LETTERS
https://doi.org/10.1038/s41588-020-0699-x

nature genetics

OPEN

Haplotype-resolved genome analyses of a heterozygous diploid potato

Qian Zhou[1,2,8], Dié Tang[1,8], Wu Huang[3], Zhongmin Yang[4], Yu Zhang[1], John P. Hamilton[5], Richard G. F. Visser[6], Christian W. B. Bachem[6], C. Robin Buell[5], Zhonghua Zhang[7,3], Chunzhi Zhang[1] and Sanwen Huang[1,3] ✉

Potato (*Solanum tuberosum* L.) is the most important tuber crop worldwide. Efforts are underway to transform the crop from a clonally propagated tetraploid into a seed-propagated, inbred-line-based hybrid, but this process requires a better understanding of potato genome. Here, we report the 1.67-Gb haplotype-resolved assembly of a diploid potato, RH89-039-16, using a combination of multiple sequencing strategies, including circular consensus sequencing. Comparison of the two haplotypes revealed ~2.1% intragenomic diversity, including 22,134 predicted deleterious mutations in 10,642 annotated genes. In 20,583 pairs of allelic genes, 16.6% and 30.8% exhibited differential expression and methylation between alleles, respectively. Deleterious mutations and differentially expressed alleles were dispersed throughout both haplotypes, complicating strategies to eradicate deleterious alleles or stack beneficial alleles via meiotic recombination. This study offers a holistic view of the genome organization of a clonally propagated diploid species and provides insights into technological evolution in resolving complex genomes.

Tetrasomic inheritance and clonal propagation via tubers are two structural challenges in *S. tuberosum* L. breeding and propagation. Genetic analyses in tetraploids are very complicated and thus genetic gains in potato breeding are limited. The widespread use of century-old varieties, such as Russet Burbank (a somatic mutant bred from a cultivar released in the 1870s in the United States) and Bintje (bred in 1904 in the Netherlands)[1], indicates that there has been little progress in developing key traits, such as yield, quality and disease resistance in modern tetraploids.

To accelerate genetic improvement in potato, several projects have been initiated to redomesticate potato from a tuber-propagated, tetraploid crop into a seed-propagated, inbred-line-based diploid crop[2–5]. To facilitate inbred line development, an improved understanding of the genome landscape of potato clones is required. While genome heterozygosity in diploid potato has been surveyed[6–9], these efforts were limited to bacterial artificial chromosome (BAC) clones and short-read sequences and lacked a genome-wide assessment of haplotype diversity.

Despite recent advances in genome assembly[10,11], construction of a haplotype-resolved genome for highly heterozygous species remains a challenge[12]. Current phasing strategies rely on the alignment of sequenced reads to a reference genome to infer regional haplotypes[13–17]; such efforts are limited by the continuity of an available reference assembly. Koren et al. have developed an alternative approach, trio binning, that can recover both parental haplotypes from an F_1 individual by partitioning parental unique reads before assembly[18], in which case parental information is required. Recently, high-throughput/resolution chromosome conformation capture (Hi-C) technology has helped to provide allele-resolved assemblies[19,20].

The heterozygous diploid potato *S. tuberosum* group Tuberosum RH89-039-16 ($2n=2x=24$, hereafter referred as to RH; Supplementary Fig. 1) has a pedigree from dihaploidized tetraploid commercial varieties, such as Katahdin, Chippewa and Primura. RH was partially assembled by the Potato Genome Sequencing Consortium (PGSC) in 2011 (refs. [6,21]). To resolve the RH genome at the haplotype level, we sequenced it using Illumina whole-genome sequencing (WGS), 10x Genomics (10xG) linked-read sequencing, Oxford Nanopore Technologies (ONT) and Hi-C technology (Supplementary Table 1). However, our attempts to de novo assemble the two haplotypes of RH using ONT reads, and scaffolding using Hi-C reads, were unsuccessful (Supplementary Fig. 2 and Supplementary Table 2). Thus, we developed an integrated strategy to generate a haplotype-resolved assembly (Fig. 1a,b and Supplementary Fig. 3). First, the diploid genome was assembled into scaffolds using Illumina reads (WGS and 10xG data; Supplementary Table 3). Second, an RH selfing population was sequenced to provide genetic information to phase the assembled fragments. Through the genetic groupings, we assigned the scaffolds into 24 linkage groups, corresponding to the 12 chromosome pairs of RH (Supplementary Figs. 4–6). Last, for each linkage group, the ONT and 10xG reads were retrieved and reassembled to generate an improved scaffold assembly. After polishing[22,23], the hybrid assembly yielded the genome draft version 1.0 (RHgv1) with a total length of 1.69 Gb and a scaffold N50 length of 920 kb.

Recently, accurate circular consensus sequencing (CCS) has provided impressive results on assembly and variant detection[24], showing its potential in resolving complex genome regions. Here, we generated 29 Gb of CCS data and assembled them using CANU[25], resulting in 1.53 Gb unitigs (contigs, split at alternate paths in the assembly graph) with an N50 size of 2.19 Mb (Supplementary Table 4).

[1]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Area, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. [2]Peng Cheng Laboratory, Shenzhen, China. [3]Key Laboratory of Biology and Genetic Improvement of Horticultural Crops of the Ministry of Agriculture, Sino-Dutch Joint Laboratory of Horticultural Genomics, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China. [4]College of Horticulture, Northwest Agriculture and Forest University, Yangling, China. [5]Department of Plant Biology, Michigan State University, East Lansing, MI, USA. [6]Plant Breeding, Wageningen University and Research, Wageningen, the Netherlands. [7]College of Horticulture, Qingdao Agricultural University, Qingdao, China. [8]These authors contributed equally: Qian Zhou, Dié Tang. ✉e-mail: huangsanwen@caas.cn

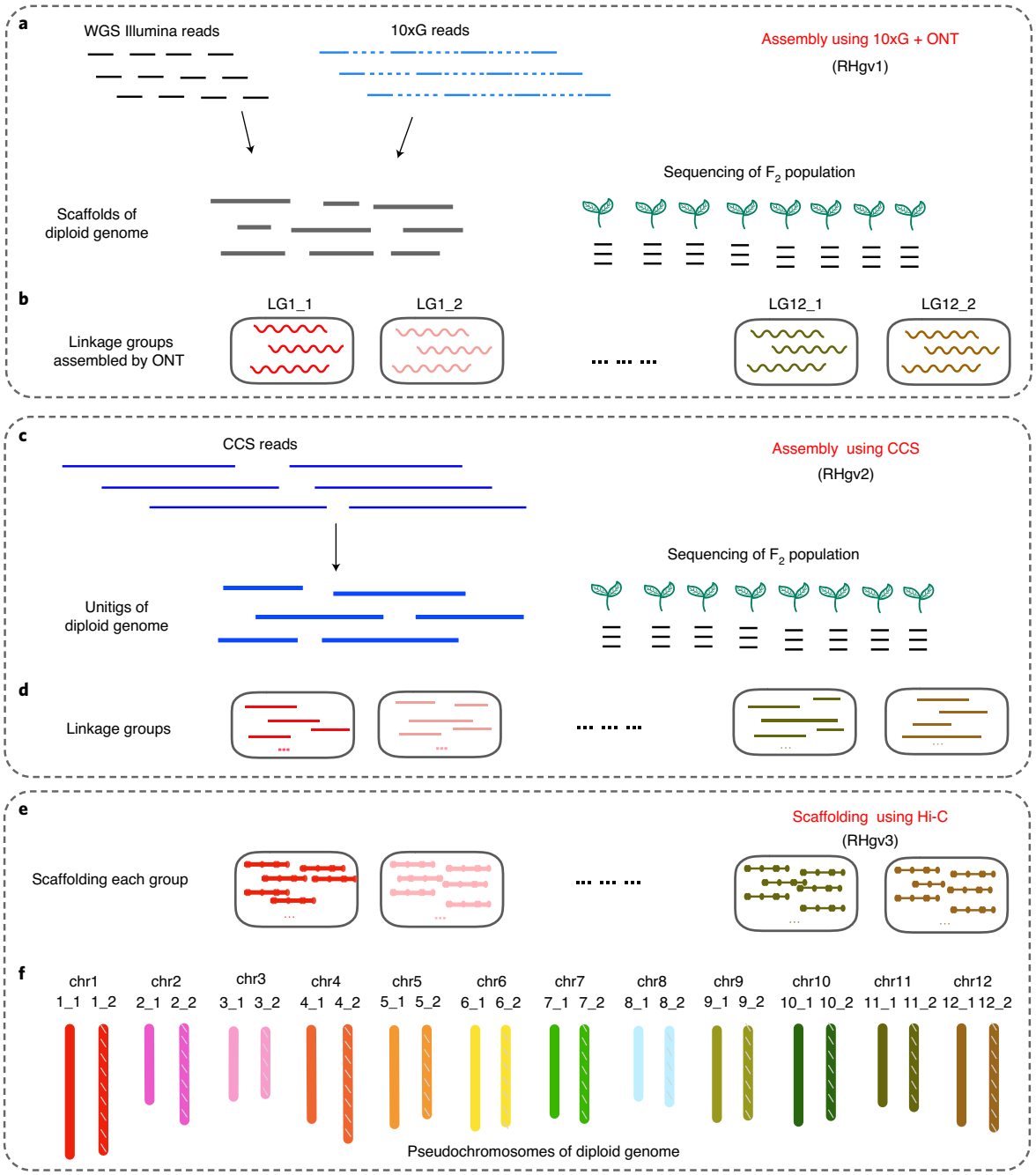


Fig. 1 | Hybrid de novo assembly and phasing of the diploid potato genome. a,**b**, The genome draft RHgv1 was assembled from WGS Illumina reads and 10xG linked reads, and the derived scaffolds were assigned into 24 haplotype-specific groups through genetic mapping based on a sequenced F_2 population. The 24 groups represent chromosomes of the diploid potato ($2n = 24$). ONT reads were aligned to each linkage group and assembled to improve scaffold contiguity. **c**,**d**, A second genome sequence, RHgv2, was assembled from CCS reads. Similarly, unitigs were assigned into 24 groups through genetic mapping. **e**,**f**, The two assemblies were merged to generate a more comprehensive genome, RHgv3. Hi-C data were used to scaffold the sequences of each group into pseudochromosomes.

Assisted by the RH selfing population, 1.31 Gb of unitigs were assigned into 24 groups, termed version 2.0 (RHgv2; Fig. 1c,d).

We assessed the accuracy of RHgv1 and RHgv2 using previously generated paired BAC-ends (BEs) and BAC clones[6]. In total, 95% and 99% of 54,902 BEs support the sequence correctness of RHgv1 and RHgv2, respectively (Supplementary Table 5). In a total of 184 BACs previously assembled and ordered, 126 and 169 BACs mapped to a single fragment on RHgv1 and RHgv2, respectively, with 113 and 152 BACs showing perfect collinearity with the scaffolds or unitigs (Supplementary Data 1 and 2). Notably, there were 10 BACs showing structural disagreements with both RHgv1 and RHgv2, while the latter two assemblies shared consistent structure, indicating potential errors in previous BAC assembly. The evaluation of base-level accuracy assessed by aligned BAC sequences was 99.127% and 99.936% for RHgv1 and RHgv2, respectively. Taken together, RHgv2 outperforms RHgv1 on both sequence continuity and accuracy.

