.

5. Stokstad, E. (2019). The new potato. Science *363*, 574–577. https://doi.org/10.1126/science.363.6427.574.

6. Zhang, C., Wang, P., Tang, D., Yang, Z., Lu, F., Qi, J., Tawari, N.R., Shang, Y., Li, C., and Huang, S. (2019). The genetic basis of inbreeding depression in potato. Nat. Genet. *51*, 374–378. https://doi.org/10.1038/s41588-018-0319-1.

7. Morrell, P.L., Buckler, E.S., and Ross-Ibarra, J. (2011). Crop genomics: advances and applications. Nat. Rev. Genet. *13*, 85–96. https://doi.org/10.1038/nrg3097.

8. Gaut, B.S., Seymour, D.K., Liu, Q., and Zhou, Y. (2018). Demography and its effects on genomic variation in crop domestication. Nat. Plants *4*, 512–520. https://doi.org/10.1038/s41477-018-0210-1.

9. Wallace, J.G., Rodgers-Melnick, E., and Buckler, E.S. (2018). On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. Annu. Rev. Genet. *52*, 421–444. https://doi.org/10.1146/annurev-genet-120116-024846.

10. Zhou, Y., Massonnet, M., Sanjak, J.S., Cantu, D., and Gaut, B.S. (2017). Evolutionary genomics of grape (*Vitis vinifera* ssp. *vinifera*) domestication. Proc. Natl. Acad. Sci. USA *114*, 11715–11720. https://doi.org/10.1073/pnas.1709257114.

11. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. *76*, 7.20.21–27.20.41. https://doi.org/10.1002/0471142905.hg0720s76.

12. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the functional effect of amino acid substitutions and indels. PLoS One *7*, e46688. https://doi.org/10.1371/journal.pone.0046688.

13. Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814. https://doi.org/10.1093/nar/gkg509.

14. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025. https://doi.org/10.1371/journal.pcbi.1001025.

15. Cooper, G.M., Stone, E.A., Asimenos, G., NISC Comparative Sequencing Program, Green, E.D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913. https://doi.org/10.1101/gr.3577405.

16. Henn, B.M., Botigué, L.R., Bustamante, C.D., Clark, A.G., and Gravel, S. (2015). Estimating the mutation load in human genomes. Nat. Rev. Genet. *16*, 333–343. https://doi.org/10.1038/nrg3931.

17. Moyers, B.T., Morrell, P.L., and McKay, J.K. (2018). Genetic costs of domestication and improvement. J. Hered. *109*, 103–116. https://doi.org/10.1093/jhered/esx069.

18. Spooner, D.M., Ghislain, M., Simon, R., Jansky, S.H., and Gavrilenko, T. (2014). Systematics, diversity, genetics, and evolution of wild and cultivated potatoes. Bot. Rev. *80*, 283–383. https://doi.org/10.1007/s12229-014-9146-y.

19. Särkinen, T., Bohs, L., Olmstead, R.G., and Knapp, S. (2013). A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. BMC Evol. Biol. *13*, 214. https://doi.org/10.1186/1471-2148-13-214.

20. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

21. Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res. *42*, D222–D230. https://doi.org/10.1093/nar/gkt1223.

22. Lisa De-Silva, D., Mota, L.L., Chazot, N., Mallarino, R., Silva-Brandão, K.L., Piñerez, L.M., Freitas, A.V., Lamas, G., Joron, M., Mallet, J., et al. (2017). North Andean origin and diversification of the largest ithomiine butterfly genus. Sci. Rep. *7*, 45966. https://doi.org/10.1038/srep45966.

23. Gagnon, E., Hilgenhof, R., Orejuela, A., McDonnell, A., Sablok, G., Aubriot, X., Giacomin, L., Gouvêa, Y., Bragionis, T., Stehmann, J.R., et al. (2022). Phylogenomic discordance suggests polytomies along the backbone of the large genus *Solanum*. Am. J. Bot. *109*, 580–601. https://doi.org/10.1002/ajb2.1827.

24. Wu, M., Kostyun, J.L., and Moyle, L.C. (2019). Genome sequence of *Jaltomata* addresses rapid reproductive trait evolution and enhances comparative genomics in the hyper-diverse Solanaceae. Genome Biol. Evol. *11*, 335–349. https://doi.org/10.1093/gbe/evy274.

25. Spalink, D., Stoffel, K., Walden, G.K., Hulse-Kemp, A.M., Hill, T.A., Van Deynze, A., and Bohs, L. (2018). Comparative transcriptomics and genomic patterns of discordance in Capsiceae (Solanaceae). Mol. Phylogenet. Evol. *126*, 293–302. https://doi.org/10.1016/j.ympev.2018.04.030.

26. Pease, J.B., Haak, D.C., Hahn, M.W., and Moyle, L.C. (2016). Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. *14*, e1002379. https://doi.org/10.1371/journal.pbio.1002379.

27. Tang, D., Jia, Y., Zhang, J., Li, H., Cheng, L., Wang, P., Bao, Z., Liu, Z., Feng, S., Zhu, X., et al. (2022). Genome evolution and diversity of wild and cultivated potatoes. Nature *606*, 535–541. https://doi.org/10.1038/s41586-022-04822-x.

28. Kono, T.J.Y., Lei, L., Shih, C.H., Hoffman, P.J., Morrell, P.L., and Fay, J.C. (2018). Comparative genomics approaches accurately predict deleterious variants in plants. G3 (Bethesda) *8*, 3321–3329. https://doi.org/10.1534/g3.118.200563.

29. Wiegand, G., and Remington, S.J. (1986). Citrate synthase: structure, control, and mechanism. Annu. Rev. Biophys. Biophys. Chem. *15*, 97–117. https://doi.org/10.1146/annurev.bb.15.060186.000525.

30. Charlesworth, D., and Willis, J.H. (2009). The genetics of inbreeding depression. Nat. Rev. Genet. *10*, 783–796. https://doi.org/10.1038/nrg2664.

31. Eyre-Walker, A., and Keightley, P.D. (2007). The distribution of fitness effects of new mutations. Nat. Rev. Genet. *8*, 610–618. https://doi.org/10.1038/nrg2146.

32. Lozano, R., Gazave, E., Dos Santos, J.P.R., Stetter, M.G., Valluru, R., Bandillo, N., Fernandes, S.B., Brown, P.J., Shakoor, N., Mockler, T.C., et al. (2021). Comparative evolutionary genetics of deleterious load in sorghum and maize. Nat. Plants *7*, 17–24. https://doi.org/10.1038/s41477-020-00834-5.

33. Shen, X., Song, S., Li, C., and Zhang, J. (2022). Synonymous mutations in representative yeast genes are mostly strongly non-neutral. Nature *606*, 725–731. https://doi.org/10.1038/s41586-022-04823-w.

34. Song, B., Buckler, E.S., Wang, H., Wu, Y., Rees, E., Kellogg, E.A., Gates, D.J., Khaipho-Burch, M., Bradbury, P.J., and Ross-Ibarra, J. (2021). Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize. Genome Res. *31*, 1245–1257. https://doi.org/10.1101/gr.266528.120.

35. Kremling, K.A.G., Chen, S.Y., Su, M.H., Lepak, N.K., Romay, M.C., Swarts, K.L., Lu, F., Lorant, A., Bradbury, P.J., and Buckler, E.S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature *555*, 520–523. https://doi.org/10.1038/nature25966.

36. 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth,

G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65. https://doi.org/10.1038/nature11632.

37. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso-2; iso-3*. Fly *6*, 80–92. https://doi.org/10.4161/fly.19695.

38. Zhou, Q., Tang, D., Huang, W., Yang, Z., Zhang, Y., Hamilton, J.P., Visser, R.G.F., Bachem, C.W.B., Robin Buell, C., Zhang, Z., et al. (2020). Haplotype-resolved genome analyses of a heterozygous diploid potato. Nat. Genet. *52*, 1018–1023. https://doi.org/10.1038/s41588-020-0699-x.

39. Farquharson, K.L. (2016). A domain in the bHLH transcription factor DYT1 is critical for anther development. Plant Cell *28*, 997–998. https://doi.org/10.1105/tpc.16.00331.

40. van Lieshout, N., van der Burgt, A., de Vries, M.E., ter Maat, M., Eickholt, D., Esselink, D., van Kaauwen, M.P.W., Kodde, L.P., Visser, R.G.F., Lindhout, P., et al. (2020). Solyntus, the new highly contiguous reference genome for potato (*Solanum tuberosum*). G3 (Bethesda) *10*, 3489–3495. https://doi.org/10.1534/g3.120.401550.

41. Bertorelle, G., Raffini, F., Bosse, M., Bortoluzzi, C., Iannucci, A., Trucchi, E., Morales, H.E., and van Oosterhout, C. (2022). Genetic load: genomic estimates and applications in non-model animals. Nat. Rev. Genet. *23*, 492–503. https://doi.org/10.1038/s41576-022-00448-x.

42. Kojima, K.-I. (1970). Mathematical Topics in Population Genetics (Springer) https://doi.org/10.1007/978-3-642-46244-3.

43. Morton, N.E., Crow, J.F., and Muller, H.J. (1956). An estimate of the mutational damage in man from data on consanguineous marriages. Proc. Natl. Acad. Sci. USA *42*, 855–863. https://doi.org/10.1073/pnas.42.11.855.

44. Hill, W.G., and Robertson, A. (1966). The effect of linkage on limits to artificial selection. Genet. Res. *8*, 269–294. https://doi.org/10.1017/S001667230800949X.

45. Otto, S.P., and Payseur, B.A. (2019). Crossover interference: shedding light on the evolution of recombination. Annu. Rev. Genet. *53*, 19–44. https://doi.org/10.1146/annurev-genet-040119-093957.

46. McMullen, M.D., Kresovich, S., Villeda, H.S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., et al. (2009). Genetic properties of the maize nested association mapping population. Science *325*, 737–740. https://doi.org/10.1126/science.1174320.

