bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar

Hequan Sun[1,2,∀], Wen-Biao Jiao[1,2,3,4,∀], Kristin Krause[1,2], José A. Campoy[2], Manish Goel[1,2], Kat Folz-Donahue[5], Christian Kukat[5], Bruno Huettel[6], and Korbinian Schneeberger[1,2,*]

[∀] These authors contributed equally.

[1]Faculty of Biology, LMU Munich, Großhaderner Str. 2, 82152 Planegg-Martinsried, Germany [2]Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany [3]Key Laboratory of Horticultural Plant Biology (Ministry of Education), Huazhong Agricultural University, 430070 Wuhan, China; [4]College of Informatics, Huazhong Agricultural University, 430070 Wuhan, China [5]FACS & Imaging Core Facility, Max Planck Institute for Biology of Ageing, 50931 Cologne, Germany [6]Max Planck-Genome-center Cologne, Carl-von-Linné-Weg 10, 50829 Cologne, Germany

*Correspondence: Korbinian Schneeberger (schneeberger@mpipz.mpg.de)

Key words: tetraploid potato, haplotyping, *de novo* assembly, single-cell sequencing, gamete

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Potato is the third most important food crop in the world. Despite its social and economic importance, the autotetraploid genome of cultivated potato has not been assembled yet. The distinct reconstruction all of four haplotypes remained an unsolved challenge. Here, we report the 3.1 Gb haplotype-resolved, chromosome-scale assembly of the autotetraploid potato cultivar, *Otava*. We assembled the genome with high-quality long reads coupled with single-cell sequencing of 717 pollen genomes and chromosome conformation capture data at a haplotyping precision of 99.6%. Unexpectedly, we found that almost 50% of the tetraploid genome were identical-by-descent with at least one of the other haplotypes. This high level of inbreeding contrasted with the extreme level of structural rearrangements encompassing nearly 20% of the genome. Overall, we annotated 148,577 gene models, where only 54% of the genes were present in all four haplotypes with an average of 3.2 copies per gene. Our work showcases how accurate assemblies of complex and partially inbred autotetraploid genomes can be generated. The newly established resource gives novel insights in the breeding history of autotetraploid potato and has the potential to change the future of genomics-assisted potato breeding.

Potato (*Solanum tuberosum* L.) is by far the most important tuber crop and is among the five most produced crops in the world. Globally more than 350 billion kilograms of potato are produced per year with an increasing trend particularly in developing countries in Asia[1]. Despite the social and economic importance, the breeding success of potato remained low over the past decades due to its highly heterozygous and tetraploid genome, which challenges usual breeding commonly applied to inbred, diploid crops[2,3].

A fundamental tool for modern breeding is the availability of reference sequences. The reference sequence for potato was generated from a double haploid plant, *DM1-3 516 R44* (*DM*), and was initially published in 2011[4] and continuously improved over the past years including a recent update based on long read sequencing[5]. Another major advancement in potato genomics was the recent assembly of a heterozygous diploid potato, *RH89-039-16* (*RH*)[6]. This haplotype-resolved genome was generated from a variety of

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

different sequencing technologies and phase information from a genetic map derived from selfed progeny[6].

However, as of now, there is no haplotype-resolved assembly of a tetraploid potato cultivar available. The latest methods for haplotype phasing include haplotype-based separation of sequencing reads based on the differences between the parental genomes[7] or based on haplotype information derived from gamete[8-11] or offspring genomes[6,12]. Similarly, chromosome conformation capture sequencing (e.g., Hi-C) could help to resolve haplotypes during or before the assembly[13-17] and has been applied to polyploids already[14,15,16]. But even though straightforward in its application, chromosome conformation capture sequencing can lead to haplotype switch errors, and requires additional efforts such genetic maps for correction[6,8,17].