To ensure the completeness of the final genome assembly, RHgv1 and RHgv2 were combined, generating a new 1.67 Gb assembly with

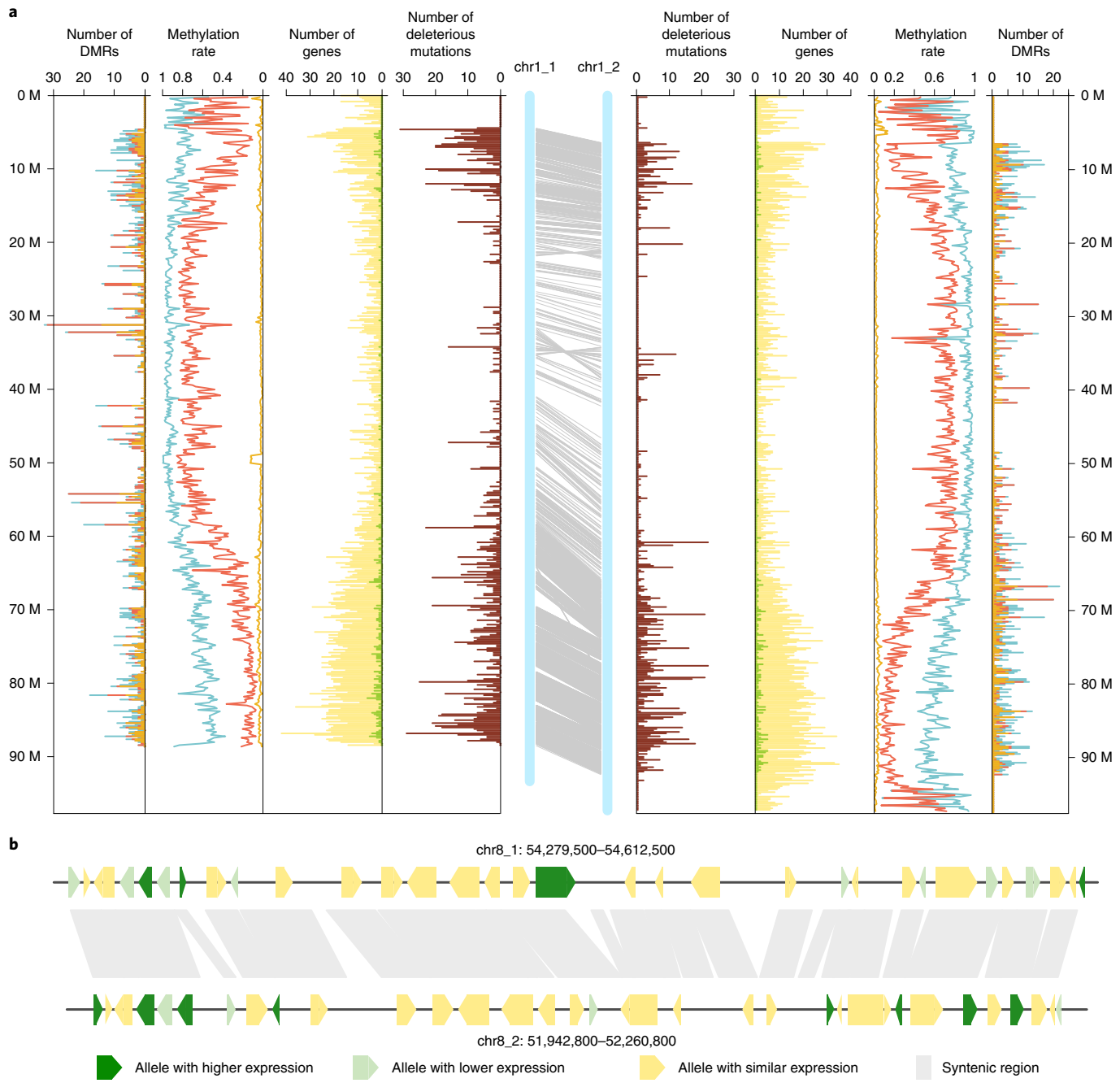


Fig. 2 | Haplotype divergence in a diploid potato genome. a, The central blue bars represent the two haplotypes of chromosome 1. The gray lines indicate paired allelic genes. Distribution of deleterious or dysfunctional mutations (brown), annotated genes (yellow), preferentially expressed alleles (green), methylation level of three contexts and differentially methylated regions are arranged symmetrically for each haplotype. Methylation level and the number of DMRs of methylated sites in CG (light blue), CHG (red) and CHH (orange) contexts are indicated by cumulative column chart. The number of DMRs on one haplotype only involves the DMRs with hypermethylation. All numbers were determined in 200-kb windows. **b**, Zoomed-in view of a syntenic block showing the mosaic pattern of preferentially expressed alleles (dark green) and alleles with lower expression (light green) on the two haplotypes.

an N50 length of 1.74 Mb (RHgv3; Supplementary Table 5). After grouping by genetic map and anchoring by the Hi-C data, 1.62 Gb of sequence constituted 24 pseudochromosomes, exceeding current assemblies of potato genomes[6,7,9] (Fig. 1e,f, Supplementary Fig. 7, Supplementary Tables 6–9 and Supplementary Data 3). Hereafter, all analyses were performed on RHgv3.

The phasing quality of the haplotype-resolved assembly was assessed on pseudochromosomes of RHgv3. The BEs were realigned to chromosomes and 46,058 (95.1%) of 48,410 aligned BEs were in same phase. A total of 1,639 BACs with unordered contigs were used to assess haplotype partitioning. Among 1,624 BACs that mapped with at least 60% of BAC length, 1,573 BACs (96.8%) aligned to a single phase of the RH assembly. Previously, Boer et al. reported a phased assembly of RH chromosome 5 based on BAC-by-BAC sequencing[21]. The alignment of 26.7-Mb haplotype RH{0} BAC minimal tiling paths (MTPs) and 25.0-Mb haplotype RH{1} MTPs with the pseudochromosome chr5_1 and chr5_2 showed that 26.0-Mb (97%) RH{0} MTPs have the best hit on chromosome 5_1

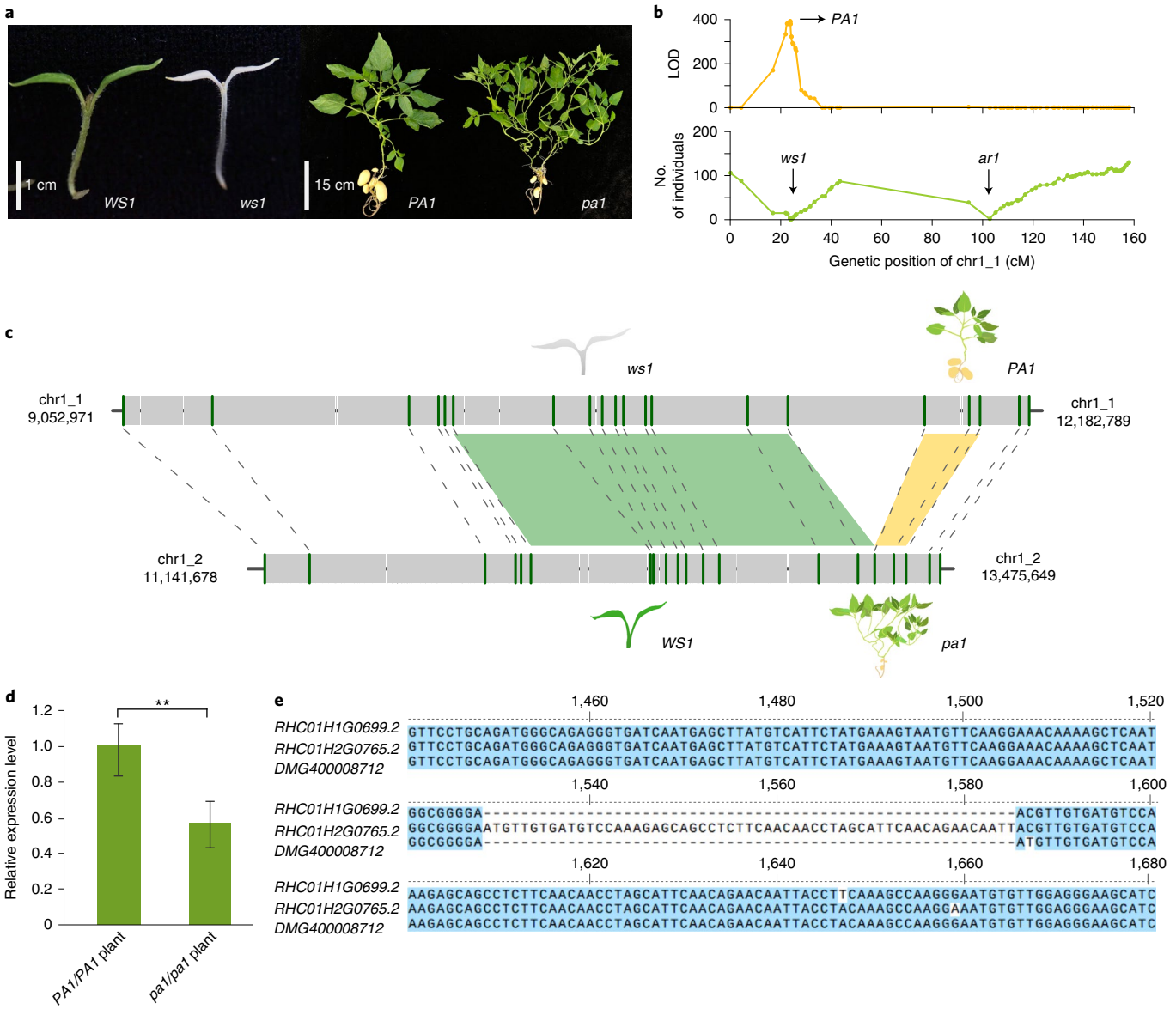


Fig. 3 | Tight linkage of two deleterious genes in the repulsion phase. a, Phenotype of normal seedling (*WS1*) and white seedling (*ws1*), normal plant architecture (*PA1*) and more branched architecture (*pa1*). **b**, Genetic mapping of *PA1* and *ws1* on chromosome 1_1. The top graph shows the likelihood of odd (LOD) value of *PA1* mapping using R/qtl software[41], and the bottom graph represents the number of individuals with a homozygous recessive allele. Dots in the graphs present the genetic markers. The *ar1* locus has been reported previously[39]. **c**, Fine mapping of *ws1* and *pa1* using indel markers (green bars). Gray segments represent the repeat elements; green and yellow blocks indicate the positions of *WS1/ws1* and *PA1/pa1*, respectively. **d**, Quantitative PCR (qPCR) result of *RHC01H1G0699.2* in normal (*PA1/PA1*) and more-branched (*pa1/pa1*) plants. Error bars represent the standard deviation from four biological replicates, and asterisks indicate significant differences between normal and more-branched plants (*t*-test, $^{**}P$ value < 0.01). **e**, CDS alignment of *RHC01H1G0699.2*, *RHC01H2G0765.2* and their homolog in DM (*DMG400008712*), showing the 57-bp indel between alleles.

and 23.9-Mb (96%) RH{1} MTPs have the best hit on chromosome 5_2. Collectively, these analyses demonstrated that the accuracy of haplotype determination is more than 95% at the chromosome level.

A total of 76,394 protein-coding genes were annotated in the RH genome, and evaluation with BUSCO[26] genes revealed that 97.0% (1,398) of 1,440 examined genes were complete, with 74.1% (1,306) duplicated. Comparative analyses among the gene models of RH, M6 (a diploid potato with an assembled genome[9]) and DM (the potato reference genome) identified 18,377, 3,842 and 10,742 lineage-specific genes in the three genomes, respectively, constituting 24.1%, 10.2% and 27.5% of their annotated genes, respectively (Supplementary Fig. 8). For example, the dominant tuber shape gene *Ro*, which was reported absent in the DM clone[27], has two homozygous copies (*RHC10H1G1859.2* and *RHC10H2G2643.2*) on two RH haplotypes and has one copy (*g7634.t1*) on M6.

To provide an accurate evaluation of the divergence between the two RH haplotypes, we identified polymorphisms between the 12 homologous chromosome pairs (Fig. 2a, Supplementary Figs. 9 and 10 and Extended Data Figs. 1–10). Based on the alignment of genes on the two haplotypes[28], 198 syntenic blocks were detected, covering 1.3 Gb (80.2%) of anchored sequence (Supplementary Table 10). Between syntenic blocks, 12,299,445 SNPs, 1,393,680 indels (~1–50 bp), 38,999 structural variants (SVs, >50 bp) and 1,878 genes showing presence and absence variation (PAV) were

identified[29,30], including 106 large SVs spanning more than 100 kb. Overall, the intragenomic diversity was estimated at ~2.1%, a level higher than the diversity among out-crossing maize lines[31]. Based on synteny and annotation, 59,907 genes (78.4% of all annotated genes) were identified as having homologs on the two haplotypes, and 20,583 pairs (41,166 genes) of those were considered as reliable allelic genes. Among them, alleles of 17,092 gene pairs showed variants within the coding sequence (CDS), including amino acid alternation and premature termination[32]. Based on amino acid conservation modeling[33,34], 4,761 and 1,753 pairs of allelic genes were predicted to have potential deleterious substitutions in one and both alleles, respectively, indicating substantial accumulation of mutations in this clonally propagated crop[35] (Supplementary Table 11).

To understand the expression landscape of allelic genes, RNA-sequencing (RNA-seq) data of ten tissues were analyzed using Kallisto pipeline[36,37]. Overall, 48,361 genes (63.3% of total) expressed in at least one tissue and 3,417 gene pairs (16.6% of allelic genes) exhibited unequal expression between two alleles, termed differentially expressed loci (DEL). The DEL were distributed randomly throughout the genome with higher and lower expressed alleles occurring alternatively on the two haplotypes (Fig. 2b and Supplementary Table 12). From methylation sequencing of mature leaves, immature small tubers (transection diameter <1 cm) and immature large tubers (transection diameter ~1–5 cm), on average, 24,929 differential methylated regions (DMRs) were identified between paired syntenic regions, resulting in allelic methylation differences in 6,345 gene pairs (30.8% of allelic genes; Supplementary Fig. 11)[38]. By comparing the DMRs with the DEL, we found the methylation difference explained only a fraction of the expression difference of alleles. For example, in immature small tuber tissue, only 292 DEL (27.5% of all DEL) showed both differential expression and methylation (Supplementary Fig. 12).