47. Roessler, K., Muyle, A., Diez, C.M., Gaut, G.R., Bousios, A., Stitzer, M.C., Seymour, D.K., Doebley, J.F., Liu, Q., and Gaut, B.S. (2019). The genomics of selfing in maize (Zea mays ssp. mays): catching purging in the act https://doi.org/10.1101/594812.

48. Daetwyler, H.D., Calus, M.P.L., Pong-Wong, R., de los Campos, G., and Hickey, J.M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics *193*, 347–365. https://doi.org/10.1534/genetics.112.147983.

49. Edwards, S.M., Sørensen, I.F., Sarup, P., Mackay, T.F.C., and Sørensen, P. (2016). Genomic prediction for quantitative traits is improved by mapping variants to gene ontology categories in *Drosophila melanogaster*. Genetics *203*, 1871–1883. https://doi.org/10.1534/genetics.116.187161.

50. Kono, T.J.Y., Liu, C., Vonderharr, E.E., Koenig, D., Fay, J.C., Smith, K.P., and Morrell, P.L. (2019). The fate of deleterious variants in a barley genomic prediction population. Genetics *213*, 1531–1544. https://doi.org/10.1534/genetics.119.302733.

51. Ramstein, G.P., and Buckler, E.S. (2022). Prediction of evolutionary constraint by genomic annotations improves functional prioritization of genomic variants in maize. Genome Biol. *23*, 183. https://doi.org/10.1186/s13059-022-02747-2.

52. Yang, J., Mezmouk, S., Baumgarten, A., Buckler, E.S., Guill, K.E., McMullen, M.D., Mumm, R.H., and Ross-Ibarra, J. (2017a). Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. PLoS Genet. *13*, e1007019. https://doi.org/10.1371/journal.pgen.1007019.

53. Ramu, P., Esuma, W., Kawuki, R., Rabbi, I.Y., Egesi, C., Bredeson, J.V., Bart, R.S., Verma, J., Buckler, E.S., and Lu, F. (2017). Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. Nat. Genet. *49*, 959–963. https://doi.org/10.1038/ng.3845.

54. Li, D., Lu, X., Zhu, Y., Pan, J., Zhou, S., Zhang, X., Zhu, G., Shang, Y., Huang, S., and Zhang, C. (2022). The multi-omics basis of potato heterosis. J. Integr. Plant Biol. *64*, 671–687. https://doi.org/10.1111/jipb.13211.

55. Fekadu, A., Petros, Y., and Zelleke, H. (2013). Genetic variability and association between agronomic characters in some potato (*Solanum tuberosum* L.) genotypes in SNNPRS, Ethiopia. Int. J. Biodivers. Conserv. *5*, 523–528. https://doi.org/10.5897/IJBC2013.0548.

56. Luthra, S.K. (2001). Heritability, genetic advance and character association in potato. J. Indian Potato Assoc. *28*, 1–3.

57. Slater, A.T., Cogan, N.O., Forster, J.W., Hayes, B.J., and Daetwyler, H.D. (2016). Improving genetic gain with genomic selection in autotetraploid potato. Plant Genome *9*, 1–15. https://doi.org/10.3835/plantgenome2016.02.0021.

58. Slater, A.T., Wilson, G.M., Cogan, N.O., Forster, J.W., and Hayes, B.J. (2014). Improving the analysis of low heritability complex traits for enhanced genetic gain in potato. Theor. Appl. Genet. *127*, 809–820. https://doi.org/10.1007/s00122-013-2258-7.

59. Li, H., Yang, X., Shang, Y., Zhang, Z., and Huang, S. (2023). Vegetable biology and breeding in the genomics era. Sci. China Life Sci. *66*, 226–250. https://doi.org/10.1007/s11427-022-2248-6.

60. Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., Cantu, D., and Gaut, B.S. (2019). The population genetics of structural variants in grapevine domestication. Nat. Plants *5*, 965–979. https://doi.org/10.1038/s41477-019-0507-8.

61. Rensing, S.A. (2014). Gene duplication as a driver of plant morphogenetic evolution. Curr. Opin. Plant Biol. *17*, 43–48. https://doi.org/10.1016/j.pbi.2013.11.002.

62. Clark, J.W., and Donoghue, P.C.J. (2018). Whole-genome duplication and plant macroevolution. Trends Plant Sci. *23*, 933–945. https://doi.org/10.1016/j.tplants.2018.07.006.

63. Frith, M.C., and Kawaguchi, R. (2015). Split-alignment of genomes finds orthologies more accurately. Genome Biol. *16*, 106. https://doi.org/10.1186/s13059-015-0670-9.

64. Houle, D., Hughes, K.A., Assimacopoulos, S., and Charlesworth, B. (1997). The effects of spontaneous mutation on quantitative traits. II. Dominance of mutations with effects on life-history traits. Genet. Res. *70*, 27–34. https://doi.org/10.1017/s001667239700284x.

65. García-Dorado, A., and Caballero, A. (2000). On the average coefficient of dominance of deleterious spontaneous mutations. Genetics *155*, 1991–2001. https://doi.org/10.1093/genetics/155.4.1991.

66. Hao, Y., Bao, W., Li, G., Gagoshidze, Z., Shu, H., Yang, Z., Cheng, S., Zhu, G., and Wang, Z. (2021). The chromosome-based genome provides insights into the evolution in water spinach. Sci. Hortic. *289*, 110501. https://doi.org/10.1016/j.scienta.2021.110501.

67. Hoshino, A., Jayakumar, V., Nitasaka, E., Toyoda, A., Noguchi, H., Itoh, T., Shin-I, T., Minakuchi, Y., Koda, Y., Nagano, A.J., et al. (2016). Genome sequence and analysis of the Japanese morning glory *Ipomoea nil*. Nat. Commun. *7*, 13295. https://doi.org/10.1038/ncomms13295.

68. Li, M., Yang, S., Xu, W., Pu, Z., Feng, J., Wang, Z., Zhang, C., Peng, M., Du, C., Lin, F., et al. (2019). The wild sweetpotato (*Ipomoea trifida*) genome provides insights into storage root development. BMC Plant Biol. *19*, 119. https://doi.org/10.1186/s12870-019-1708-z.

69. Wu, S., Lau, K.H., Cao, Q., Hamilton, J.P., Sun, H., Zhou, C., Eserman, L., Gemenet, D.C., Olukolu, B.A., Wang, H., et al. (2018). Genome sequences of two diploid wild relatives of cultivated sweetpotato reveal targets for genetic improvement. Nat. Commun. *9*, 4580. https://doi.org/10.1038/s41467-018-06983-8.

70. Yang, J., Moeinzadeh, M.-H., Kuhl, H., Helmuth, J., Xiao, P., Haas, S., Liu, G., Zheng, J., Sun, Z., Fan, W., et al. (2017b). Haplotype-resolved sweet potato genome traces back its hexaploidization history. Nat. Plants *3*, 696–703. https://doi.org/10.1038/s41477-017-0002-z.

71. Barchi, L., Rabanus-Wallace, M.T., Prohens, J., Toppino, L., Padmarasu, S., Portis, E., Rotino, G.L., Stein, N., Lanteri, S., and Giuliano, G. (2021). Improved genome assembly and pan-genome provide key insights into eggplant domestication and breeding. Plant J. *107*, 579–596. https://doi.org/10.1111/tpj.15313.

72. Bolger, A., Scossa, F., Bolger, M.E., Lanz, C., Maumus, F., Tohge, T., Quesneville, H., Alseekh, S., Sørensen, I., Lichtenstein, G., et al. (2014). The genome of the stress-tolerant wild tomato species *Solanum pennellii*. Nat. Genet. *46*, 1034–1038. https://doi.org/10.1038/ng.3046.

73. Bombarely, A., Moser, M., Amrad, A., Bao, M., Bapaume, L., Barry, C.S., Bliek, M., Boersma, M.R., Borghi, L., Bruggmann, R., et al. (2016). Insight into the evolution of the Solanaceae from the parental genomes of *Petunia hybrida*. Nat. Plants *2*, 16074. https://doi.org/10.1038/nplants.2016.74.

74. Cao, Y.L., Li, Y.L., Fan, Y.F., Li, Z., Yoshida, K., Wang, J.Y., Ma, X.K., Wang, N., Mitsuda, N., Kotake, T., et al. (2021). Wolfberry genomes and the evolution of *Lycium* (Solanaceae). Commun. Biol. *4*, 671. https://doi.org/10.1038/s42003-021-02152-8.

75. Edwards, K.D., Fernandez-Pozo, N., Drake-Stowe, K., Humphry, M., Evans, A.D., Bombarely, A., Allen, F., Hurst, R., White, B., Kernodle, S.P., et al. (2017). A reference genome for *Nicotiana tabacum* enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. BMC Genomics *18*, 448. https://doi.org/10.1186/s12864-017-3791-6.

76. Kim, S., Park, J., Yeom, S.-I., Kim, Y.-M., Seo, E., Kim, K.-T., Kim, M.-S., Lee, J.M., Cheong, K., Shin, H.-S., et al. (2017). New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. Genome Biol. *18*, 210. https://doi.org/10.1186/s13059-017-1341-9.

77. Kim, S., Park, M., Yeom, S.I., Kim, Y.M., Lee, J.M., Lee, H.A., Seo, E., Choi, J., Cheong, K., Kim, K.T., et al. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. Nat. Genet. *46*, 270–278. https://doi.org/10.1038/ng.2877.

78. Lu, J., Luo, M., Wang, L., Li, K., Yu, Y., Yang, W., Gong, P., Gao, H., Li, Q., Zhao, J., et al. (2021). The *Physalis floridana* genome provides insights into the biochemical and morphological evolution of *Physalis* fruits. Hortic. Res. *8*, 244. https://doi.org/10.1038/s41438-021-00705-w.

79. Molitor, C., Kurowski, T.J., Fidalgo de Almeida, P.M., Eerolla, P., Spindlow, D.J., Kashyap, S.P., Singh, B., Prasanna, H.C., Thompson, A.J., and Mohareb, F.R. (2021). *De novo* genome assembly of *Solanum sitiens* reveals structural variation associated with drought and salinity tolerance. Bioinformatics *37*, 1941–1945. https://doi.org/10.1093/bioinformatics/btab048.