Genome assembly of a tetraploid potato

We generated an assembly of the autotetraploid genome of *S. tuberosum*, cultivar *Otava*, using high-quality long PacBio HiFi reads (30x per haplotype) using *hifiasm*[18] (Fig. 1a; Supplementary Table 1; Supplementary Figure 1-4; Methods). The initial assembly consisted of 6,366 contigs with an N50 of 2.1 Mb. While the total assembly size of 2.2 Gb was much larger than the estimated haploid genome size of ~840 Mb, it accounted only for ~65% of the tetraploid genome size indicating that one third of the genome collapsed during the assembly (Supplementary Figure 2). A sequencing depth histogram across the contigs featured four distinct peaks, which originated from regions with either one, two, three, or four (collapsed) haplotype(s) (Fig. 1b). While most of the contigs represented only one haplotype (referred to as *haplotigs*) and accounted for 1.5 Gb (68%) of the assembly, contigs representing two, three or even four collapsed haplotypes (referred to as *diplotigs*, *triplotigs* or *tetraplotigs*) still made up 470 Mb (21%), 173 Mb (8%) or 43 Mb (2%). Regions with even higher coverages were virtually absent (9.4 Mb, 0.4%).

As there is no straight forward solution to untangle collapsed contigs after the assembly, we restarted the genome assembly, this time with four separated read sets each derived from one of the four haplotypes. In diploids, such a read separation prior to the

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

assembly can be performed by sorting the reads according to their similarity to the parental genomes (trio binning)[7]. But as autotetraploid individuals inherit two haplotypes through both the maternal and paternal lineages, this cannot be applied for autotetraploid genomes. Alternatively, the reads can also be separated using the haplotypes found in gamete genomes (gamete binning)[8]. While this is straightforward with haploid gametes from diploid individuals, tetraploid potato develops diploid gametes, which again does not separate individual haplotypes. However, as the pairing of the two haplotypes in a diploid gamete is random in potato, we speculated that it might be possible to gain information on individual haplotypes (and thus to separate the reads into four distinct sets) if we sequence a sufficient number of diploid gametes.

To test if gamete binning can be applied for the genome assembly of *Otava*, we sequenced the genomes from 717 pollen nuclei with Illumina short reads with an average sequence coverage of 0.18x (Supplementary Figure 5; Methods) and aligned each of the 717 read sets against the initial assembly. As defining a high-density SNP list can be difficult in an autotetraploid genome, we defined “coverage markers” (using average alignment depth in 50 kb windows) to assess if a genomic region was present in a pollen genome or not (Methods).

A coverage marker will be covered by reads if one of the two haplotypes of a pollen carries the region of the coverage marker. With this, we could assess the presence/absence pattern (PAP) of a coverage marker across all the 717 pollen genomes (Supplementary Figure 6). Closely linked markers featured highly similar PAPs, as most pollen genomes carried the same pair of haplotypes at two neighboring loci. We used similarities between PAPs to cluster the contigs into 48 groups representing the four haplotypes of all 12 chromosomes (Supplementary Figure 7-8; Methods). Haplotigs were assigned to single clusters. Diplotigs, triplotigs and tetraplotigs represented multiple haplotypes and were assigned to two, three or four of the clusters (Methods).

Once the contigs were assigned to haplotypes, also the PacBio HiFi reads could be assigned to these haplotypes based on their alignments against the contigs. Reads aligned

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

to diplotigs, triplotigs or tetraplotigs were randomly assigned to one of the respective haplotypes. With this, more than 99.9% of the non-organellar PacBio HiFi reads could be assigned to one of the 48 read sets (Supplementary Figure 9; Methods). Assembling the read sets using *hifiasm* resulted in 48 haplotype-resolved assemblies with an average N50 of 7.1 Mb and a total size of 3.1 Gb (92% of the tetraploid genome). Finally, we used Hi-C short read data (70x per haplotype) to scaffold the contigs of each assembly to a chromosome-scale, haplotype-resolved assembly (Supplementary Figure 10; Methods).