Through the analysis of an RH selfing population, 25.7% of genomic regions (430.8 Mb) exhibited strong segregation distortion (SD; χ^2 test, $P < 0.001$; Supplementary Fig. 13). Large-effect recessive deleterious mutations are the main cause of zygotic selection, which caused 71.4% of the SD regions. Using the selfed progeny of RH, we identified several loci affecting survival (white seedling 1 (*ws1*), abnormal rooting 1 (*ar1*), lethal allele 2 (*la2*)) or growth vigor (plant architecture 1 (*pa1*), plant architecture 2 (*pa2*) and weak vigor 1 (*wv1*)). Except for *wv1* (Supplementary Fig. 14), the other five loci have been previously reported[39]. Here, we relocated these loci on the phased RH genome, clarifying which haplotype contained the dominant or recessive allele (Fig. 3a,b and Supplementary Fig. 15)[40,41]. All six loci were located in the SD regions (Supplementary Table 13). Generally, large-effect deleterious mutations are relatively dispersed in the genome, which could be removed by sexual selection.

However, two recessive detrimental alleles, *ws1* and *pa1*, are tightly linked but in repulsion on the short arm of chromosome 1. Phenotype-based selection will not be sufficient to break the linkage of these two alleles, as demonstrated in practical breeding efforts; therefore, additional genetic analysis of this locus is required. In the mapping of *pa1* and *ws1*, using 880 F_2 progeny, the two loci remain linked (Fig. 3b). Genotyping of an additional 1,200 F_2 individuals identified two recombinant plants, which delimited *ws1* and *pa1* into two adjacent regions. The *WS1* locus was mapped to a 1.19 Mb region (chr1_2: 12,061,229–13,249,054) that contained 76 annotated genes, including six DEL and 30 allelic genes harboring exonic variants. The *PA1* locus was mapped to a 191-kb region (chr1_1: 11,821,758–12,012,928) containing 11 genes (Fig. 3c). Among them, one gene, *RHC01H1G0699.2*, encoding the ETHYLENE INSENSITIVE3 (EIN3) protein was identified as a candidate. *AtEIN3* has been reported to regulate plant growth in *Arabidopsis thaliana*[42] by acting as a transcriptional regulator in the ethylene signaling pathway. *RHC01H1G0699.2* showed decreased expression in more-branched plants (*pa1*/*pa1*) than in normal plants (*PA1*/*PA1*; Fig. 3d), and this expression pattern was consistent with observations on the *A. thaliana ein3* mutant. Between the dominant allele *RHC01H1G0699.2* and the recessive allele *RHC01H2G0765.2*, there was a 57-bp insertion that might result in additional translation of 19 amino acids in RHC01H2G0765.2 (Fig. 3e and Supplementary Figs. 16 and 17). Genome-assisted analyses of these two large-effect deleterious alleles provide tools to break the tight linkage in the repulsive phase for a better inbred line from RH.

In the current study, we combined multiple sequencing technologies to achieve the de novo assembly and haplotype determination of the heterozygous diploid potato. Compared with short reads or the longer but more error-prone ONT reads, CCS reads generated higher resolution and accuracy in differentiating haplotypes, which is particularly useful in resolving complex genomes. However, for a complex genome like diploid potato, there is still no tool that can build near-complete haplotypes from long-read sequencing or Hi-C sequencing without the assistance of genetic information, which requires further improvement in assembly algorithms.

The haplotype-resolved genome of the diploid potato provides a holistic view of the genome organization of a clonally propagated, heterozygous plant species. Haplotype-resolved identification of deleterious mutations, especially tightly linked genes in repulsion, provides insights into purging mutation burden by efficient molecular selection and/or genome-editing technologies[43]. As such, this study could facilitate the exploitation of heterosis, using inbred lines with complementary haplotypes, which is the core of diploid potato breeding.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0699-x.

Received: 10 June 2019; Accepted: 24 August 2020;
Published online: 28 September 2020

References

1. Ramulu, K. S., Dijkhuis, P. & Roest, S. Phenotypic variation and ploidy level of plants regenerated from protoplasts of tetraploid potato (*Solanum tuberosum* L. cv. 'Bintje'). *Theor. Appl. Genet.* **65**, 329–338 (1983).
2. Lindhout, P. et al. Towards F_1 hybrid seed potato breeding. *Potato Res.* **54**, 301–312 (2011).
3. Jansky, S. H. et al. Reinventing potato as a diploid inbred line-based crop. *Crop Sci.* **56**, 1412–1422 (2016).
4. Li, Y., Li, G., Li, C., Qu, D. & Huang, S. Prospects of diploid hybrid breeding in potato. *Chin. Potato J.* **27**, 96–99 (2013).
5. Stokstad, E. Breeders seek a breakthrough to help farmers facing an uncertain future. *Science* **363**, 574–577 (2019).
6. Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189–195 (2011).
7. Aversano, R. et al. The *Solanum commersonii* genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives. *Plant Cell* **27**, 954–968 (2015).
8. Hardigan, M. A. et al. Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proc. Natl Acad. Sci. USA* **114**, 9999–10008 (2017).
9. Leisner, C. P. et al. Genome sequence of M6, a diploid inbred clone of the high-glycoalkaloid-producing tuber-bearing potato species *Solanum chacoense*, reveals residual heterozygosity. *Plant J.* **94**, 562–570 (2017).
10. Jiao, W. B. & Schneeberger, K. The impact of third-generation genomic technologies on plant genome assembly. *Curr. Opin. Plant Biol.* **36**, 64–70 (2017).
11. Jiao, W. B. et al. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res.* **27**, 778–786 (2017).
12. Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D. & Stromvik, M. V. Current strategies of polyploid plant genome sequence assembly. *Front. Plant Sci.* **9**, 1660 (2018).

13. Biernacka, J. M. et al. Assessment of genotype imputation methods. *BMC Proc.* **3**, S5 (2009).
14. Ullah, E. et al. Comparison and assessment of family- and population-based genotype imputation methods in large pedigrees. *Genome Res.* **29**, 125–134 (2019).
15. Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
16. Mostovoy, Y. et al. A hybrid approach for de novo human genome sequence assembly and phasing. *Nat. Methods* **13**, 587–590 (2016).
17. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* **27**, 757–767 (2017).
18. Koren, S. et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018).
19. Zhang, J. et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* **50**, 1565–1573 (2018).
20. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
21. de Boer, J. M. et al. Homologues of potato chromosome 5 show variable collinearity in the euchromatin, but dramatic absence of sequence similarity in the pericentromeric heterochromatin. *BMC Genomics* **16**, 374 (2015).
22. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
23. Wang, J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
24. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
25. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
26. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
27. Wu, S. et al. A common genetic mechanism underlies morphological diversity in fruits and other plant organs. *Nat. Commun.* **9**, 4734 (2018).
28. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
29. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
30. Harris, R. S. *Improved Pairwise Alignment of Genomic DNA*. PhD thesis, The Pennsylvania State University (2007).
31. Sun, S. et al. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nat. Genet.* **50**, 1289–1295 (2018).
32. Cingolani, P. et al. A program for annotating and predicting the effects of single-nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
33. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
34. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
35. Ramu, P. et al. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* **49**, 959–963 (2017).
36. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–528 (2016).
37. Ramírez-González, R. H. et al. The transcriptional landscape of polyploid wheat. *Science* **361**, eaar6089 (2018).
38. Xi, Y. & Li, W. BSMAP: whole-genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232 (2009).
39. Zhang, C. et al. The genetic basis of inbreeding depression in potato. *Nat. Genet.* **51**, 374–378 (2019).
40. Meng, L., Li, H., Zhang, L. & Wang, J. QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* **3**, 269–283 (2015).
41. Arends, D., Prins, P., Jansen, R. C. & Broman, K. W. R/qtl: high-throughput multiple QTL mapping. *Bioinformatics* **26**, 2990–2992 (2010).
42. Munne-Bosch, S., Simancas, B. & Muller, M. Ethylene signaling cross-talk with other hormones in *Arabidopsis thaliana* exposed to contrasting phosphate availability: differential effects in roots, leaves and fruits. *J. Plant Physiol.* **226**, 114–122 (2018).
43. Kremling, K. A. G. et al. Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* **555**, 520–523 (2018).

Methods

Genome, transcriptome and methylome sequencing. To construct an Illumina sequencing library, genomic DNA was extracted from RH leaves by using the cetyltrimethylammonium bromide (CTAB) method[44]. The library was sequenced on the Illumina HiSeq 2500 platform, generating 155 Gb of 250-nucleotide paired-end reads with an insert size of ~400 bp.

About 1.2 ng high molecular weight DNA (>50 kb) was isolated and loaded for 10xG library construction, following the manufacturer's recommended protocols (https://support.10xgenomics.com/de-novo-assembly/library-prep/doc/user-guide-chromium-genome-reagent-kit-v1-chemistry/). The 10xG library was sequenced on the Illumina HiSeq X Ten platform, yielding 122 Gb of 150-nucleotide paired-end data.

In construction of the ONT library, an optimized protocol for long plant DNA enrichment was applied[45]. The library was constructed using LSK108 kit (SQK-LSK108, Oxford) and sequenced using 38 R9.4 flow cells on the Nanopore GridION X5 sequencer. The base calling was performed using Albacore in MinKNOW package, and 10.7 million nanopore reads with an N50 length of 25.3 kb were available for assembly.

For CCS, genomic DNA was extracted from in vitro seedlings using the DNeasy Plant Mini Kit (Qiagen). The integrity of the DNA was determined with the Agilent 4200 Bioanalyzer (Agilent Technologies). Genomic DNA (15 μg) was sheared using g-Tubes (Covaris) and concentrated with AMPure PB magnetic beads. Two SMRTbell libraries were constructed using the Pacific Biosciences SMRTbell Template Prep Kit 1.0. The libraries were size selected on a BluePippin system for molecules with a size of 11 Kb, followed by primer annealing and the binding of SMRTbell templates to polymerases with the DNA/Polymerase Binding Kit. Libraries sequencing was carried out on the Pacific Bioscience Sequel II platform (Annoroad Gene Technology) and 29-Gb CCS reads with an N50 size of 13 kb were generated using ccs software v.3.0.0 (https://github.com/pacificbiosciences/unanimity/).

The Hi-C libraries were constructed at Annoroad Gene Technology using the in situ method[46]. DNA from in vitro seedlings of RH was digested with MboI using the standard Hi-C library preparation protocol. The Hi-C libraries were sequenced on an Illumina HiSeq X Ten platform, yielding 150 Gb of data.

The selfing population (S_1 population, equivalent to an F_2 population) of RH was constructed by forced self-pollination[39] and 880 F_2 individuals were sequenced at ~1× depth using an Illumina HiSeq X Ten platform. On average, ~2 Gb of data were obtained from each individual.

Samples from the young leaf, mature leaf, stem, perianth, anther, carpel, stolon, immature small tuber (transection diameter of <1 cm), immature big tuber (transection diameter of 1–5 cm) and root tissue were collected for transcriptome sequencing. All tissues were isolated and sequenced in three biological replicates. Total RNA was extracted from the samples using the TIANGEN Kit with DNase I and processed for the library construction using NEBNext UltraTM RNA Library Prep Kit. Following the removal of low-quality data, ~3 Gb of 150-nucleotide paired-end data for each sample were used for further RNA-seq analysis.

In addition to the transcriptome sequencing, samples from three tissues—mature leaf, immature small tuber (transection diameter <1 cm) and immature large tuber (transection diameter of 1–5 cm)—were used for whole-genome bisulfite sequencing with three biological replicates. Genomic DNA was extracted using the CTAB method and fragmented to ~200–300 bp by sonication before library construction. The barcoded DNA was treated twice with bisulfite using the EZ DNA Methylation-Gold Kit (Zymo Research) according to the manufacturer's instructions. The libraries were sequenced on a HiSeq X Ten platform and 50 Gb of 150-nucleotide paired-end data were generated.

Hierarchical assembly and phasing of diploid potato genome using 10xG and ONT data. *Whole-genome de novo assembly.* The WGS Illumina reads were assembled using DISCOVAR de novo (https://software.broadinstitute.org/software/discovar/blog/), resulting in 1.3-Gb sequences with a scaffold N50 size of 14.9 kb. The 10xG reads were assembled using Supernova (https://github.com/10XGenomics/supernova/) with the 'megabubbles' output; 1.58 Gb of assembled sequence data were generated.