80. Paajanen, P., Kettleborough, G., López-Girona, E., Giolai, M., Heavens, D., Baker, D., Lister, A., Cugliandolo, F., Wilde, G., Hein, I., et al. (2019). A critical comparison of technologies for a plant genome sequencing project. GigaScience *8*, giy163. https://doi.org/10.1093/gigascience/giy163.

81. Pham, G.M., Hamilton, J.P., Wood, J.C., Burke, J.T., Zhao, H., Vaillancourt, B., Ou, S., Jiang, J., and Buell, C.R. (2020). Construction of a chromosome-scale long-read reference genome assembly for potato. GigaScience *9*, giaa100. https://doi.org/10.1093/gigascience/giaa100.

82. Powell, A.F., Feder, A., Li, J., Schmidt, M.H., Courtney, L., Alseekh, S., Jobson, E.M., Vogel, A., Xu, Y., Lyon, D., et al. (2022). A *Solanum lycopersicoides* reference genome facilitates insights into tomato specialized metabolism and immunity. Plant J. *110*, 1791–1810. https://doi.org/10.1111/tpj.15770.

83. Sierro, N., Battey, J.N.D., Bovet, L., Liedschulte, V., Ouadi, S., Thomas, J., Broye, H., Laparra, H., Vuarnoz, A., Lang, G., et al. (2018). The impact of genome evolution on the allotetraploid *Nicotiana rustica* – an intriguing story of enhanced alkaloid production. BMC Genomics *19*, 855. https://doi.org/10.1186/s12864-018-5241-5.

84. Sierro, N., Battey, J.N.D., Ouadi, S., Bovet, L., Goepfert, S., Bakaher, N., Peitsch, M.C., and Ivanov, N.V. (2013). Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. Genome Biol. *14*, R60. https://doi.org/10.1186/gb-2013-14-6-r60.

85. Song, B., Song, Y., Fu, Y., Kizito, E.B., Kamenya, S.N., Kabod, P.N., Liu, H., Muthemba, S., Kariba, R., Njuguna, J., et al. (2019). Draft genome sequence of *Solanum aethiopicum* provides insights into disease resistance, drought tolerance, and the evolution of the genome. GigaScience *8*, giz115. https://doi.org/10.1093/gigascience/giz115.

86. Wu, M., Haak, D.C., Anderson, G.J., Hahn, M.W., Moyle, L.C., and Guerrero, R.F. (2021). Inferring the genetic basis of sex determination from the genome of a dioecious nightshade. Mol. Biol. Evol. *38*, 2946–2957. https://doi.org/10.1093/molbev/msab089.

87. Xu, S., Brockmöller, T., Navarro-Quezada, A., Kuhl, H., Gase, K., Ling, Z., Zhou, W., Kreitzer, C., Stanke, M., Tang, H., et al. (2017). Wild tobacco genomes reveal the evolution of nicotine biosynthesis. Proc. Natl. Acad. Sci. USA *114*, 6133–6138. https://doi.org/10.1073/pnas.1700073114.

88. Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. *11*, 1432. https://doi.org/10.1038/s41467-020-14998-3.

89. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. Nat. Methods *18*, 170–175. https://doi.org/10.1038/s41592-020-01056-5.

90. Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. *3*, 95–98. https://doi.org/10.1016/j.cels.2016.07.002.

91. Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., and Aiden, E.L. (2017). *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science *356*, 92–95. https://doi.org/10.1126/science.aal3327.

92. Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. *20*, 275. https://doi.org/10.1186/s13059-019-1905-y.

93. Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360. https://doi.org/10.1038/nmeth.3317.

94. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295. https://doi.org/10.1038/nbt.3122.

95. Stanke, M., and Morgenstern, B. (2005). Augustus: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. *33*, W465–W467. https://doi.org/10.1093/nar/gki458.

96. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. *33*, 6494–6506. https://doi.org/10.1093/nar/gki937.

97. Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016). BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics *32*, 767–769. https://doi.org/10.1093/bioinformatics/btv661.

98. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

99. Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinform. *12*, 491. https://doi.org/10.1186/1471-2105-12-491.

100. Blum, M., Chang, H.Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., et al. (2021). The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. *49*, D344–D354. https://doi.org/10.1093/nar/gkaa977.

101. Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I.T., Novak, A.M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., et al. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature *587*, 246–251. https://doi.org/10.1038/s41586-020-2871-y.

102. Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274. https://doi.org/10.1093/molbev/msu300.

103. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

104. Chen, Y., Song, W., Xie, X., Wang, Z., Guan, P., Peng, H., Jiao, Y., Ni, Z., Sun, Q., and Guo, W. (2020). A collinearity-incorporating homology inference strategy for connecting emerging assemblies in the Triticeae tribe as a pilot practice in the plant pangenomic era. Mol. Plant *13*, 1694–1708. https://doi.org/10.1016/j.molp.2020.09.019.

105. Lovell, J.T., Sreedasyam, A., Schranz, M.E., Wilson, M.A., Carlson, J.W., Harkess, A., Emms, D., Goodstein, D., and Schmutz, J. (2022). GENESPACE: syntenic pan-genome annotations for eukaryotes https://doi.org/10.1101/2022.03.09.483468.

106. Alexa, A., and Rahnenfuhrer, J. (2022). topGO: Enrichment Analysis for Gene Ontology R Package Version 2.50.0. https://doi.org/10.18129/B9.bioc.topGO.

107. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

108. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

109. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

110. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509.

111. Rohde, P.D., Fourie Sørensen, I., and Sørensen, P. (2020). qgg: an R package for large-scale quantitative genetic analyses. Bioinformatics *36*, 2614–2615. https://doi.org/10.1093/bioinformatics/btz955.

112. Clifford, D., and McCullagh, P. (2006). The regress function. R News *6*, 6–10.

113. Potato Genome Sequencing Consortium, Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., et al. (2011). Genome sequence and analysis of the tuber crop potato. Nature *475*, 189–195. https://doi.org/10.1038/nature10158.

114. Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature *485*, 635–641. https://doi.org/10.1038/nature11119.

115. Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., et al. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. Nature *606*, 527–534. https://doi.org/10.1038/s41586-022-04808-9.

116. Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. Methods *58*, 268–276. https://doi.org/10.1016/j.ymeth.2012.05.001.

117. Cheng, C.Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S., and Town, C.D. (2017). Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J. *89*, 789–804. https://doi.org/10.1111/tpj.13415.

118. Wu, Y., Johnson, L., Song, B., Romay, C., Stitzer, M., Siepel, A., Buckler, E., and Scheben, A. (2022). A multiple genome alignment workflow shows the impact of repeat masking and parameter tuning on alignment of functional regions in plants. Plant Genome *15*, e20204. https://doi.org/10.1002/tpg2.20204.

119. Wilf, P., Carvalho, M.R., Gandolfo, M.A., and Cúneo, N.R. (2017). Eocene lantern fruits from Gondwanan Patagonia and the early origins of Solanaceae. Science *355*, 71–75. https://doi.org/10.1126/science.aag2737.

120. Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. *40*, e49. https://doi.org/10.1093/nar/gkr1293.

121. Bombarely, A., Rosli, H.G., Vrebalov, J., Moffett, P., Mueller, L.A., and Martin, G.B. (2012). A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. Mol. Plant Microbe Interact. *25*, 1523–1530. https://doi.org/10.1094/MPMI-06-12-0148-TA.

122. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

123. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

124. Hardigan, M.A., Laimbeer, F.P.E., Newton, L., Crisovan, E., Hamilton, J.P., Vaillancourt, B., Wiegert-Rininger, K., Wood, J.C., Douches, D.S., Farré, E.M., et al. (2017). Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. Proc. Natl. Acad. Sci. USA *114*, E9999–E10008. https://doi.org/10.1073/pnas.1714380114.

125. Sun, H., Jiao, W.B., Krause, K., Campoy, J.A., Goel, M., Folz-Donahue, K., Kukat, C., Huettel, B., and Schneeberger, K. (2022). Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. Nat. Genet. *54*, 342–348. https://doi.org/10.1038/s41588-022-01015-0.

126. Hoopes, G., Meng, X., Hamilton, J.P., Achakkagari, S.R., de Alves Freitas Guesdes, F., Bolger, M.E., Coombs, J.J., Esselink, D., Kaiser, N.R., Kodde, L., et al. (2022). Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity. Mol. Plant *15*, 520–536. https://doi.org/10.1016/j.molp.2022.01.003.

127. Bao, Z., Li, C., Li, G., Wang, P., Peng, Z., Cheng, L., Li, H., Zhang, Z., Li, Y., Huang, W., et al. (2022). Genome architecture and tetrasomic inheritance of autotetraploid potato. Mol. Plant *15*, 1211–1226. https://doi.org/10.1016/j.molp.2022.06.009.

128. Schrinner, S.D., Mari, R.S., Ebler, J., Rautiainen, M., Seillier, L., Reimer, J.J., Usadel, B., Marschall, T., and Klau, G.W. (2020). Haplotype threading: accurate polyploid phasing from long reads. Genome Biol. *21*, 252. https://doi.org/10.1186/s13059-020-02158-1.

129. Mátyás, L. (1999). Generalized Method of Moments Estimation (Cambridge University Press) https://doi.org/10.1057/978-1-349-95189-5_2486.

130. Simmons, M.J., and Crow, J.F. (1977). Mutations affecting fitness in *Drosophila* populations. Annu. Rev. Genet. *11*, 49–78. https://doi.org/10.1146/annurev.ge.11.120177.000405.