The sizes of the four haplotypes of each chromosome were highly consistent to each other as well as to those of the *DM* and *RH* assemblies[4,5,6] except for the consistently shorter assemblies of LG10 (Fig. 1c). Comparison with the *DM* assembly showed high levels of synteny suggesting that the chromosomes have been assembled correctly (Supplementary Figure 11-12). To evaluate the haplotyping accuracy of the tetraploid assembly in more depth, we sequenced the parental cultivars of *Otava*, called *Stieglitz* and *Hera*, with Illumina short reads with 40x genome coverage. Comparing the parental genome specific *k*-mers, we found that each of the 48 assemblies included almost exclusively *k*-mers from one or the other parent implying a haplotyping accuracy of 99.6% (Fig. 1d; Methods).

Integrating *ab initio* predictions, protein and RNA-seq read[4,5,6] alignments, we annotated 148,577 gene models across all haplotypes with an overall *BUSCO*[19] completeness score of 97.3%, which is highly comparable to the annotations of the *RH* and *DM* assemblies[5,6] (Supplementary Table 2-3; Methods). Repetitive sequences made up 66% of the assembly with LTR retrotransposons as the most abundant class and rDNA clusters of up to 600 kb in size, which were assembled without any gaps (Supplementary Table 4-5; Methods). The distribution of genes and repeats along the chromosome followed the typical distribution of plant genomes with high gene and low repeat densities at the distal parts of the chromosome, while in the peri-centromeric regions the gene densities were low and the repeat densities were high (Fig. 2).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The genomic footprints of inbreeding

A histogram of sequence differences within 10kb windows between the haplotypes revealed two separated peaks implying the presence of highly similar as well as highly different regions (Fig. 3a). The divergent regions averaged 1 SNP per 17 bp, while nearly 50% of the regions were without differences (Fig. 3a). This extreme similarity between some of the regions suggested that they were recently inherited from a common ancestor. In fact, the pedigree of many of the cultivated potatoes, including *Otava*, contain cultivars that occur more than once in their ancestry[20,21] (Supplementary Figure 1). Common ancestors in different lineages of the pedigree leads to inbreeding and results in regions which are identical-by-descent (IBD) between their haplotypes (Fig. 3b-c; Supplementary Figure 13-24; Methods).

Overall, almost 50% of the tetraploid genome of *Otava* were included in IBD blocks and were shared by either two, three or in rare cases even by four haplotypes (Fig. 3b-d). Individual IBD blocks varied in size and reached up to 41.6 Mb, while IBD blocks in the peri-centromeres were significantly larger as compared to the IBD blocks in the distal parts of the chromosomes (Supplementary Figure 13-25). Even though it is possible that long IDB blocks were recently introduced and were not broken up by meiotic recombination yet, it is more likely that these extremely long IBD blocks exist due to local suppression of meiotic recombination in the peri-centromeres (Fig. 3c). Using the accumulated mutation rates in the IBD blocks as an estimate of their age showed that long IBD blocks weren't younger as compared to short IBD blocks (Supplementary Figure 25b).

Extreme sequence differences and their influence on genes

The highly similar IBD blocks were contrasted by high levels of structural rearrangements in the non-shared regions of the genome (Fig. 3; Supplementary Figure 13-24; Methods). Inversions, duplications, and translocations made up 3.8% to 42.9% of each of the haplotypes (or 19.3% of the genome) depending on the abundance of IDB regions in the respective haplotypes (Fig. 3d). Excluding IBD regions, structural rearrangements made up 15.0% to 65.8% of each chromosome. In addition to these high levels of structural

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

differences, each haplotype included another 11.0% to 42.5% of unique sequence that could not be aligned to the other haplotypes (Fig. 3d). This amount of structural variation and haplotype-specific sequence was much higher than what has been reported for any other crop species, supporting earlier suggestions that wild introgressions were part of the domestication history of potato[22].

Overall, we found 661 structural variations longer than 100 kb which all were supported by the contiguity of the assembled contigs or Hi-C contact signals, including 220 duplications, 207 translocations and 234 inversions (Supplementary Table 6; Supplementary Fig. 10,13-24,26). While comparable in number, inversions were much larger than the other types of rearrangements and reached sizes of up to 12.4 Mb (Fig. 3f). Although these large inversions were mostly located in the peri-centromeric regions where genes occur at low density, they still harbored nearly 5% of all genes (7,958 out of 148,577). Meiotic crossover events within the pollen genomes were virtually absent in the inversions, indicating that these regions are likely to introduce large segregating haplotypes among cultivated potato (Fig. 3c).