Genetic genotyping and grouping. In the pipeline, the 880 F_2 selfing progeny were sequenced at ~1×–2× coverage for genetic mapping (Supplementary Fig. 5). First, for each progeny, the sequenced reads were mapped to the assembled RH scaffolds using BWA-MEM[47]. For each scaffold, the raw number of mapped reads with mapping quality >50 was normalized according to the scaffold length, the total assembled length and the total mapped reads of this progeny. Then the normalized read number for all scaffolds was transformed to genotype scores. The genetic groups were built using the software JoinMap (v4.0)[48]. A total of 1,408 Mb sequences were genotyped and grouped into 12 linkage groups, corresponding to the 12 chromosomes of monoploid potato.

For each linkage group, we applied the R function *hclust* (method = 'ward.D2', $k = 2$) to separate each group into two clusters, corresponding to the two haplotypes of the diploid potato. To assign the residual 310 Mb of scaffolds, which displayed obscure read distribution and failed in genotype calling, we calculated the correlation between grouped scaffolds (target) and residual scaffolds (query) on the number of mapped reads using the *cor* function in R (Supplementary Fig. 5). If the query scaffold and the target scaffold shared a similar pattern (correlation > 0.7) on read distribution in the population, they were deemed to belong to same linkage group. For each query scaffold, we determined its group using two criteria: (1) the top two correlation values with target scaffolds should be larger than 0.7 and (2) the top two target scaffolds showing the highest correlation values should be located on the same group. After this process, 117.8 Mb of residual scaffolds were assigned to 24 linkage groups. In total, 1.52 Gb of 1.7 Gb sequences were grouped into 24 clusters, accounting for 90% of the assembled genome.

Simplified reassembly within each group. One effective way to simplify the assembly for a complex genome is to dilute the genome into multiple parts and separately assemble each part. In this project, we leveraged a similar simplification by separately reassembling the 24 clusters. First, the ONT long reads were mapped to the scaffolds using minimap2 (ref. [49]). Second, the reads belonging to each genetic group were retrieved and assembled into contigs using SMARTdenovo (https://github.com/ruanjue/smartdenovo/). Only reads with properly paired mapping and less than two mismatched bases reads were collected for the reassembly. Third, the contigs were polished iteratively using Racon[22] and Pilon[23]. Last, the 10xG reads were aligned to the contigs using Long Ranger (https://support.10xgenomics.com/genome-exome/software/downloads/latest) to generate scaffolds using the ARCS + LINKS pipeline[50], which increased the assembly continuity from contig N50 length of 636 kb to 921 kb. The hybrid assembly yielded the genome draft RHgv1.

Genome assembly and phasing using PacBio CCS reads. A total of 29 Gb CCS reads were assembled using Canu (v1.91)[25] with the parameter --pacbio-hifi. Canu generated two assemblies composed of contigs and unitigs (Supplementary Table 4), and the unitig assembly consisted of the contigs that split at any alternative paths in the assembly graph. The contig assembly had longer continuity but more chimeric fragments as revealed in the genetic mapping analysis. To avoid the mis-joining of two haplotypes, the unitig assembly rather than the contig assembly was chosen for the subsequent analysis. The unitigs were then polished iteratively using two rounds of Pilon[23] with ~150 Gb of WGS Illumina data, generating the genome draft RHgv2.

Similarly, the sequenced reads of RH selfing progeny were mapped to unitigs of RHgv2 to perform genetic grouping. Because the unitigs were relatively long (N50 = 2 Mb), windows with a size of 200 kb rather than the whole unitig were used. If the adjacent windows of one unitig showed contrary read distribution, the unitig was defined as chimeric and broken between windows; 40 chimeric unitigs with a total length of 95 Mb were broken. In total, 1.31 Gb of 1.53 Gb sequences were assigned to 24 linkage groups.

After merging, 141 Mb sequences and 5,252 annotated genes of RHgv1 were added to the RHgv2, yielding a 1.67 Gb genome draft with 1.54 Gb sequences assigned to 24 groups, termed RHgv3 (Supplementary Table 6). The sequences from the RHgv1 and RHgv2 assemblies were named as ontctg* and unitig* in the AGP file, respectively (Supplementary Data 3).

Construction of pseudochromosomes. As no approach generated satisfactory results on the RH genome, we introduced the group information derived from the genetic mapping to assist the Hi-C application on chromosome-level assembly. The process was performed on RHgv3 including three steps as follows:

1. Align. The 24 previously determined groups were divided into two haplotypes to generate two pseudohaploid genome drafts. The 132 Mb sequences that could not be assigned to any group were added to two pseudohaploid genomes. Total Hi-C reads were aligned to each pseudohaploid genome using HiC-Pro[51] to calculate the contact frequency. This step yielded two bam files for the two pseudohaploid genomes.
2. Rescue. Using the bam file as input, the *rescue* function in ALLHiC[20] was applied to assign unplaced sequences to known groups. Because the 132 Mb unplaced sequences were added to two pseudohaploid genomes and processed twice, the rescued results were redundant. For every unplaced sequence, we considered its best Hi-C signal density to decide the group to which it belonged. After this step, the sequence content of 24 groups was updated with an extra 75.6 Mb sequences assigned to proper groups.
3. Optimize and build. For each pseudohaploid genome, using the bam file and the updated group file as input, the *optimize* function in ALLHiC decided the order and orientation of scaffolds for each group; thus, the *build* function generated fasta sequences on that basis. By performing this step, we identified the pseudochromosomes for 24 groups. The order and orientation of scaffolds on chromosomes are provided in Supplementary Data 3.

Genome assembly assessment. The BAC clones and BEs of RH were downloaded from http://solanaceae.plantbiology.msu.edu/pgsc_download.shtml (ref. [6]) to assess the assembly.

Assess scaffold assembly. Using Sanger technology, the 54,902 paired BEs were sequenced with the average length of 714 nucleotides[6]. The BEs were aligned to

the assembled scaffolds using BLASR (v1.3.1)[52], and those aligned with >98% query coverage and >98% identity were considered as the successful alignments. A reasonable distance between end sequences (set to ~30–300 kb) and correct orientation (± for each of them), when mapped on to the genome, were used as criteria for assessing the correctness of the assemblies.

The 184 BACs that assembled with ordered contigs and a total length of 21,734,426 bp were aligned to RH scaffolds using MUMmer (v3.23)[29]. The alignment was filtered using the cutoff criteria: identity >98% and alignment length >2 kb. Some BACs completely aligned with single scaffolds, while others were fragmented or repeatedly mapped to multiple scaffolds. To evaluate the correctness of scaffolds or contigs, only the BACs that mapped to single scaffolds of genome assembly were used in the statistics. The mapping structure between BACs and scaffolds was manually checked to determine if the alignments were complete and collinear (Supplementary Data 1 and 2). SNPs and indels were identified from the alignment between BACs and scaffolds using the *show-snps* function in MUMmer.

Assess phasing quality. The BE sequences and 1,639 RH BACs with a total length of 205 Mb and an average size of 125 kb were aligned to the pseudochromosomes of RHgv3 using BLASR (v5.1) to assess the haplotypes. Because most of the BACs contained only unordered contigs, we considered the alignment length and identity when selecting the best hits for BACs. Only alignments with a mapQV of >50 and identity of >95% were retained for downstream analyses.

A total of 55.4 Mb nonredundant BAC MTPs with an N50 length of 336 kb were extracted from the assembly of the previously published RH chromosome 5 (ref. [21]). MTPs were aligned to the RH chromosomes using MUMmer (v3.23)[29] and filtered using the criteria of identity >95% and alignment length >2 kb. For each MTP, only the best hit was considered to compare the phases.

Genome annotation. Repeat-sequence masking was performed using RepeatMasker (v4.0.6) with default parameters. The reference repeat libraries included plant short fragment repeats and DM annotated repeats[6]. The RNA-seq data were aligned to the reference genome with HISAT2 (v2.0.4) and assembled using StringTie (v1.2.2)[53–55]. Trinity (v2.4.0) was used to assemble transcripts with (--genome_guided_max_intron 15,000 --genome_guided_bam --min_kmer_cov 2 --trimmomatic --normalize_reads) and without (--min_kmer_cov 2 --trimmomatic --normalize_reads --no_bowtie) reference guidance. To perform ab initio gene prediction, PASA (v2.2.0)[56] was used to build the coding region model. This PASA step utilized assembled transcripts from Trinity as the library and trained the model. The ab initio predictions included SNAP (--categorize 100, --export 1,000 and --plus)[57], AUGUSTUS (v2.7)[58] and GlimmerHMM (v3.0.4)[59]. Two Trinity assemblies combined with ab initio gene prediction results were fed into EVM software (v1.1.1) to merge into a final gene set.

The annotated CDSs of the RH, DM and M6 genomes were aligned using BLAT[60], and the homologous genes were screened in each genome using coverage of >75% and identity of >75% as the criteria.

Haplotype comparison and diversity analysis. To identify the homologous regions between two haplotypes, we applied the MCScanX package[28] to construct the syntenic blocks based on well-aligned genes. We screened the syntenic regions according to the following criteria: (1) paired regions must be on homologous haplotypes, (2) one segment should not be larger than three times the length of its counterpart and (3) aligned regions must cover over 50% of the whole region. Regions meeting these criteria were trusted as syntenic regions. One gene and its best homologous gene on the complementary haplotype were considered as allelic genes.

The syntenic regions were then subjected to LASTZ (v1.02.00)[30] with the parameters --chain --format=diff --matchcount=3,000 --rdotplot --strand=plus/minus --ambiguous=n. The homologous chromosomes were aligned using MUMmer (4.0)[29], and the SVs were detected from the differences reported by the *show-diff* function. To reduce the number of false positives, we only identified the PAV genes in syntenic regions and defined a PAV gene as one that lacked a homolog at the complementary haplotype, while its surrounding genes had homologs that were arranged in good collinearity between two haplotypes.

The SNPs and indels between haplotypes were annotated using SnpEff[32]. To detect the mutations that were potentially deleterious, we aligned the RH chromosomes to the potato reference genome DM using LASTZ and performed in silico prediction on the SNPs through the 'sorting intolerant from tolerant' (SIFT) algorithm[33,34]. The underlying premise of this algorithm is based on the evolutionary conservation of the amino acid within protein families: highly conserved positions tend to be intolerant to substitution, whereas those within a low degree of conservation tolerate most substitutions.

Gene expression analysis. The allele-specific mapping of Kallisto[36] was used in the comparison of homologous expression in polyploid wheat[37], and we applied the software to the RNA-seq data to obtain the expression levels in transcripts per million (TPM) of genes on both haplotypes. Only genes that showed <30% variance of expression levels in biological replicates were retained for further analysis. Genes with a summed TPM value of >1 for all tissues were taken as expressed genes. We then tested the expression difference between allelic genes with the *binom.test* using an R script.

Methylation analysis. The whole-genome bisulfite sequencing reads from each sample were mapped to the RH genome using BSMAP[38], allowing only unique mapping and mismatches of up to 4%. Positional DNA methylation levels were computed using the methratio.py script in the BSMAP package. To define differentially methylated positions (DMPs) between two haplotypes, we compared the methylation level of pairwise C sites in syntenic regions using Fisher's exact test. We empirically used the reads depth of ≥5, CG difference of <0.4, CHG difference of <0.2, CHH difference of <0.1 and a P value < 0.01 derived from two-tailed Fisher's exact test to screen the DMPs[61,62]. For each tissue, only the DMPs supported by all the three replicates were retained for further analysis. Then, the DMPs with the same content were collapsed into a DMR, only if the distance on the chromosome of the nearest two DMPs was less than 100 bp.

Gene mapping of six loci related to inbreeding depression. *Mapping based on the sequencing.* To construct the genetic map, we scanned the chromosome using a 300-kb window, and the windows were genotyped as markers using the method described above. For each of the given 24 linkage groups, the markers were ordered on the genetic map using IciMapping (v4.0)[40] with the parameters: LOD ≥ 3 and algorithm = nnTwOpt.

For *pa1*, *pa2* and *wv1*, the regular genetic mapping was performed using R/qtl[41] (https://www.rqtl.org/) with the *cim* function, and the candidate interval was defined by the peak LOD bin and its adjacent two bins.

For the loci controlling growth vigor, *ar1*, *la2* and *ws1*, homozygous recessive genotypes were lethal and absent in the selfing population, impeding the effectiveness of the regular linkage mapping. Thus, we localized *ar1*, *la2* and *ws1* by screening the regions that excluded the homozygous recessive genotype.