131. Huber, C.D., Durvasula, A., Hancock, A.M., and Lohmueller, K.E. (2018). Gene expression drives the evolution of dominance. Nat. Commun. *9*, 2750. https://doi.org/10.1038/s41467-018-05281-7.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
PacBio CCS, Hi-C, and RNA sequencing data	This study	BioProject: PRJNA839598
Genome assemblies	This study	NGDC:https://ngdc.cncb.ac.cn/search/?dbId=gwh&q=PRJCA010759
Public Convolvulaceae genome assemblies	Hao et al.[66]; Hoshino et al.[67]; Li et al.[68]; Wu et al.[69]; Yang et al.[70]	N/A
Public Solanaceae genome assemblies	Barchi et al.[71]; Bolger et al.[72]; Bombarely et al.[73]; Cao et al.[74]; Edwards et al.[75]; Kim et al.[76]; Kim et al.[77]; Lu et al.[78]; Molitor et al.[79]; Paajanen et al.[80]; Pham et al.[81]; Powell et al.[82]; Sierro et al.[83]; Sierro et al.[84]; Song et al.[85]; Tang et al.[27]; Wu et al.[86]; Wu et al.[24]; Xu et al.[87]	N/A
Software and algorithms		
GenomeScope2.0 v1.0.0	Ranallo-Benavidez et al.[88]	https://github.com/tbenavi1/genomescope2.0
Hifiasm v0.16.1-r375	Cheng et al.[89]	https://github.com/chhylp123/hifiasm
juicer v1.6	Durand et al.[90]	https://github.com/aidenlab/juicer
3d-dna v180922	Dudchenko et al.[91]	https://github.com/aidenlab/3d-dna
BUSCO v5.2.2	Simão et al.[20]	https://busco.ezlab.org
EDTA v1.9.4	Ou et al.[92]	https://github.com/oushujun/EDTA
HISAT2 v2.2.1	Kim et al.[93]	https://github.com/DaehwanKimLab/hisat2
StringTie v1.13	Pertea et al.[94]	https://ccb.jhu.edu/software/stringtie
AUGUSTUS v3.3.3	Stanke and Morgenstern[95]	https://github.com/Gaius-Augustus/Augustus
GeneMark-ET v4.68_lic	Lomsadze et al.[96]	http://exon.gatech.edu/GeneMark
BRAKER2 v2.1.5	Hoff et al.[97]	https://github.com/Gaius-Augustus/BRAKER
cd-hit-est v 4.8.1	Li and Godzik[98]	https://github.com/weizhongli/cdhit/
MAKER2 v2.31.11	Holt and Yandell[99]	https://www.yandell-lab.org/software/maker.html
InterProScan v5.53-87.0	Blum et al.[100]	http://www.ebi.ac.uk/interpro
Cactus v2.0.3	Armstrong et al.[101]	https://github.com/ComparativeGenomicsToolkit/cactus
IQ-TREE v2.0.6	Nguyen et al.[102]	http://www.iqtree.org
PAML v4.9	Yang[103]	http://abacus.gene.ucl.ac.uk/software/paml.html
GeneTribe v1.2.0	Chen et al.[104]	https://chenym1.github.io/genetribe
GENESPACE v0.9.3	Lovell et al.[105]	https://github.com/jtlovell/GENESPACE
GERP++	Davydov et al.[14]	http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html
R package topGO	Alexa et al.[106]	https://bioconductor.org/packages/topGO
BWA v0.7.17-r1188	Li and Durbin[107]	https://bio-bwa.sourceforge.net/
GATK v4.2.3.0	McKenna et al.[108]	https://gatk.broadinstitute.org
minimap2 v2.21-r1071	Li[109]	https://github.com/lh3/minimap2
BCFtools v1.13	Li[110]	https://github.com/samtools/bcftools
R package qgg v1.0	Rohde et al.[111]	https://github.com/psoerensen/qgg
R package regress	Clifford and McCullagh[112]	https://github.com/kbroman/regress

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Sanwen Huang (huangsanwen@caas.cn).

Materials availability

This study did not generate new unique plant materials.

Data and code availability

- Genome assemblies and annotations of the newly assembled Solanaceae genomes are available at https://ngdc.cncb.ac.cn/gwh with BioProject accession number PRJCA010759. All sequence data generated in this study have been deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) with BioProject accession number PRJNA839598.
- All codes were deposited at https://github.com/yywyaoyaowu/SolEvo_PotatoDele.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Plant growth conditions

The Solanaceae seeds were grown in the greenhouses of the Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen (22°36’N and 114°30’E), Guangdong province, China in the spring and summer of 2021.

METHOD DETAILS

Sample selection and sequencing

We selected a total of 100 accessions from 92 species, comprising five from Convolvulaceae[66–70] and 95 from Solanaceae including 61 *Solanum* species,[24,27,71–87,113–115] representing the major clades of the Solanaceae phylogeny (Table S1). We performed Pacific Biosciences (PacBio) high-fidelity (HiFi) sequencing of 38 Solanaceae accessions (32 species, of which 22 are *Solanum* species), comprising 30 diploids and eight polyploids, on the Sequel II platform, using the circular consensus sequencing (CCS) mode (Table S1). The CCS program (https://github.com/PacificBiosciences/ccs) was applied to generate HiFi reads for 38 genomes. Additionally, we constructed Hi-C sequencing libraries for 20 of these species (Table S1), for which the restriction enzyme *Mbo* I/*Hind* III was used to digest their genomic DNA, using a previously described library preparation protocol.[116] These libraries were subsequently sequenced on an Illumina HiSeq X Ten platform to generate paired-end reads of 150-bp length. We extracted total RNA from roots, young leaves, mature leaves, leaf buds, stems, flower buds, flowers, sepals, young fruits and mature fruits of 32 species to prepare RNA sequencing libraries (Table S1). The resulting 204 libraries were sequenced on the DNBSEQ-T7 system, representing most tissues in most samples (Table S1).

Genome assembly of the 38 Solanaceae accessions

GenomeScope2.0[88] was used to estimate genome heterozygosity. Hifiasm (v0.16.1-r375)[89] was deployed to assemble the genomes of the 38 sequenced accessions. To achieve a monoploid assembled content, the hifiasm parameter “-l” was tuned for different species (0 or 3), based on their estimated genome heterozygosity. We constructed pseudo-chromosomes of the 20 accessions with available Hi-C data. We then mapped Hi-C reads to the assembled contigs using juicer (v1.6)[90] with default parameters, and applied the 3d-dna pipeline[91] (v180922) to order and orient these contigs into chromosome-level scaffolds with parameters “-I 15000 -r 0”, followed by thorough manual curation. We employed Benchmarking Universal Single-Copy Orthologs (BUSCO, v5.2.2)[20] to assess the completeness in genic regions using the Solanales_odb10 database (for Solanaceae species).

Repetitive element annotation

For each species, we used the Extensive *de novo* TE Annotator (EDTA v1.9.4)[92] to identify transposable elements (TEs), generating non-redundant TE libraries that were used to mask repeats.

Prediction of protein-coding genes

To predict coding-gene models, we applied a uniform pipeline combining evidence from *ab initio* prediction, homology search and transcript expression. RNA-seq reads were mapped to the assembled genome using HISAT2 (v2.2.1)[93] with the “–dta” parameter, followed by genome-guided transcript assembly by StringTie (v1.13).[94] Hidden Markov models (HMM) of the two software packages used for *ab initio* gene prediction, AUGUSTUS (v3.3.3)[95] and GeneMark-ET (v4.68_lic),[96] were trained by BRAKER2 (v2.1.5)[97] utilizing the assembled transcripts as hints. The parameters set in BRAKER2 were “–nocleanup –softmasking”.

We combined human-curated, high-confidence plant protein sequences downloaded from the UniProt Swiss-Prot database (https://www.uniprot.org/downloads) with published amino-acid sequences of tomato, potato, eggplant, chili pepper and *Arabidopsis thaliana*[76,81,113–115,117]; we excluded potential redundance using cd-hit-est (v 4.8.1)[98] with default parameters, serving as the homologous proteins. We considered the assembled transcripts from StringTie as evidence for transcript expression. Putative gene structures inferred by AUGUSTUS (v3.3.3)[95] and GeneMark-ET (v4.68_lic)[96] were polished and synthesized by MAKER2 (v2.31.11)[99] to generate the final gene annotations. The longest transcript of each predicted gene model was considered as its representative. InterProScan (v5.53-87.0)[100] was used to predict protein functional domains using the parameters "-cli -iprlookup -tsv -appl Pfam".

Whole-genome alignment

To deploy whole-genome alignments of the 100 genomes (92 species), we first inferred a phylogeny based on mash distances from their genome assemblies, with repeats being soft-masked by mashtree (v1.2.0) (https://github.com/lskatz/mashtree) which is incorporated in the msa_pipeline (v1.0).[118] We then processed the phylogeny and the soft-masked genomes in Progressive Cactus (v2.0.3)[101] to obtain genome-wide alignments, choosing the reference genome of potato, *Solanum tuberosum* group Phureja DM1-3 516 R44 v6.1 (DM v6.1)[81,113] for downstream analysis.

Phylogenetic analyses

To reconstruct the phylogeny of the 100 accessions (five from Convolvulaceae and 95 from Solanaceae, see Table S2), the consensus tree topology was inferred by IQ-TREE (v2.0.6),[102] using the alignment of four-fold degenerate sites from the whole-genome multiple alignments. To obtain 'local' phylogenies, we split whole-genome alignment blocks into 1-Mb sliding windows with 200-Kb step size, followed by tree-topology inference for each window using IQ-TREE with the parameter "-m 012345". We randomly selected 500 local (i.e. window-based) trees for visualization, using an R script modified from https://zenodo.org/record/3401692#.YNrvJ6e76XQ.