Pairwise allelic divergence of the genes ranged from 0 to 140 differences per kb and included identical as well as divergent alleles. The average pairwise difference of the divergent alleles was 18 differences per kb (Fig. 4a). Moreover, due to the high sequence differences, only 53.6% of the genes were present in all four haplotypes. The remaining 46.4% of the genes were present in three (20.0%), two (15.9%) or even only one (10.5%) of the haplotypes (Fig. 4b) with an average of 3.2 copies per gene. In addition, the coding sequences of some of these copies were identical to each other. For example, only 3,066 (15.4%) of the genes with four copies also featured four distinct alleles. In consequence, even though each gene featured 3.2 copies, there were only 1.9 distinct alleles per gene (Fig. 4b).

While it was expected to find identical alleles within shared haplotypes, only ~45% of the identical alleles were actually within IDB blocks. To test if the high number of identical alleles between the otherwise different haplotypes was indicative of selection, we tested

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

whether the genes with identical alleles were enriched for specific functions. This revealed a significant enrichment for genes with GO terms involving photosynthesis, chlorophyll binding and translation (Fig. 4c) suggesting a selection-induced loss of allelic diversity through the optimization of plant performance.

The non-functional alleles of the genes were randomly distributed throughout the genome implying that a ploidy reduction of the tetraploid genome would lead to a significant gene loss**.** In fact, the doubled-monoploid *DM*[1,5] or diploid *RH*[6], which both were derived from tetraploid cultivars, carried 5,901 (15.8%) or 3,245 (8.7%) less genes as compared to the tetraploid genome. The gene family with highest percentage of genes with presence/absence variation (45.4%; 316 out of 696 genes) were the NLR resistance genes (Supplementary Table 7), which are known for their high intraspecies variability[23,24].

Conclusions

Here we reported the first haplotype-resolved assembly of an autotetraploid potato. Leveraging high-quality, long reads and single-cell genotyping of diploid gametes, we were able to reconstruct the sequences of all four haplotypes. This revealed the high levels of structural variation between the haplotypes, which were much higher as compared to the diversity commonly found within species. This supported earlier suggestions that the diverse haplotypes might have been introgressed from wild species during domestication[22].

The high level of sequence differences was contrasted by widespread IBD blocks, which were most likely introduced by the common usage of related genotypes during cultivation, even though we cannot exclude that some of these blocks might have been formed via double reduction during meiosis[25]. The similarity of the IBD blocks was the reason for the abundant collapsed regions in the initial assembly. As these regions were almost identical, it was not possible to assemble them from the sequence data alone. IBD blocks are a widespread phenomenon in many crops or livestock in general, though the challenges associated with the high similarity between haplotypes can be solved by using the power of genetics and analyzing individual gamete genomes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The abundance of IBD blocks also implied that the maximal allelic diversity of the tetraploid genome was not reached, even though the high yield and yield stability of potato is supposed to be promoted by the effects of heterosis, which itself is based on non-additive interactions of diverse alleles[26]. Whether the high abundance of shared alleles suggests that the effects of heterosis could still be optimized by increasing the number of polymorphic alleles or if this indicates that the limits of heterosis were already reached remains to be seen.

Over the past years, considerable success has been made in re-domesticating potato from a clonally-propagated, tetraploid crop into a seed-propagated, diploid crop to increase reproduction rate, decrease costs in storage and transportation, and improve disease control[2,27,28,29]. However, the random distribution of loss-of-function alleles in tetraploid potato can lead to the accelerated manifestation of inbreeding depression in the diploid genomes, when they are derived from tetraploids[6,30]. Haplotype-resolved assemblies of autotetraploids like the one presented here have the potential to support the design of optimal haplotypes by avoiding the combination of known incompatibility alleles[31].