Fine mapping using indel markers. To identify the recombinant plants of *ws1* and *pa1*, we sowed another 1,200 selfed seeds from RH on culture medium, and the seedlings were genotyped using the newly designed heterozygous indel primers in the candidate region (Supplementary Table 14).

Reverse transcription qPCR analysis of RHC01H1G0699.2. Total RNA was extracted from the leaves of *PA1*/*PA1* and *pa1*/*pa1* plants at the seedling stage using the TIANGEN kit with DNase I. The RNA was reverse transcribed using PrimeScript RT reagent kit with gDNA Eraser (Takara). Reverse transcription qPCR analysis was conducted with the StepOnePlus System (Applied Biosystems) using TB Green Premix Ex Taq GC (Takara). The reaction procedure was 95 °C for 30 s, 95 °C for 5 s and 60 °C for 30 s for 40 cycles. Actin was used as the internal control gene. All analyses were conducted with four biological replicates. The relative gene expression levels were calculated using the $2-\Delta Ct$ method, and a t-test was performed to compare the results of *PA1*/*PA1* and *pa1*/*pa1* plants.

Statistical analysis. The χ^2 test statistic was performed using the *chisq.test* function in R. The expression difference between alleles was determined using the *binom.test* function in R (parameters: p = 0.5, alternative = two.sided, conf. level = 0.95). Two-tailed Student's t-tests were calculated using *t.test* in R. The two-tailed Fisher's test was performed using the *fisher.test* function in R.

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

Data availability

The final RH genome assembly (RHgv3), annotation and a genome browser are available at Spud DB (https://solanaceae.plantbiology.msu.edu/rh_potato_download.shtml/). This whole-genome shotgun project has been deposited at GenBank under accession numbers JACDXL000000000 and JACDXM000000000. The raw sequencing data have been deposited in the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra/) under BioProject accession number PRJNA573826. The data have also been submitted to the Chinese National Genomics Data Center (https://bigd.big.ac.cn/) under accession number CRA002005. Source data are provided with this paper.

Code availability

The custom pipelines and scripts used in the project are deposited in GitHub (https://github.com/zhouqiansolab/Haplotype-resolved-potato-genome/).

References

44. Gawel, N. J. & Jarret, R. L. A modified CTAB DNA extraction procedure for Musa and Ipomoea. *Plant Mol. Biol. Report.* **9**, 262–266 (1991).
45. Schmidt, M. H. et al. De novo assembly of a new *Solanum pennellii* accession using nanopore sequencing. *Plant Cell* **29**, 2336–2348 (2017).
46. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* **72**, 65–75 (2015).

47. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997v2 (2013).
48. Van Ooijen, J. W. Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genet. Res.* **93**, 343–349 (2011).
49. Li, H. & Birol, I. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
50. Yeo, S. et al. ARCS: scaffolding genome drafts with linked reads. *Bioinformatics* **34**, 725–731 (2018).
51. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
52. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
53. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
54. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
55. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
56. Haas, B. J. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
57. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
58. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
59. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
60. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).
61. Zhang, Y. et al. Large-scale comparative epigenomics reveals hierarchical regulation of non-CG methylation in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **115**, E1069–E1074 (2018).
62. Wang, L. et al. Comparative epigenomics reveals evolution of duplicated genes in potato and tomato. *Plant J.* **93**, 460–471 (2018).

Acknowledgements

We thank J. Ruan, X. Zhang, G. Zhu and W. Lucas for project discussions and critical comments. We thank the AGIS CAAS-YNNU Joint Academy of Potato Sciences for greenhouse assistance. This work was supported by the Agricultural Science and Technology Innovation Program (ASTIP-CAAS) and the National Natural Science Foundation of China (31601360). This work was also supported by the Chinese Ministry of Agriculture and Rural Affairs and the Shenzhen municipal (The Peacock Plan grant no. KQTD2016113010482651 to S.H.) and Dapeng district governments. This work was supported by a grant from the U.S. National Science Foundation (ISO-1237969) to C.R.B.

Author contributions

S.H. and Q.Z. designed the experiments and wrote the manuscript. Q.Z. and D.T. performed the majority of bioinformatics analyses. W.H. assisted in bioinformatics analyses and Y.Z. contributed to methylome analysis. Z.Y. carried out the mapping of *ws1* and *pa1*. C.Z., Z.Z., C.R.B., J.P.H., C.W.B.B. and R.G.F.V. analyzed the data and revised the manuscript. S.H., C.Z. and Z.Z. coordinated the project.

Competing interests

The authors declare no competing interests.

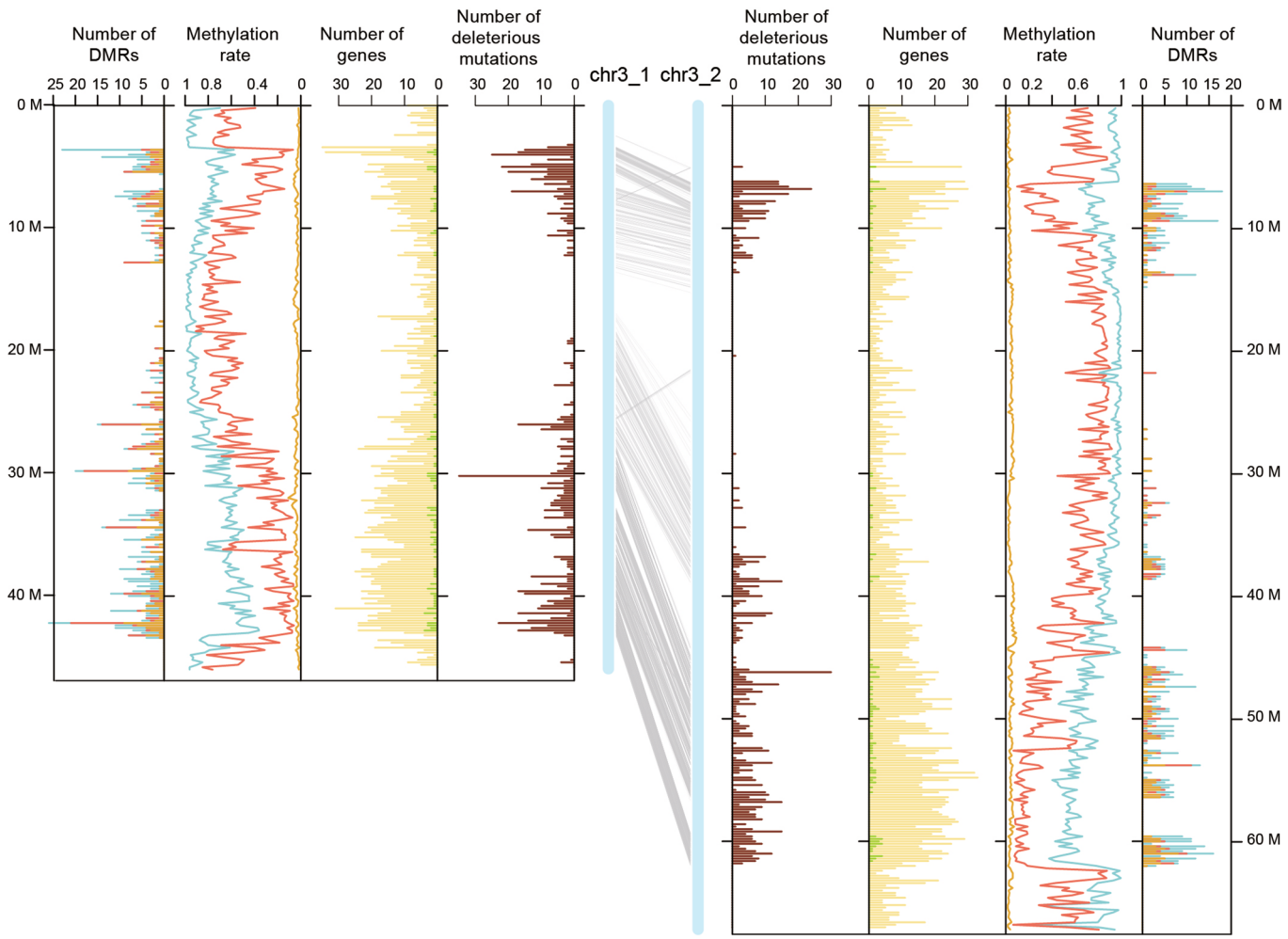
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-020-0699-x.

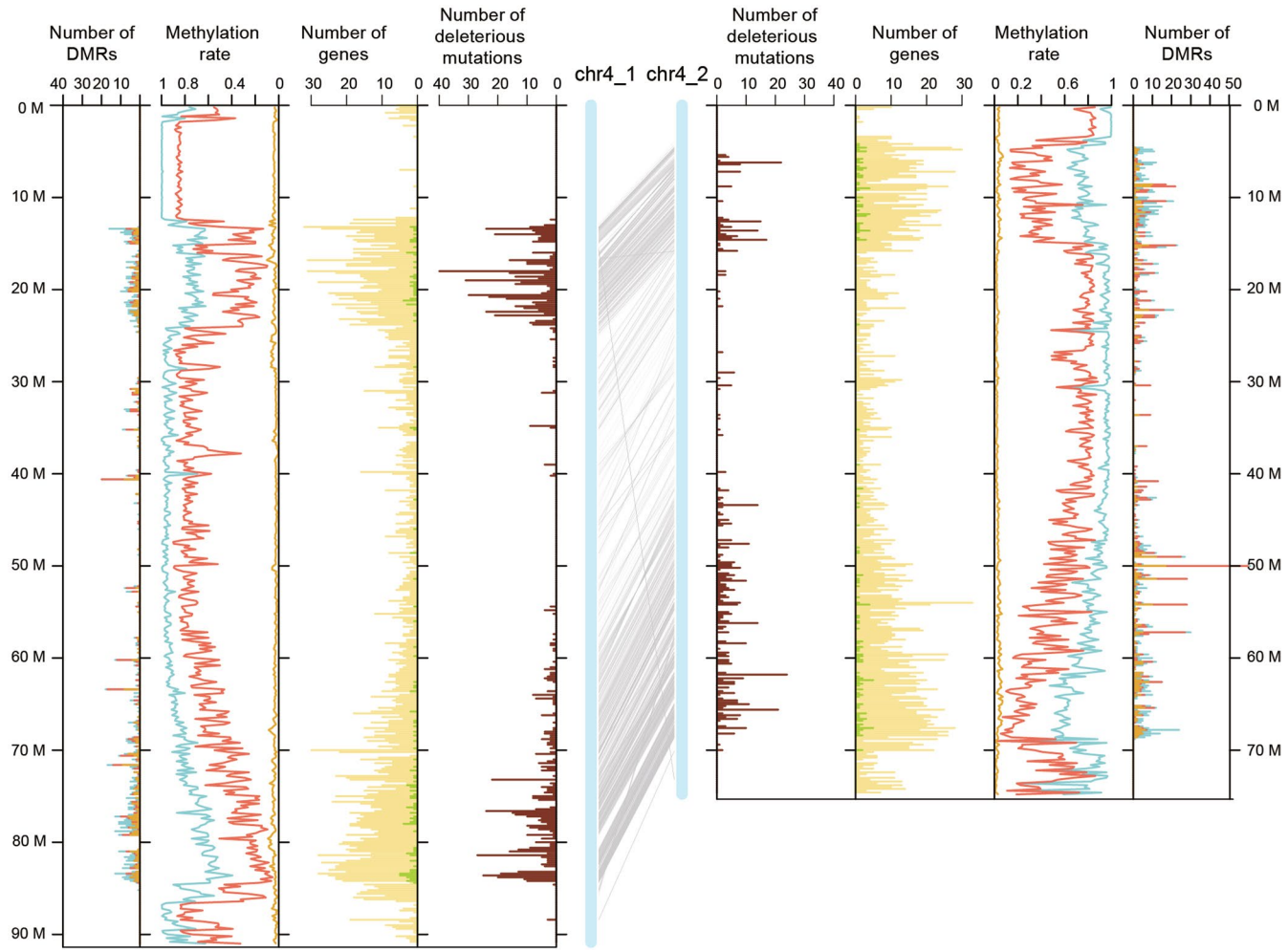
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0699-x.

Correspondence and requests for materials should be addressed to S.H.