The total branch length is the sum of the branch lengths of all the 100 accessions. The increase of the total branch length obtained in this study (4.05) is compared with that of the previously reported whole-genome phylogeny (0.71) of *Solanum* section *Petota* from Tang et al.[27]

Estimation of divergence times

We removed gap positions and sequences containing unknown nucleotides among the 100 accessions from the multiple alignment of four-fold degenerate sites. We then performed maximum-likelihood estimation of substitution rate per site using the baseml program in the PAML package (version 4.9),[103] based on the general time-reversible (GTR) nucleotide substitution model. Bayesian estimation of divergence times was deployed using the mcmctree program, manually setting the gamma prior for the overall substitution rate. We set the fossil calibration point to the stem node of the Berry clade of Solanaceae (51.2–53.2 MYA), following Wilf et al.[119] and De-Silva et al.[22]

Detection of genomic synteny

To assess gene synteny between each pair of the 100 accessions, we applied the MCScanX[120] algorithm incorporated in GeneTribe (v1.2.0).[104] The resulting syntenic gene pairs were converted into genome-wide syntenic blocks with their coordinates retained. A whole-genome synteny plot was generated using GENESPACE (v0.9.3).[105] We observed large differences in syntenic gene proportions between the previously released genome of *Nicotiana benthamiana*[121] and the genome of the same species assembled in this study (17% versus 78%; Table S2). Given that our genome of *N. benthamiana* was assembled using PacBio HiFi and Hi-C data, achieving significantly higher continuity than the previously reported one based on short-read technology (54 Mb versus 89 Kb in terms of contig N_{50} length), this difference in syntenic proportions should be due to our assembly's markedly improved continuity. This reasoning would also explain the low degrees of genomic synteny observed for other *Nicotiana* species (Table S2), with all of their genomes being built using short-read sequencing data.[75,83,84,87]

Identification of constrained sites

Utilizing the whole-genome alignment and the inferred phylogeny, we computed the degree of genomic (evolutionary) constraint in terms of GERP score of each of the nucleotides present in the 100-genome alignment with GERP++.[14,15] We defined sites with GERP scores 2–2.75, 2.75-3.5 and ≥ 3.5 as mildly, moderately and strongly evolutionarily constrained sites, respectively. GERP ≥ 2, GERP ≥ 2.75 and GERP ≥ 3.5 were our 'mild', 'moderate' and 'high' thresholds for constrained sites and deleterious variants respectively. The results based on our 'moderate' threshold were mainly reported in the main text.

We calculated the constrained proportion of coding genes (the number of constrained sites per CDS divided by the total length of that CDS in bp) for each of the 32,917 genes. Those genes with the top 1% of constrained proportions were used for GO enrichment analysis using the "topGO" R package[106] (https://bioconductor.org/packages/topGO).

Identification of deleterious variants in the diploid diversity panel

Variants (i.e. those with SNPs) within the diploid potato diversity panel at the constrained sites were considered as deleterious variants. The workflow for SNP calling and filtering was as follows: clean reads were mapped onto the *S. tuberosum* group Phureja DMv6.1 genome using BWA-mem software (v6.0.2)[107] with default parameters. The sam file was converted to bam and sorted by SAMtools (v0.7.17)[122] software. SAMtools was used to remove any duplicate reads. The variants were called by GATK (v.4.2.3.0) HaplotypeCaller,[108] and SNPs were further filtered using the following criteria: "QD<2.0| | FS>60.0| | MQ<40.0| | SOR>3.0| | MQRankSum<−12.5| | ReadPosRankSum<−8.0"; then the bi-allelic SNPs were filtered by VCFtools (v0.1.16)[123] software using the following criteria: "–minDP 4 –maxDP 100 –minGQ 10 –minQ 30 –max-missing 0.5 –min-alleles 2 –max-alleles 2".

The main manuscript presents the results of moderately (GERP score ≥ 2.75 and <3.5) and highly deleterious variants (GERP score ≥ 3.5) using the moderate threshold (GERP score ≥ 2.75, representing the top 0.6% of the genome-wide distribution across all 58,597,787 SNPs). A total of 97% of these deleterious variants for which minor alleles in the diversity panel are also non-major alleles across the 100 Solanaceae genomes, suggesting that evolutionarily derived mutations generally coincide with the minor allele variants within the potato diversity panel. The results using the mild (GERP ≥ 2) and high (GERP ≥ 3.5) thresholds are presented in the supplemental information (Figures S3D and S3F; Table S3).

Of the deleterious variants, the minor alleles within the diploid potato diversity panel and within the 100 Solanaceae genomes were assumed to represent deleterious alleles.

Estimation of deleterious mutation burdens

By summing the GERP scores over all deleterious alleles present in a given individual, we computed the so-called 'deleterious mutation burden' for each individual of the diploid potato diversity panel. The deleterious burden in homozygous state (homozygous burden, B_Homo, recessive burden) for each accession was obtained by summing the GERP scores of the genome-wide inferred deleterious mutations encountered in homozygous state. Likewise, the deleterious burden in heterozygous state (heterozygous burden, B_Heter) for each accession was calculated by summing the GERP scores for the genome-wide inferred deleterious mutations found in heterozygous state. The genetic deleterious burden (genetic burden, B_Genetic; numerically equals additive burden) for each individual, representing the deleterious burden potentially transmitted to its offspring and related to the offsprings' fitness, was calculated by summing the homozygous burden and the heterozygous burden multiplied by 0.5.[41–43] The factor 0.5 represents the 50% probability that a heterozygous deleterious mutation is transmitted to any given offspring. The expressed deleterious burden (expressed burden, B_Expressed), representing all exposed deleterious effects and related to the bearer's fitness, is the sum of the homozygous burden and the heterozygous burden multiplied by the estimated average dominance coefficient, *h*.[41–43] Here, the genome-wide average *h* is 0.1, which is estimated from a linear-mixed model by the method of moments and corroborated by a maximum-likelihood method (see details below on the linear mixed model; Figure S5D). This value is congruent with previous experimental studies.[64,65]

If all heterozygous deleterious mutations were recessive (*h* = 0), the expressed burden equals the homozygous deleterious burden (recessive burden); if all heterozygous deleterious mutations were additive (*h* = 0.5), the expressed burden equals the additive burden (numerically equals genetic burden, see the equations below). The masked deleterious burden rests on the heterozygous deleterious burden, contingent on the average *h* values and thus the degree of mutations' recessivity. The masked deleterious burden and expressed burden jointly constitute the genetic burden. Correlation coefficients between deleterious burdens were estimated by fitting a linear model using the "lm()" function, and outliers were detected and removed by boxplot.stats in R.

$$\mathrm{B\_}Homo = \sum_{\mathrm{i=1}}^{\mathrm{L(homo)}} \mathrm{GERP_i}$$

$$\mathrm{B\_}Heter = \sum_{\mathrm{j=1}}^{\mathrm{L(heter)}} \mathrm{GERP_j}$$

$$B\_Genetic(individual) = B\_Homo + B\_Heter * 0.5$$

$$B\_Expressed\ (individual) = \sum_{\mathrm{i=1}}^{\mathrm{L(homo)}} \mathrm{GERP_i} + \sum_{j=1}^{\mathrm{L(heter)}} \mathrm{GERP_j} * h_j$$

$$= B\_Homo + B\_Heter * h$$

$$B\_Mask(individual) = B\_Heter * (0.5 - h)$$

Note: h_j is the dominance coefficient for heterozygous deleterious site j; h is the genome-wide average dominance coefficient, which was estimated as 0.1 in the linear fixed model by the method of moments and corroborated by maximum likelihood method computed at the moderate threshold (GERP score $\geq$ 2.75; see details below on the linear mixed model).

Identification of deleterious mutations in tetraploid potatoes

To identify deleterious mutations in tetraploid potatoes, we downloaded DNA-seq data of 43 cultivated tetraploid potato accessions from previous studies.[124–128] We used the same workflow to perform SNP calling and filtering as for our diploid dataset and then merged the SNP set with the SNPs from the diploid diversity panel. The homozygous SNPs were defined when only one allele was supported with reads, and heterozygous SNPs were defined when both reference allele and alternative allele were supported with reads. The homozygous and heterozygous deleterious mutations for each individual were calculated by using the same pipeline as for the diploid diversity panel.

Determination of heterozygous genomic regions of RH and RH10-15

RH10-15 is the only progeny of the F_4 inbred population of RH89-039-16 (RH) with a proportion of genomic regions still being heterozygous. To determine the extent of these regions, we mapped HiFi reads of RH10-15 and RH[27] to the potato *S. tuberosum* group Phureja DM v6.1 reference genome using minimap2 (v2.21-r1071)[109] with the parameter "-ax map-hifi". We then identified SNPs using the subcommands "mpileup -Ou" and "call -m" embedded in BCFtools (v1.13),[110] followed by applying a set of filtering criteria: 5 $\leq$ read depth (DP) $\leq$ 200, mapping quality (MQ) $\geq$ 40, variant quality (QUAL) $\geq$ 30 and Phred-scaled *p*-value using Fisher's exact test to detect strand bias (FS) < 60. Heterozygous SNPs were defined as variants with the SNP-index (i.e. number of reads supporting the alternate allele/total number of reads mapped at this position) ranging from 0.3 to 0.7. We calculated the number of heterozygous SNPs using 50-Kb windows with 5-Kb step size, and windows with heterozygous proportion > 2% were merged and regarded as heterozygous genomic regions. We quantified heterozygous proportion as the number of heterozygous SNPs divided by the total number of aligned sites. We applied the same approach to the RH genome to delineate its heterozygous genomic regions.

Genomic prediction

To estimate the improvement of genomic prediction by phylogenomic information, we calculated the deleterious burden in an F_2 population consisting of n = 1,064 individuals with bin-genotype data and several agronomic traits available. The grandparents of this population are the two inbred lines A6-26 and E4-63. Because the genotypes of this population use E4-63 as the reference genome, we mapped Illumina reads of A6-26 to the genome of E4-63 with BWA (0.7.17-r1188)[107] and called SNPs via BCFtools (v1.13),[110] followed by applying a set of filtering criteria: 3 $\leq$ DP $\leq$ 100, MQ $\geq$ 40, QUAL $\geq$ 30 and FS < 60. We regard SNPs at sites with GERP scores $\geq$ 2.75 as harboring potentially deleterious variants, considering each of the non-major alleles in the 100-Solanaceae genome alignment as potentially deleterious alleles. The deleterious mutation burden of each of the 2,603 bins from the F_2 population (the so-called 'bin burden') was calculated by summing the GERP scores of all inferred deleterious alleles within each bin j: $\mathbf{b}_j = \sum_{k=1\{SNP\ k\ in\ bin\ j\}}^{L} \delta_k \mathbf{m}_k$, where δ_k is the GERP score for SNP k and $\mathbf{m}_k$ indicates the presence of the deleterious allele at SNP k across individuals ($\mathbf{m}_k$=1 for deleterious allele, otherwise $\mathbf{m}_k$=0). To characterize genome-wide deleterious mutation burdens for each individual, we summed the deleterious mutation burdens over bins, distinguishing homozygous and heterozygous states: the genome-wide homozygous burden ($\mathbf{b}_{hom}$) and heterozygous burden ($\mathbf{b}_{het}$) is contributed solely by deleterious alleles in the homozygous and heterozygous states, respectively.