Of course, this new possibility to assemble autotetraploid genomes does not eliminate all breeding-related problems that result from the tetraploid nature of potato. However, being able to reconstruct the four haplotypes of cultivated potato is a breakthrough for modern genomics-assisted breeding strategies, and ultimately has the power to increase the breeding success of potato in the future.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Methods

Plant material was grown at Max Planck Institute for Plant Breeding Research (Cologne, Germany). The genome of *Otava* was sequenced with PacBio HiFi Sequel II platform with four SMRTcells. DNA extracted from individual pollen nuclei was prepared with 10x Genomics CNV kits and subsequently sequenced with Illumina sequencing. Barcodes were corrected using *cellranger* (10x Genomics). Short/long reads were aligned using *bowtie2*[32]/*minimap2*[33]. BAM, VCF file processing and sequencing depth analysis were performed using *samtools*[34] and *bedtools*[35]. PacBio sequence reads were assembled using *hifiasm*[18], and genome annotation was performed following a previous pipeline[8]. Structural variations were identified using *SyRI*[36] based on *minimap2* genome alignments. More details and other related methods are provided in the Supplementary Information.

Acknowledgements

The authors would like to thank Christiane Gebhardt (MPI-PZ, Cologne, Germany) and Benjamin Stich (HHU, Düsseldorf, Germany) for helpful discussions, Birgit Walkemeier and Christine Sänger (both MPI-PZ, Cologne, Germany) for help with plant cultivation, Christine Brandt and Klaus J. Dehmer (both IPK, Groß Lüsewitz, Germany) for providing material, Pádraic J. Flood (WUR, Wageningen, The Netherlands) for comments on the manuscript as well as Saurabh Pophaly (MPI-PZ, Cologne, Germany) for help in data management. This work was funded by the “Humboldt Research Fellowship for Experienced Researchers” (Alexander von Humboldt Foundation) (J.A.C.), the Marie Skłodowska-Curie Individual Fellowship PrunMut (789673) (J.A.C.), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2048/1–390686111, and the European Research Council (ERC) Grant “INTERACT” (802629) (K.S.).

Author contributions

H.S. and K.S. developed the project. K.K., J.A.C., K.F-D., C.K. and B.H. generated data. H.S., W-B.J., and M.G. performed all data analysis. H.S. and K.S. wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Competing interests

The authors declare no competing interests.

Data availability

High-throughput sequencing data as well as the genome assembly and gene annotation of *Otava* will be made available through NCBI upon publication of this work under Bioproject PRJNA726019.

Code availability

Upon publication customed scripts supporting this work will be available at *github.com/schneeberger-lab/GameteBinning_tetraploid*.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

REFERENCES

1. The Food and Agriculture Organization (FAO). http://www.fao.org/faostat/en/#data/QV (2021).
2. Jansky, S. H. et al. Reinventing potato as a diploid inbred line-based crop. *Crop Sci.* **56**, 1412-1422 (2016).
3. Douches, D.S., Maas, D., Jastrzebski, K., Chase, R. W. Assessment of Potato Breeding Progress in the USA over the Last Century. *Crop Sci*. **36**,1544-1552 (1996).
4. The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-195 (2011).
5. Pham, G.M. & Hamilton, J.P. *et al*. Construction of a chromosome-scale long-read reference genome assembly for potato. *GigaScience* **9**, 1-11 (2020).
6. Zhou, Q. & Tang, D. *et al*. Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat. Genet.* **52**, 1018-1023 (2020).
7. Koren, S. & Rhie, A. *et al*. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174-1182 (2018).
8. Campoy, J.A. & Sun, H.Q. *et al*. *Gamete binning*: chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes. *Genome Biol.* **21**, 306 (2020).
9. Li, R. & Qu, H. *et al*. Inference of chromosome-length haplotypes using genomic data of three or a few more single gametes. *Mol Biol Evol.* **37**, 3684-3698 (2020).
10. Kirkness, E.F. *et al*. Sequencing of isolated sperm cells for direct haplotyping of a human genome. *Genome Res.* **23**, 826-832 (2013).
11. Shi, D., Wu, J., Tang, H. & Yin, H. *et al*. Single-pollen-cell sequencing for gamete-based phased diploid genome assembly in plants. *Genome Res.* **29**, 1889-1899 (2019).
12. Zhou, C. *et al*. Assembly of whole-chromosome pseudomolecules for polyploid plant genomes using outbred mapping populations. *Nat. Genet*. **52**, 1256-1264 (2020).
13. Garg, S. *et al.* Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