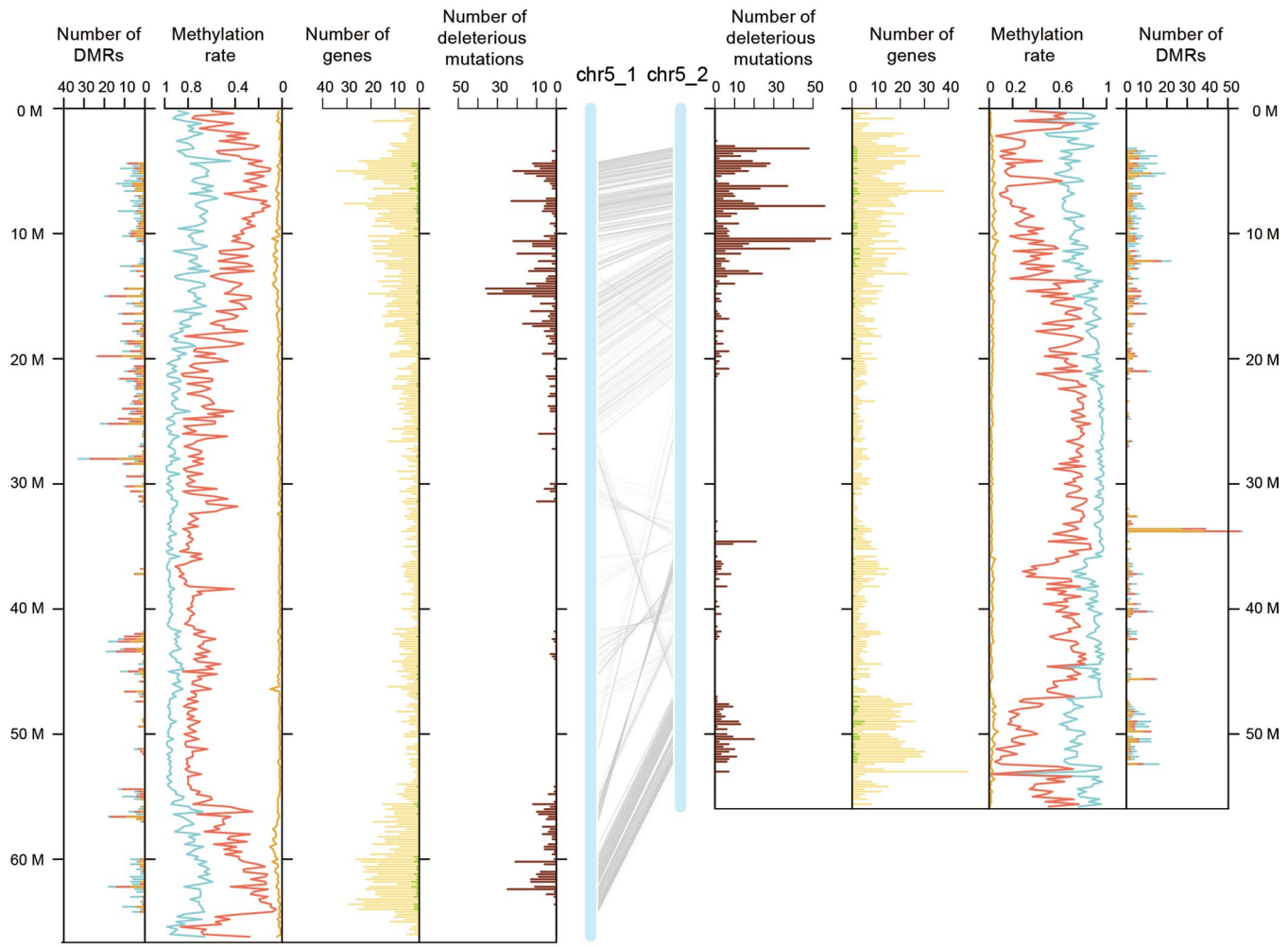
Reprints and permissions information is available at www.nature.com/reprints.



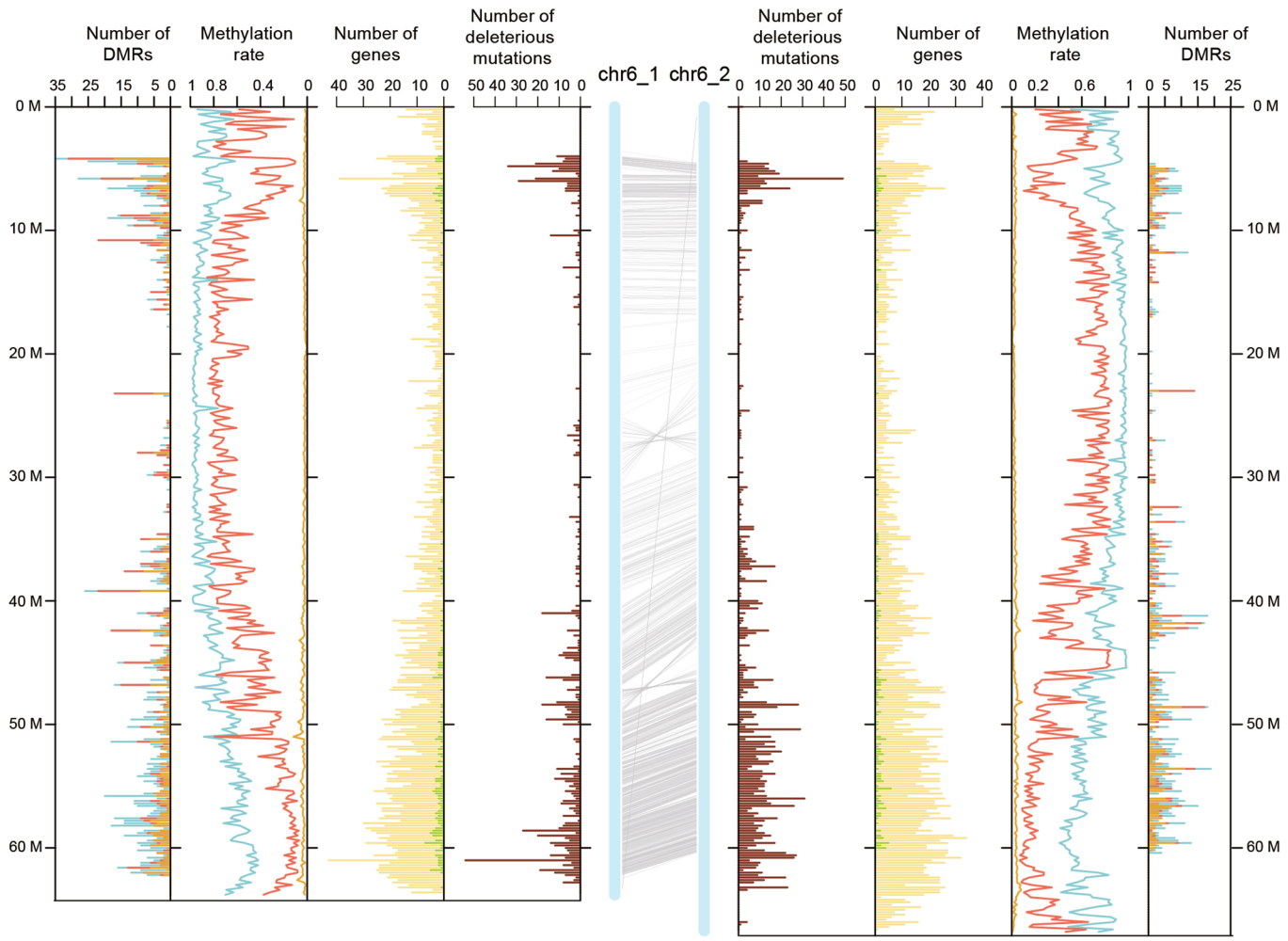
Extended Data Fig. 1 | Haplotype alignment of RH chromosome 3. The central blue bars represent the two haplotypes of chromosome 3 with the gray lines indicating the paired allelic genes. The distribution of deleterious or dysfunctional mutations (brown), annotated genes (yellow), preferentially expressed alleles (green), methylation level of three contexts and differentially methylated regions are arranged symmetrically for each haplotype. The methylation level and the number of DMRs of methylated sites in CG (light blue), CHG (red) and CHH (orange) contexts are indicated by cumulative column chart. Number of DMRs on one haplotype only involves the DMRs with hyper-methylation. All of the numbers were determined in 200 kb windows.



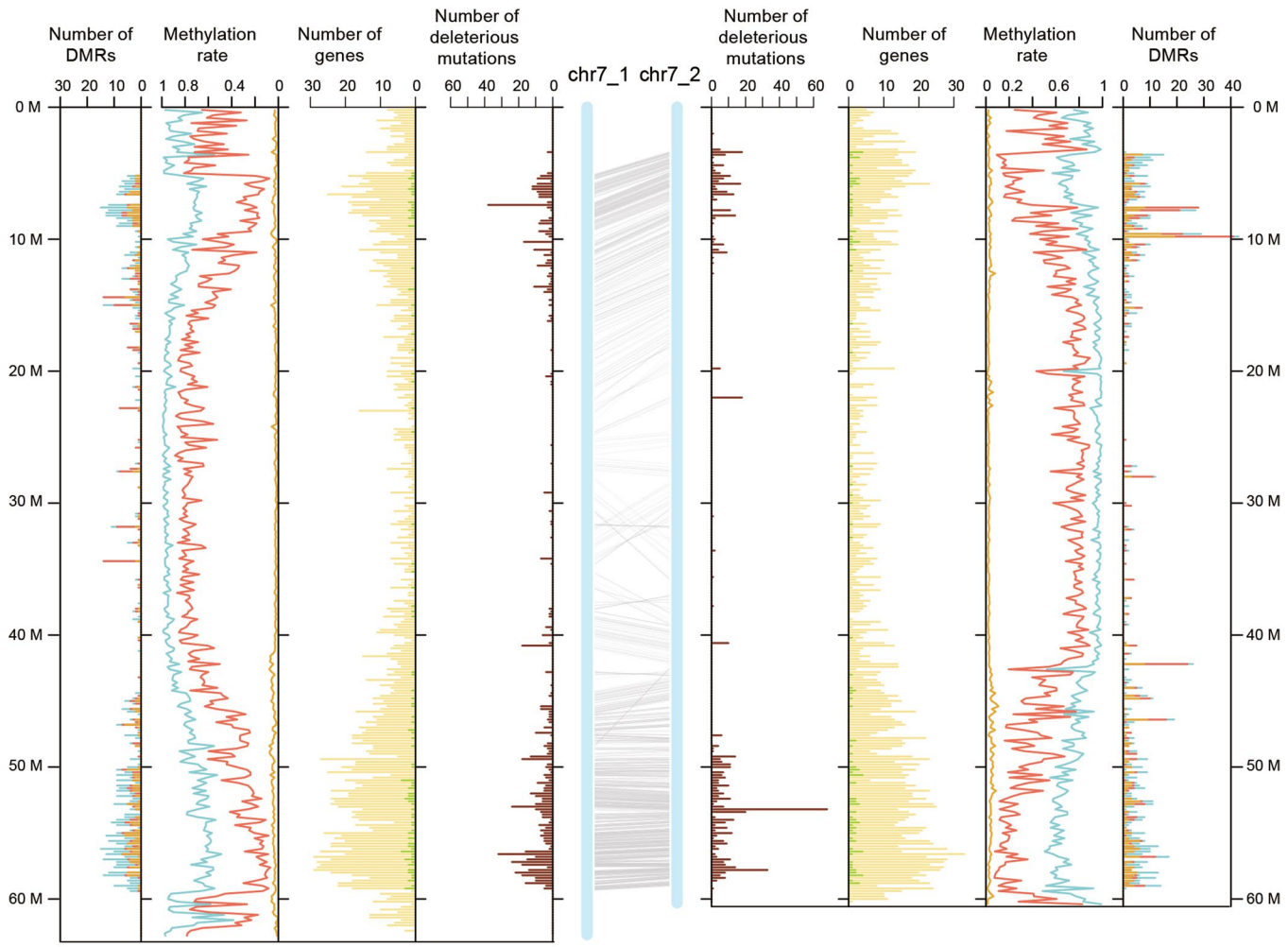
Extended Data Fig. 2 | Haplotype alignment of RH chromosome 4. The central blue bars represent the two haplotypes of chromosome 4 with the gray lines indicating the paired allelic genes. The distribution of deleterious or dysfunctional mutations (brown), annotated genes (yellow), preferentially expressed alleles (green), methylation level of three contexts and differentially methylated regions are arranged symmetrically for each haplotype. The methylation level and the number of DMRs of methylated sites in CG (light blue), CHG (red) and CHH (orange) contexts are indicated by cumulative column chart. Number of DMRs on one haplotype only involves the DMRs with hyper-methylation. All of the numbers were determined in 200 kb windows.