In the 'baseline' genomic prediction model, we calculated the additive genomic relationship matrix (GRM) by bin-genotype:

$$\mathbf{y} = \mu + \mathbf{u} + \mathbf{e}$$

$$\mathbf{u} \sim N\left(\mathbf{0}, \mathbf{G}\sigma_u^2\right)$$

$$\mathbf{G} = \frac{\mathbf{XX}^{\mathrm{T}}}{\sum_{j=1}^{2603} 2p_j\left(1 - p_j\right)}$$

where $\mathbf{y}$ is the vector of phenotypic scores, μ is the intercept (grand mean), $\mathbf{u}$ is the vector of additive genomic effects, $\mathbf{e}$ is the vector of residuals, $\mathbf{X}$ is the 1064 × 2603 matrix of reference allele counts (0, 1, or 2) at each bin, p_j is the allele frequency at bin j (j = 1,..., 2603), and $\mathbf{G}$ is the GRM by bin genotypes.

In the deleterious mutation model, we estimated the effects of genome-wide homozygous burden and heterozygous burden as fixed effects and the additive genomic effects as random effects in a linear mixed model[112]:

$$\mathbf{y} = \mu + \alpha_{hom}\mathbf{b}_{hom} + \alpha_{het}\mathbf{b}_{het} + \mathbf{u} + \mathbf{e}$$

$$\mathbf{u} \sim N\left(\mathbf{0}, \mathbf{G}\sigma_u^2\right)$$

where $\mathbf{b}_{hom}$ and $\mathbf{b}_{het}$ are the genome-wide homozygous and heterozygous burdens (as described above), with fixed effects α_{hom} and α_{het}, respectively. Genomic prediction models were fitted by the *greml* function in the *qgg* package in R (v1.0)[111] and were evaluated by five-fold cross-validation replicated 20 times. Genomic prediction accuracy was then computed by the average of squared Pearson correlation coefficients (r^2) between predicted and observed phenotypic scores.

To test the hypothesis that actual GERP scores are more useful than random SNP scores in genomic prediction, we generated random SNP scores by permuting GERP scores among genomic windows of 500,000 SNPs to account for linkage disequilibrium among contiguous polymorphisms. For each permutation, we used the random SNP scores to generate the genome-wide burden indices $\mathbf{b}_{hom}$ and $\mathbf{b}_{het}$. Then, we estimated their effect in the weighted deleterious mutation model and computed its prediction accuracy in cross-validation, as described above with the same training datasets. The prediction accuracy achieved by the actual deleterious burden was deemed statistically significant if it was higher than the top 5% of prediction accuracy achieved by the random SNP scores from 100 permutations. These significance thresholds are all lower than the observed prediction accuracy ($P < 0.05$).

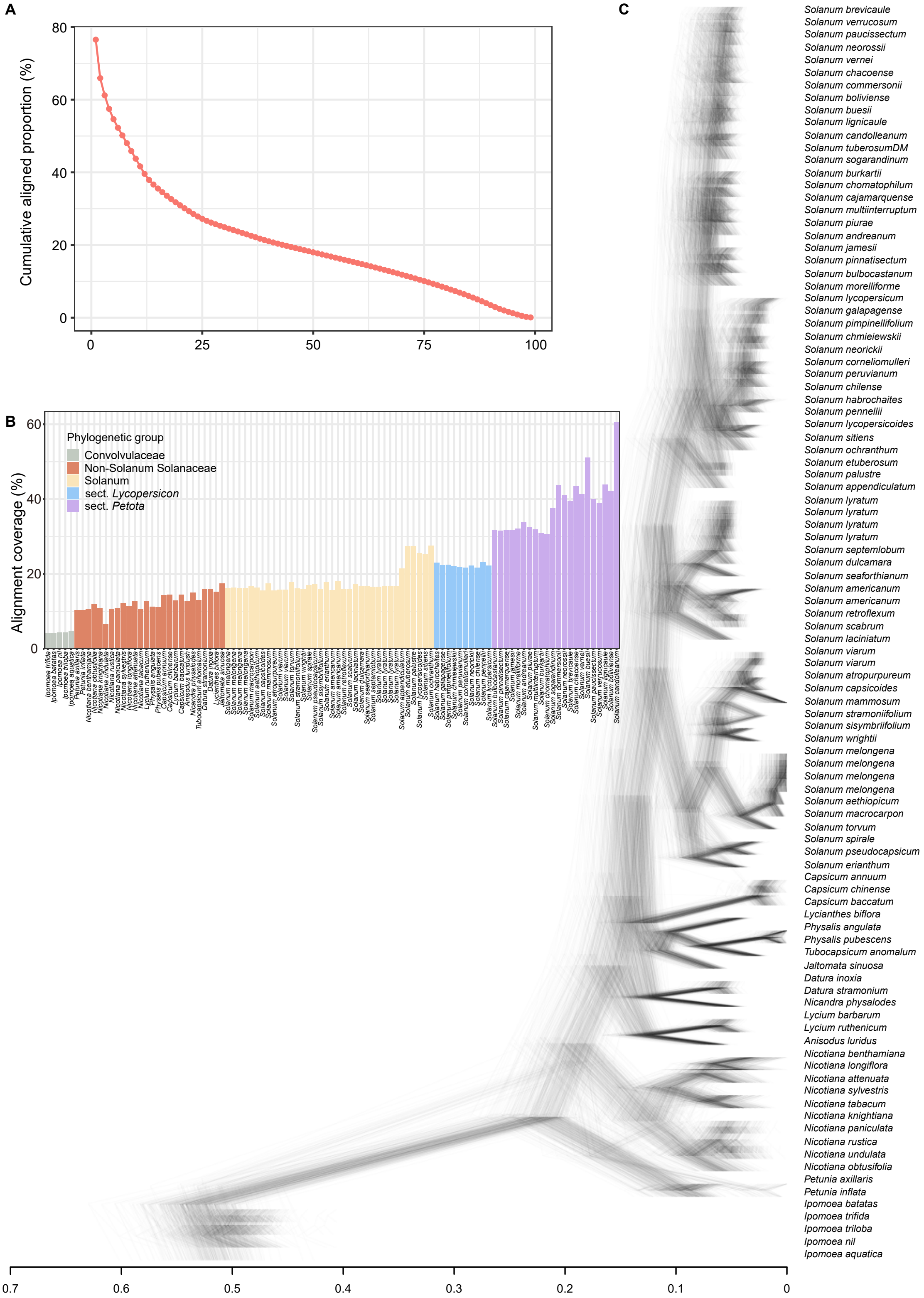
Notwithstanding the difficulties to obtain accurate estimates of the genome-wide average dominance coefficient (h), it is an important component for estimating the effect of heterozygous burden and expressed burden (the fitness burden). Here, we used the method of moments[129] to estimate the genome-wide average dominance coefficient (h) of deleterious mutations by estimating α_{het} and α_{hom}: $\widehat{h} = \frac{\widehat{\alpha_{het}}}{\widehat{\alpha_{hom}}}$ in the weighted deleterious mutation model for yield in the F_2 population. h was estimated to be 0.1 at the moderate threshold (GERP score $\geq$ 2.75). We also inferred h using the maximum likelihood method with the *regress* package[112] in R: we applied the expressed burden as the fixed effect in the weighted deleterious mutation model by using a series of h values ranging from 0 to 0.5 with a step size of 0.005, thus obtaining the optimized h by maximum likelihood. This yielded results similar to those estimated by the method of moments ($h = 0.115$ at the moderate threshold, Figure S5D), and this value is congruent with previous experimental studies,[64,65] further corroborating this roughly estimated h. Moreover, h tends to decrease as the deleterious threshold (GERP score) increases (Figure S5D), consistent with classical results.[130,131]

QUANTIFICATION AND STATISTICAL ANALYSIS

The statistical details of analysis applied in this paper are provided alongside in the results section and Figure Legends. Statistical analyses were performed in R 4.1.2.