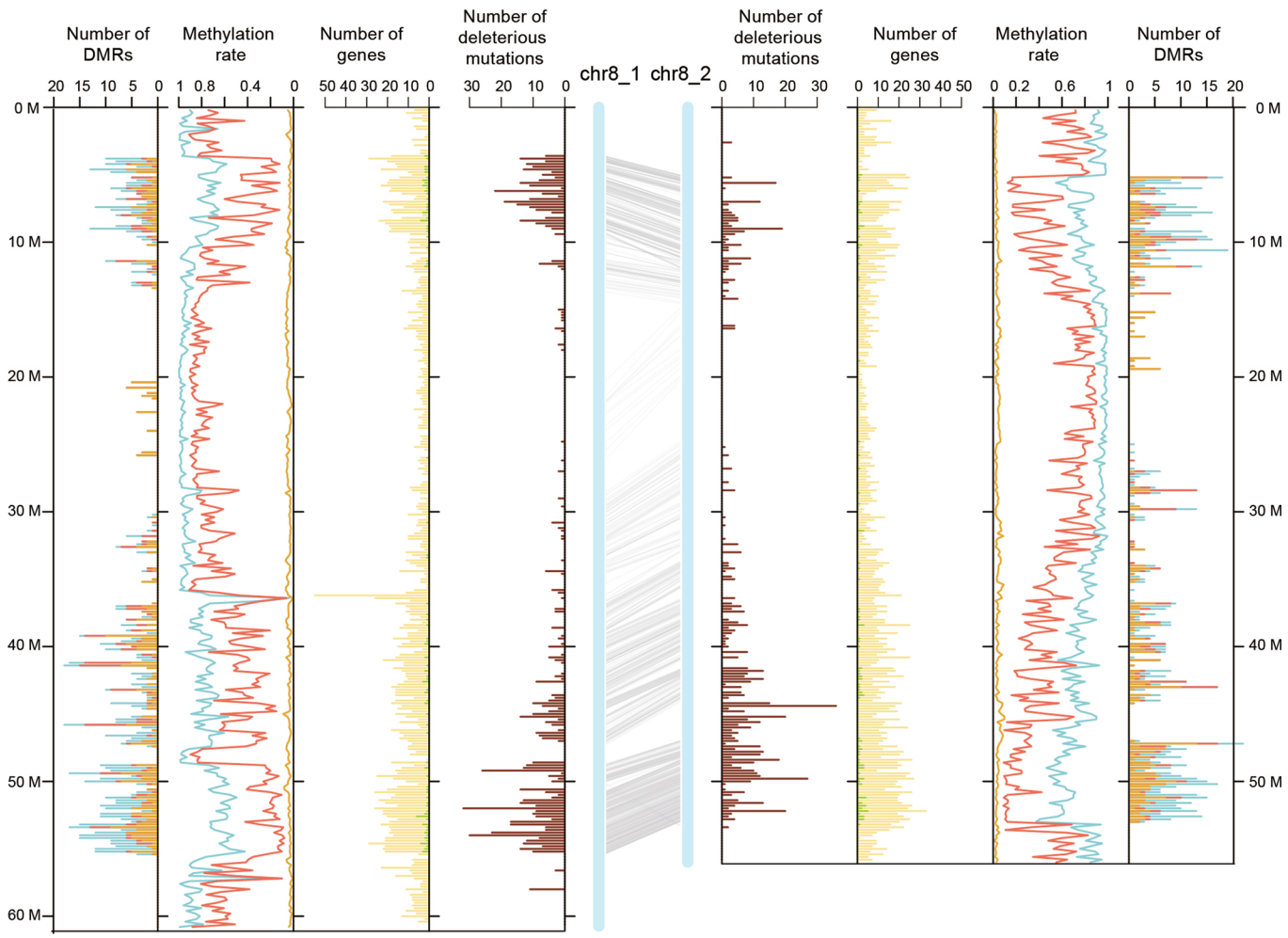
Extended Data Fig. 3 | Haplotype alignment of RH chromosome 5. The central blue bars represent the two haplotypes of chromosome 5 with the gray lines indicating the paired allelic genes. The distribution of deleterious or dysfunctional mutations (brown), annotated genes (yellow), preferentially expressed alleles (green), methylation level of three contexts and differentially methylated regions are arranged symmetrically for each haplotype. The methylation level and the number of DMRs of methylated sites in CG (light blue), CHG (red) and CHH (orange) contexts are indicated by cumulative column chart. Number of DMRs on one haplotype only involves the DMRs with hyper-methylation. All of the numbers were determined in 200 kb windows.



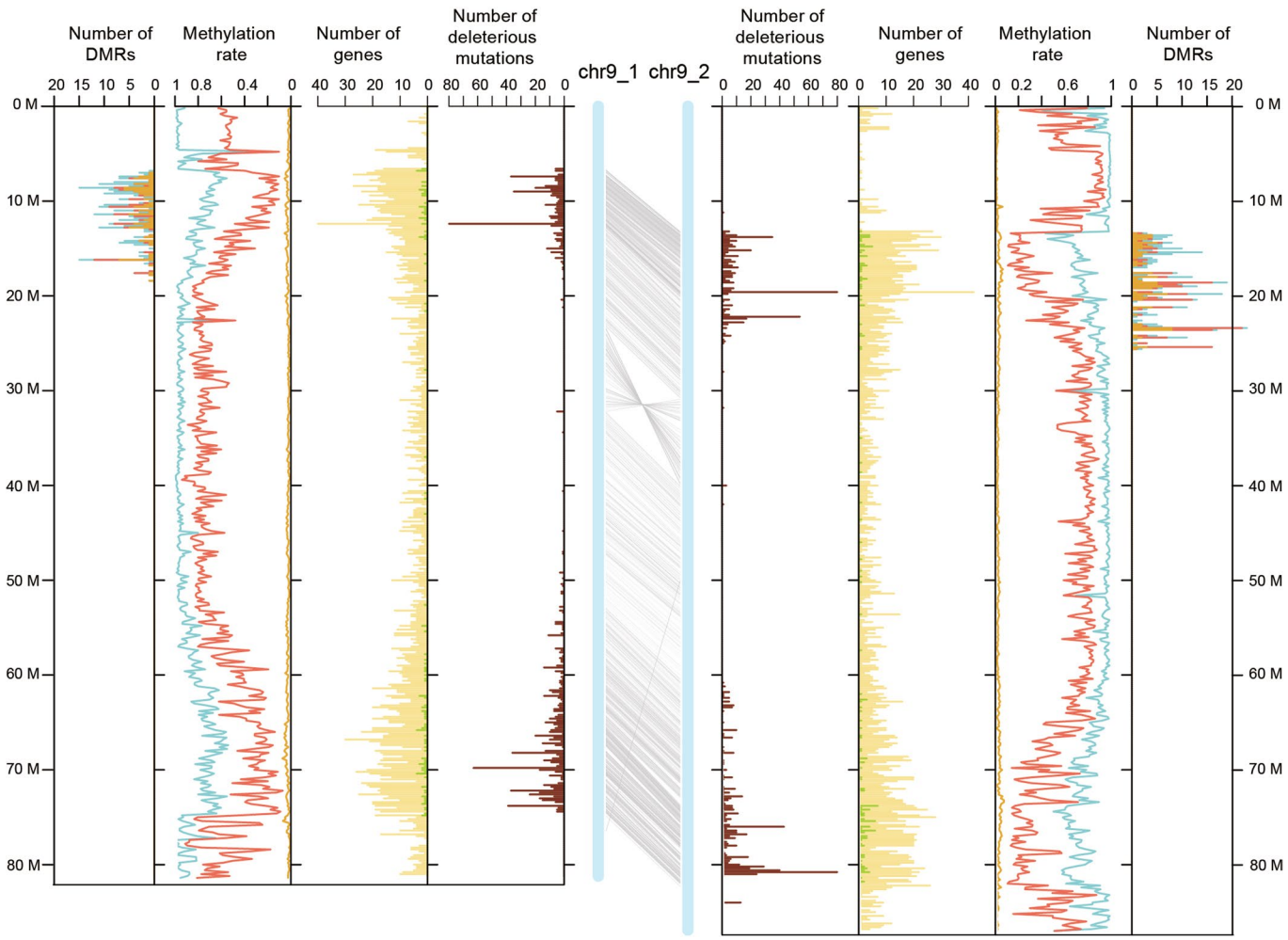
Extended Data Fig. 4 | Haplotype alignment of RH chromosome 6. The central blue bars represent the two haplotypes of chromosome 6 with the gray lines indicating the paired allelic genes. The distribution of deleterious or dysfunctional mutations (brown), annotated genes (yellow), preferentially expressed alleles (green), methylation level of three contexts and differentially methylated regions are arranged symmetrically for each haplotype. The methylation level and the number of DMRs of methylated sites in CG (light blue), CHG (red) and CHH (orange) contexts are indicated by cumulative column chart. Number of DMRs on one haplotype only involves the DMRs with hyper-methylation. All of the numbers were determined in 200 kb windows.



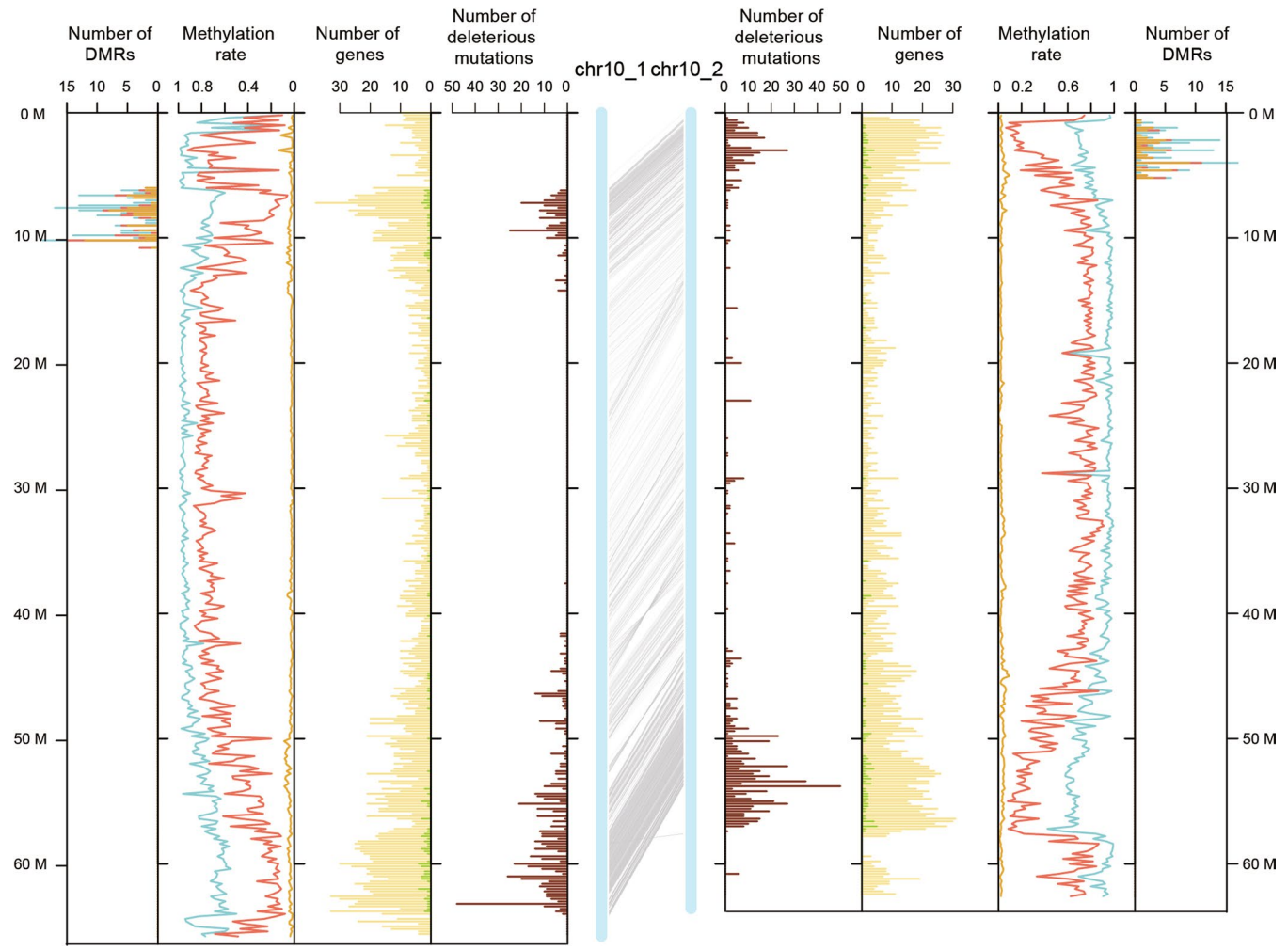
Extended Data Fig. 5 | Haplotype alignment of RH chromosome 7. The central blue bars represent the two haplotypes of chromosome 7 with the gray lines indicating the paired allelic genes. The distribution of deleterious or dysfunctional mutations (brown), annotated genes (yellow), preferentially expressed alleles (green), methylation level of three contexts and differentially methylated regions are arranged symmetrically for each haplotype. The methylation level and the number of DMRs of methylated sites in CG (light blue), CHG (red) and CHH (orange) contexts are indicated by cumulative column chart. Number of DMRs on one haplotype only involves the DMRs with hyper-methylation. All of the numbers were determined in 200 kb windows.



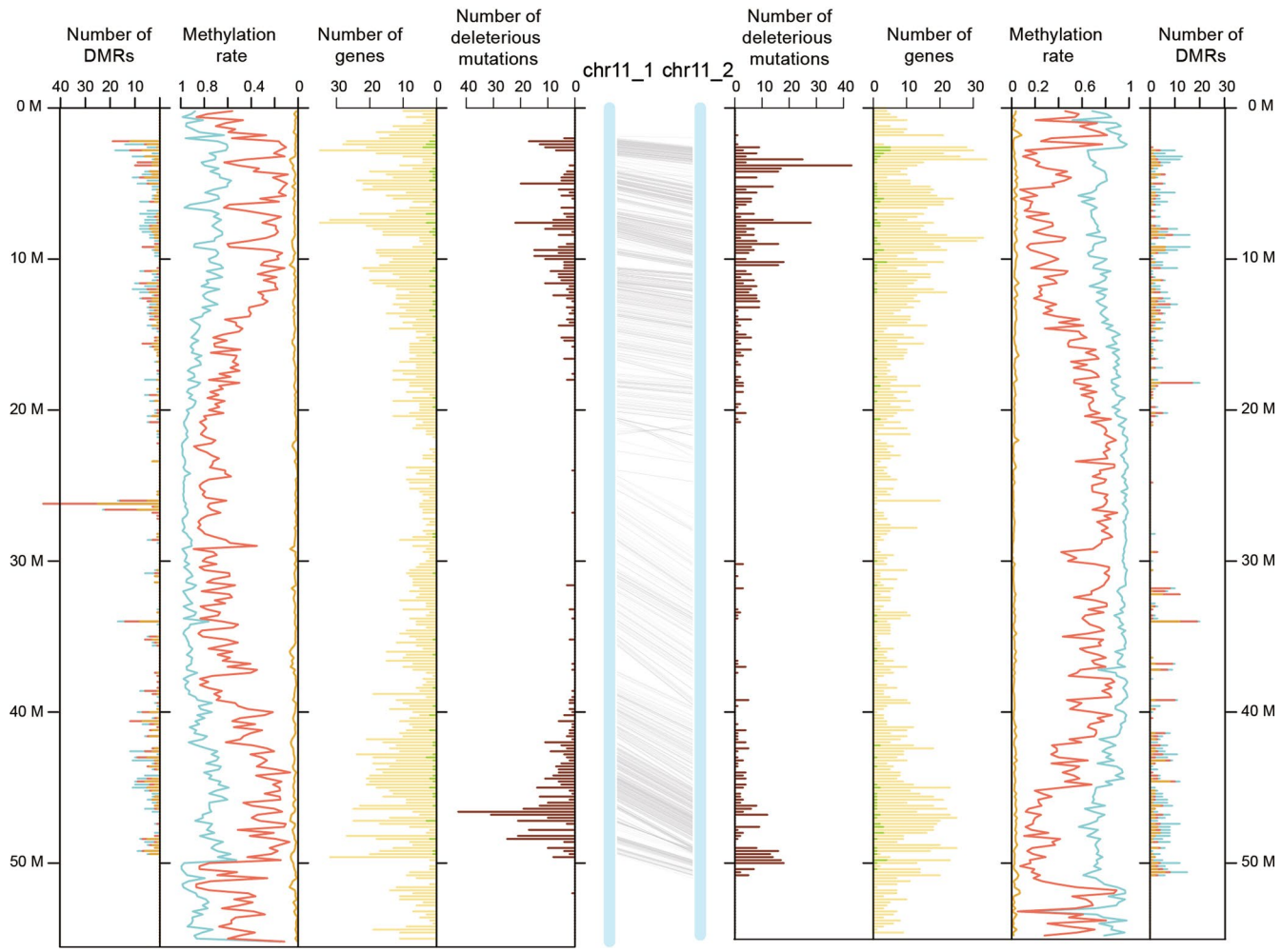
Extended Data Fig. 6 | Haplotype alignment of RH chromosome 8. The central blue bars represent the two haplotypes of chromosome 8 with the gray lines indicating the paired allelic genes. The distribution of deleterious or dysfunctional mutations (brown), annotated genes (yellow), preferentially expressed alleles (green), methylation level of three contexts and differentially methylated regions are arranged symmetrically for each haplotype. The methylation level and the number of DMRs of methylated sites in CG (light blue), CHG (red) and CHH (orange) contexts are indicated by cumulative column chart. Number of DMRs on one haplotype only involves the DMRs with hyper-methylation. All of the numbers were determined in 200 kb windows.



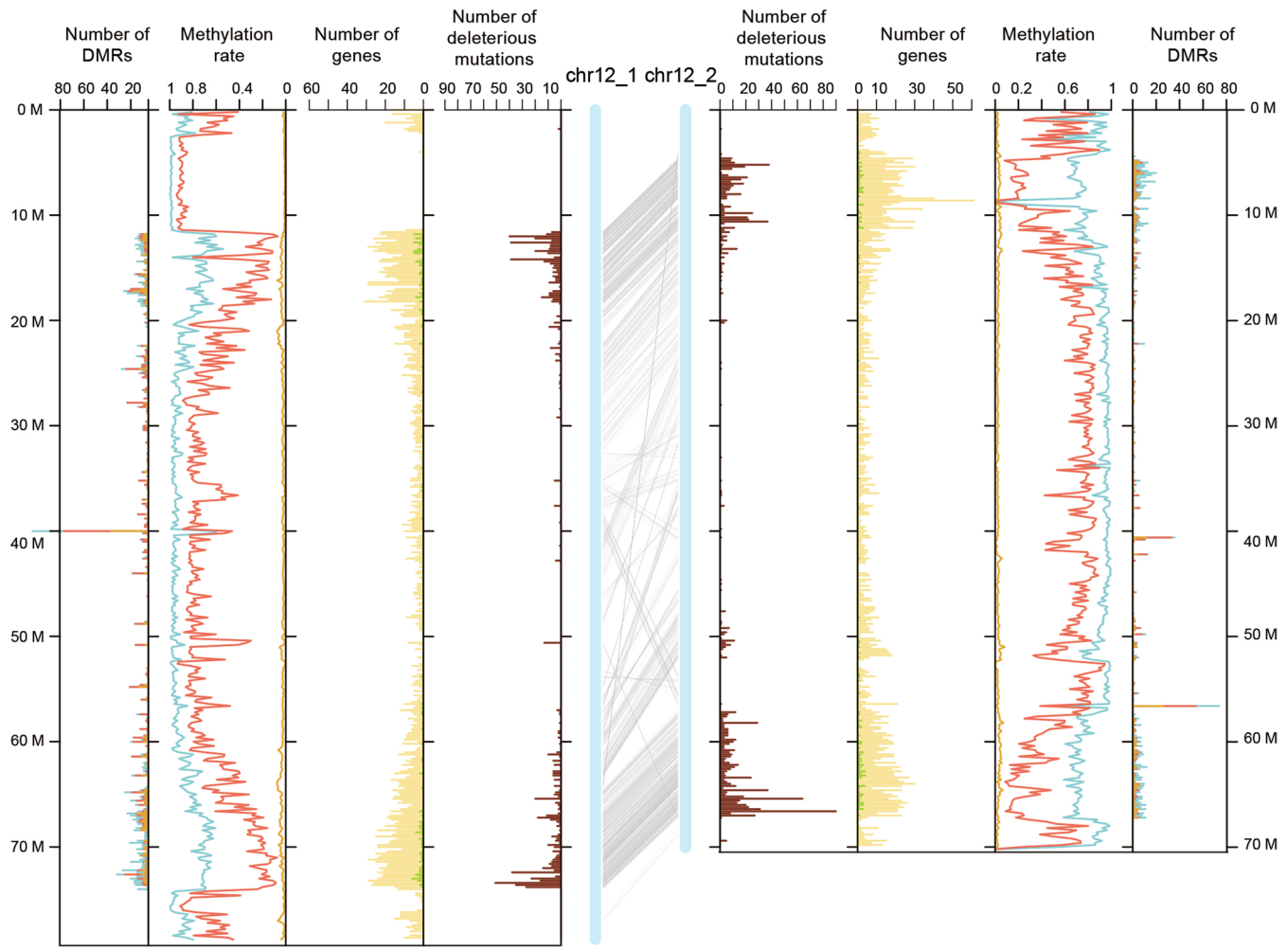
Extended Data Fig. 7 | Haplotype alignment of RH chromosome 9. The central blue bars represent the two haplotypes of chromosome 9 with the gray lines indicating the paired allelic genes. The distribution of deleterious or dysfunctional mutations (brown), annotated genes (yellow), preferentially expressed alleles (green), methylation level of three contexts and differentially methylated regions are arranged symmetrically for each haplotype. The methylation level and the number of DMRs of methylated sites in CG (light blue), CHG (red) and CHH (orange) contexts are indicated by cumulative column chart. Number of DMRs on one haplotype only involves the DMRs with hyper-methylation. All of the numbers were determined in 200 kb windows.



Extended Data Fig. 8 | Haplotype alignment of RH chromosome 10. The central blue bars represent the two haplotypes of chromosome 10 with the gray lines indicating the paired allelic genes. The distribution of deleterious or dysfunctional mutations (brown), annotated genes (yellow), preferentially expressed alleles (green), methylation level of three contexts and differentially methylated regions are arranged symmetrically for each haplotype. The methylation level and the number of DMRs of methylated sites in CG (light blue), CHG (red) and CHH (orange) contexts are indicated by cumulative column chart. Number of DMRs on one haplotype only involves the DMRs with hyper-methylation. All of the numbers were determined in 200 kb windows.



Extended Data Fig. 9 | Haplotype alignment of RH chromosome 11. The central blue bars represent the two haplotypes of chromosome 11 with the gray lines indicating the paired allelic genes. The distribution of deleterious or dysfunctional mutations (brown), annotated genes (yellow), preferentially expressed alleles (green), methylation level of three contexts and differentially methylated regions are arranged symmetrically for each haplotype. The methylation level and the number of DMRs of methylated sites in CG (light blue), CHG (red) and CHH (orange) contexts are indicated by cumulative column chart. Number of DMRs on one haplotype only involves the DMRs with hyper-methylation. All of the numbers were determined in 200 kb windows.



Extended Data Fig. 10 | Haplotype alignment of RH chromosome 12. The central blue bars represent the two haplotypes of chromosome 12 with the gray lines indicating the paired allelic genes. The distribution of deleterious or dysfunctional mutations (brown), annotated genes (yellow), preferentially expressed alleles (green), methylation level of three contexts and differentially methylated regions are arranged symmetrically for each haplotype. The methylation level and the number of DMRs of methylated sites in CG (light blue), CHG (red) and CHH (orange) contexts are indicated by cumulative column chart. Number of DMRs on one haplotype only involves the DMRs with hyper-methylation. All of the numbers were determined in 200 kb windows.

nature research

Corresponding author(s): Sanwen Huang

Last updated by author(s): Jul 30, 2020

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data.
Data analysis	All of softwares used in this study have been fully described including specifying versions in the Methods. The published tools have been cited and the open source has been described with the website links. The custom pipelines and scripts used in the project are deposited in GitHub https://github.com/zhouqiansolab/Haplotype-resolved-potato-genome.

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

Data

Policy information about availability of data

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

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

The genome assembly, annotation and a genome browser are available at SpudDB (http://solanaceae.plantbiology.msu.edu). This Whole Genome Shotgun project has been deposited at GenBank under the accession JACDXL000000000 and JACDXM000000000. The raw sequencing data has been deposited in the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra), with BioProject accession number PRJNA573826. The data has also been submitted to the Chinese National Genomics Data Center (https://bigd.big.ac.cn), with accession number CRA002005.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	For genome assembly and comparative analysis of haplotypes, a single sample of RH is enough. For genetic groups construction, we applied 880 selfed progeny and for fine mapping of genes. To locate the candidate genes for PA1 and WS1, we applied another 1200 progeny, which provides sufficient segregations in genetic mapping.
Data exclusions	No data was excluded.
Replication	For transcriptome sequencing, methylome sequencing and RT-qPCR analysis, the analyses were conducted on three replicates for each tissue.
Randomization	The sequencing samples of RH were collected randomly from the cultured seedlings and plants grown in green house.
Blinding	A blinded-experiment is not needed in this study because there is no comparing analysis between different groups.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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