CellPress
OPEN ACCESS

Supplemental figures



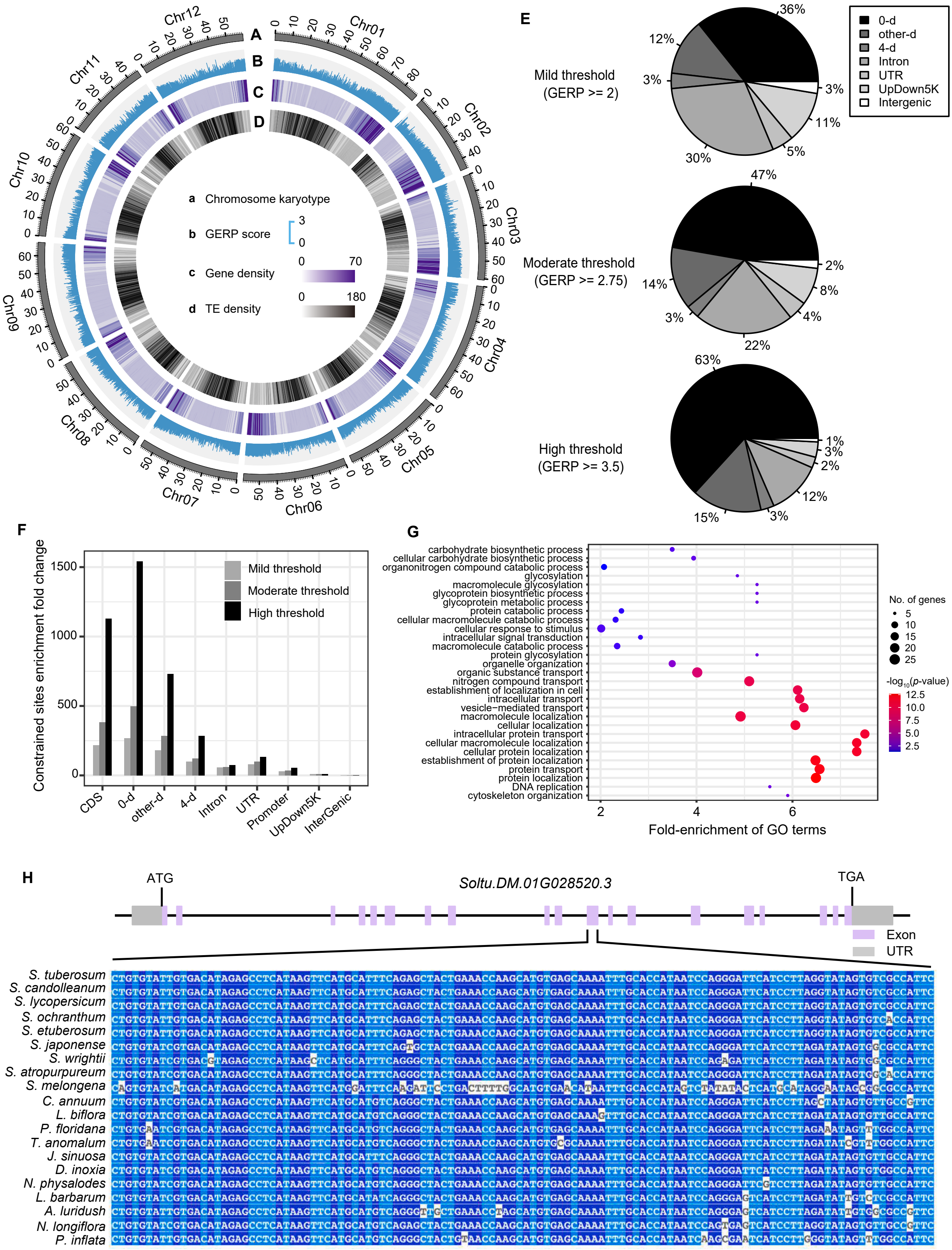
(legend on next page)

Figure S1. Summary of the 100-way whole-genome multiple alignment and phylogenetic trees of 500 randomly chosen genomic windows from 95 Solanaceae and five Convolvulaceae outgroup genomes, related to Figure 1

(A) Cumulative alignment proportion. The x axis counts the number of genomes, the y axis the alignment proportions with at least the given number of species aligned.

(B) Alignment coverage of 100 genomes from five phylogenetic groups. Section *Petota*, species from *Solanum* section *Petota*; section *Lycopersicon*, species from *Solanum* section *Lycopersicon*; *Solanum*, species from *Solanum* that do not belong to sections *Petota* or *Lycopersicon*.

(C) Trees for genomic windows are depicted by gray lines. The 1-Mb windows with 200-kb step size were randomly selected across the entire genome. The x axis indicates branch length.



(legend on next page)

Figure S2. Genome-wide distribution of evolutionary constraint in potato, related to Figure 2

(A) Karyotype of the 12 potato chromosomes.

(B) Histograms indicate the distribution of GERP scores by splitting the genome into 73,135 non-overlapping 10-kb windows.

(C) Heatmaps illustrate gene density, represented as the number of genes per 500 kb.

(D) Heatmaps illustrate transposable element (TE) density in terms of number of TEs (size >1 kb) per 500 kb.

(E) Relative composition of evolutionarily constrained sites among different genomic regions for three different GERP-score thresholds.

(F) Enrichment of evolutionarily constrained sites in functional regions compared with intergenic regions for different GERP-score thresholds. UpDown5K, 5-kb upstream and downstream of genes. UTR, untranslated region. Promoter, 1-kb upstream of CDS.

(G) Gene ontology enrichments (biological processes) of the top 1% of constrained genes in the potato genome.

(H) A detailed alignment of the 11th exon of *Soltu.DM.01G028520.3*, an evolutionarily constrained gene.

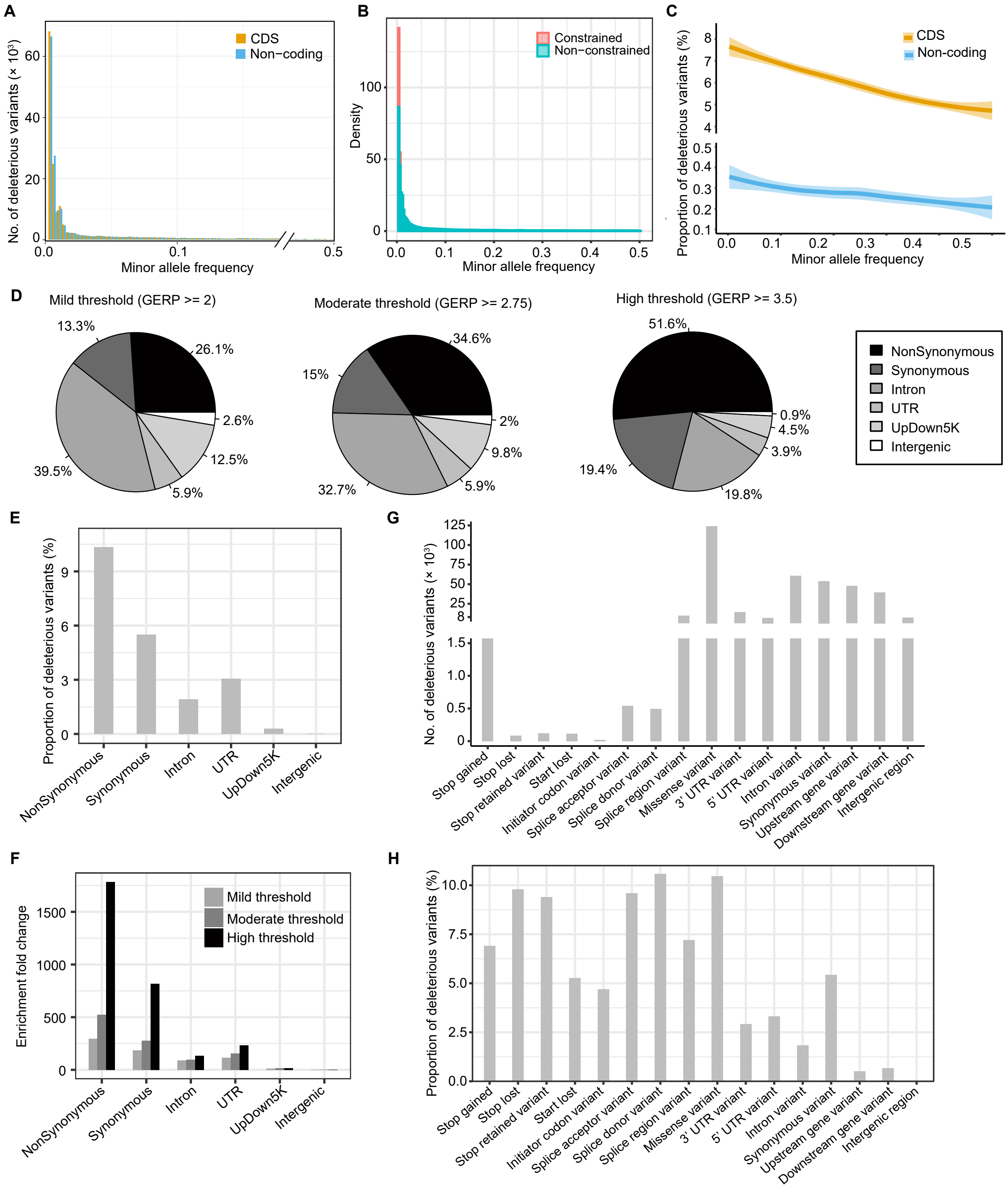


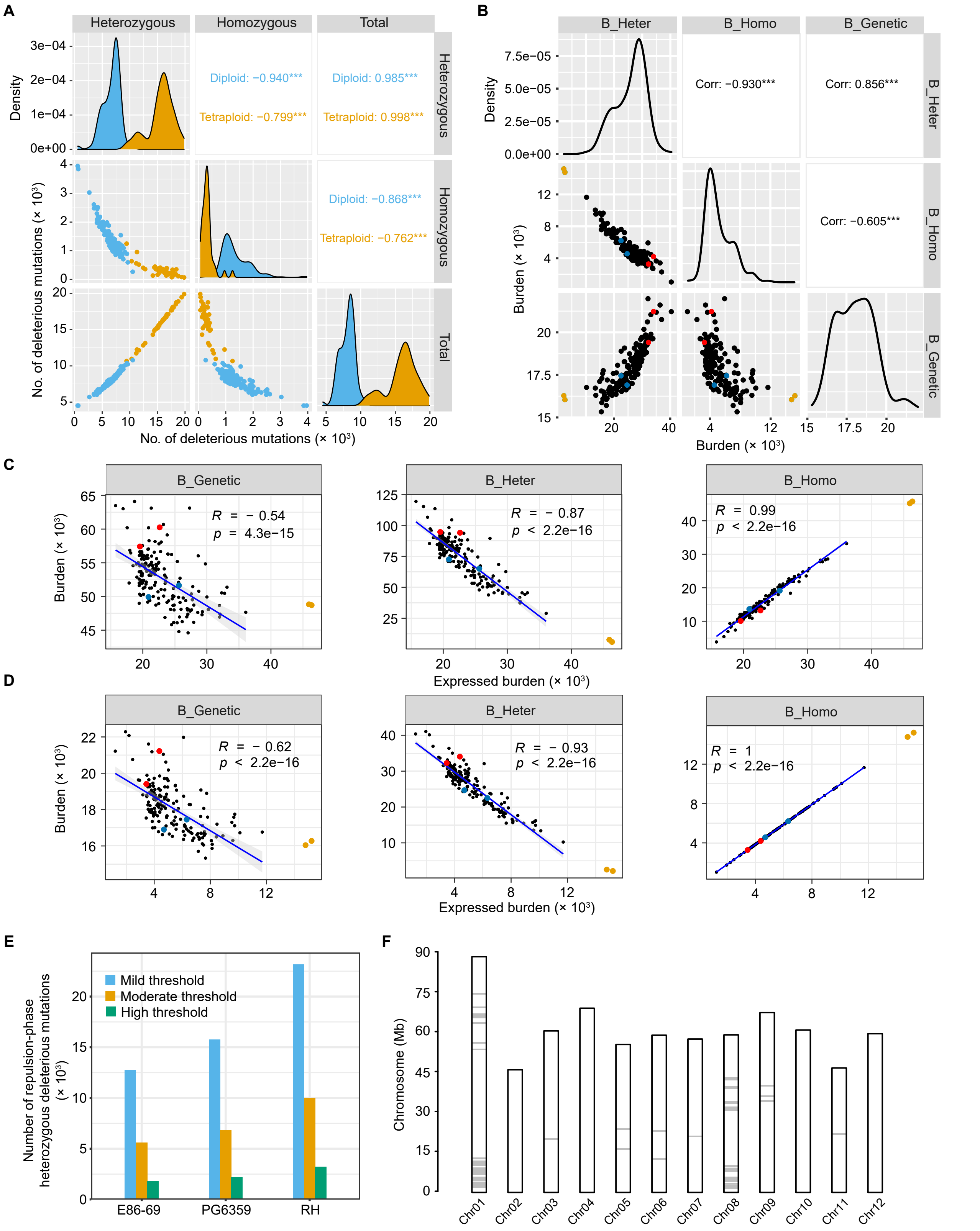
Figure S3. Deleterious variants within the diploid potato diversity panel, related to Figure 3

(A) Distributions of minor allele frequency of inferred deleterious variants in coding (CDS) and non-coding regions in the diploid potato diversity panel.

(B) Density plot of allele-frequency spectrum of SNPs at constrained and non-constrained sites in the diploid potato diversity panel.

(legend continued on next page)

(C) Decrease in the proportion of deleterious variants per SNP with increasing minor allele frequency for coding (CDS) and non-coding regions in the diploid potato diversity panel.
(D) Partitioning of all inferred deleterious variants into those at nonsynonymous sites, synonymous sites, and variants in non-coding genomic regions. UpDown5K, 5-kb upstream and downstream of genes. UTR, untranslated regions for different thresholds.
(E) The proportion of deleterious variants per SNP, calculated as the number of deleterious mutations divided by the number of all SNPs, within nonsynonymous and synonymous sites and variants in non-coding genomic regions, respectively.
(F) The enrichment fold change of proportion of deleterious variants per SNP in nonsynonymous sites, synonymous sites, and variants in non-coding genomic regions compared with that in intergenic regions for different thresholds. UpDown5K, 5-kb upstream and downstream of genes. UTR, untranslated region.
(G) The numbers of different types of deleterious variants in the diploid potato diversity panel. Note the interrupted y axis scale.
(H) Proportion of deleterious variants per SNP in different types of variants.



(legend on next page)

Figure S4. The deleterious mutation burden in potato, related to Figure 4

(A) Number of homozygous, heterozygous and total deleterious mutations among tetraploid and diploid potatoes (GERP score ≥2.75). Correlation coefficients (*R*) among deleterious mutations are also shown. ***p < 0.001 in Pearson correlation tests. Tetraploid potato lines are highlighted with orange dots, and diploid potato lines are denoted by blue dots.

(B) Correlations among genetic burden (B_Genetic), homozygous burden (B_Homo) and heterozygous burden (B_Heter) at the highly deleterious threshold (GERP score ≥3.5). Correlation coefficients (*R*) among different burdens are also shown. ***p < 0.001 in Pearson correlation tests. Accessions RH and C10-20 are highlighted with red dots, accessions PG6359 and E86-69 are denoted by blue dots, and the two inbred lines A6-26 and E4-63 are marked by orange dots.

(C and D) Correlations among expressed burden (B_Expressed; x axis) and genetic burden (B_Genetic), homozygous burden (B_Homo), and heterozygous burden (B_Heter) at the moderate threshold (C, GERP score ≥ 2.75) and the high threshold (D, GERP score ≥ 3.5), respectively. Correlation coefficients (*R*) among different burdens are also shown.

(E) Inferred numbers of repulsion-phase heterozygous deleterious mutations in the genomes of RH, PG6359 and E86-69.

(F) Genomic coordinates of heterozygous genomic regions in RH10-15. Regions in gray shade denote heterozygous genomic regions, using the coordinate system of the potato reference genome *S. tuberosum* group Phureja DM v6.1.

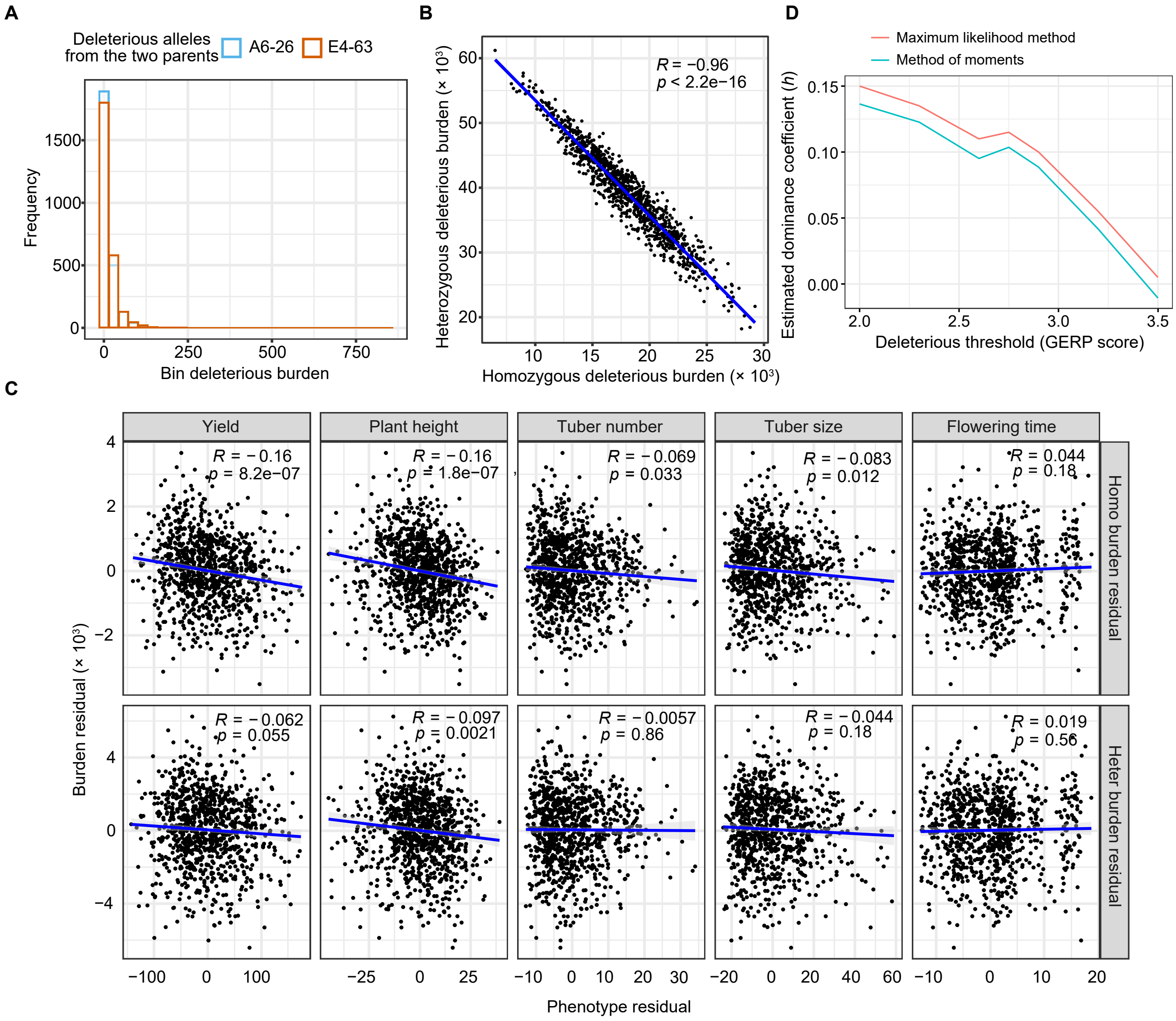


Figure S5. Distribution of deleterious mutation burden in the F_2 population, related to Figure 5

(A) Distribution of deleterious burden of each bin. A6-26 indicates the deleterious alleles contributed by parent A6-26, and E4-63 shows the deleterious alleles contributed by the other parent, E4-63 (moderate threshold).

(B) Correlation between the burdens of homozygous (x axis) and heterozygous (y axis) deleterious mutations among the F_2 individuals (moderate threshold). Pearson's correlation coefficient (R) and the corresponding p value computed with Pearson correlation test are indicated.

(C) Partial correlations between the five phenotypes and the per-individual burdens of homozygous (upper panel) and heterozygous (lower panel) deleterious mutations among F_2 individuals (moderate threshold). "Homo burden residual": the residual of homozygous burden after fitting heterozygous burden. "Heter burden residual": the residual of heterozygous burden after fitting homozygous burden. Phenotype residual: the residual of phenotype after fitting the heterozygous and homozygous burdens, respectively. Pearson's correlation coefficients (R) are indicated; p values were computed with Pearson correlation tests. Gray shadows represent the 95% confidence intervals, estimated by fitting a linear model using the lm() function in R.

(D) The genome-wide average dominance coefficient (h) estimated by the method of moments (green) and maximum-likelihood method (red) based on a range of deleterious thresholds.