14. Zhang, J., Zhang, X., Tang, H. & Zhang, Q. *et al*. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* **50**, 1565-1573 (2018).
15. Chen, H., Zeng, Y., Yang, Y., Huang, L., Tang, B. & Zhang, H. et al. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat Commun* **11**, *2494* (2020).
16. 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).
17. Linsmith, G. *et al*. Pseudo-chromosome-length genome assembly of a double haploid "*Bartlett*" pear (*Pyrus communis L.*). *GigaScience* **8**, 1-17 (2019).
18. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li, H. Haplotype-resolved de novo assembly with phased assembly graphs. *Nat Methods* **18**, 170-175 (2021).
19. Simão, F. A. and Waterhouse, R. M. *et al*. *BUSCO*: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).
20. Hutten, R.C.B. and Berloo, R. van. An online potato pedigree database. URL: http://www.plantbreeding.wur.nl/PotatoPedigree/ (2001).
21. Berloo, R. van, Hutten, R.C.B, Eck, H.J. van and Visser, R.G.F. An online potato pedigree database resource. *Potato research* **50**, 45-57 (2007).
22. Hardigan, M.A., Laimbeer, F.P.E., Newton, L., Crisovan, E., Hamilton, J.P., Vaillancourt, B. et al. Genome diversity of tuber-bearing Solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proc Natl Acad Sci U S A* **114**, E9999-E10008 (2017). doi: 10.1073/pnas.1714380114.
23. Van de Weyer, A.L., Monteiro, F., Furzer, O.J., et al. A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana. *Cell*. **178**, 1260-1272 (2019).
24. Seong, K., Seo, E., Witek, K., Li, M., Staskawicz, B. Evolution of NLR resistance genes with noncanonical N-terminal domains in wild tomato species. *New Phytol.* **227**, 1530-1543 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

25. Bourke, P.M., Voorrips, R.E., Visser, R.G., Maliepaard, C. The double-reduction landscape in tetraploid potato as revealed by a high-density linkage map. *Genetics*, 201:853-863 (2015).

26. J. Muthoni, H. Shimelis, R. Melis. Production of hybrid potatoes: Are heterozygosity and ploidy levels important? *Australian Journal of Crop Science* **13**, 687-694 (2019).

27. Lindhout, P. *et al*. Towards F1 Hybrid Seed Potato Breeding. *Potato Res.* **54**, 301-312 (2011).

28. Ye, M. & Peng, Z. et al. Generation of self-compatible diploid potato by knockout of S-RNase. *Nature Plants* **4**, 651-654 (2018).

29. Li, Y., Li, G., Li, C., Qu, D. & Huang, S. Prospects of diploid hybrid breeding in potato. *Chin. Potato J.* **27**, 96-99 (2013).

30. Zhang, C., Wang, P., Tang, D. et al. The genetic basis of inbreeding depression in potato. *Nat Genet.* **51**, 374-378 (2019).

31. Lian. Q., Tang, D., Bai, Z., Qi, J., Lu, F., Huang, S., Zhang, C. Acquisition of deleterious mutations during potato polyploidization. *J Integr Plant Biol.* **61**, 7-11 (2019).

32. Langmead, B., and Salzberg, S. L. Fast gapped-read alignment with *Bowtie* 2. *Nature methods* **9**, 357-359 (2012).

33. Li, H. *Minimap2*: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(18), 3094-100 (2018).

34. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

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

36. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. *SyRI*: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 1-13 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.15.444292; this version posted May 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

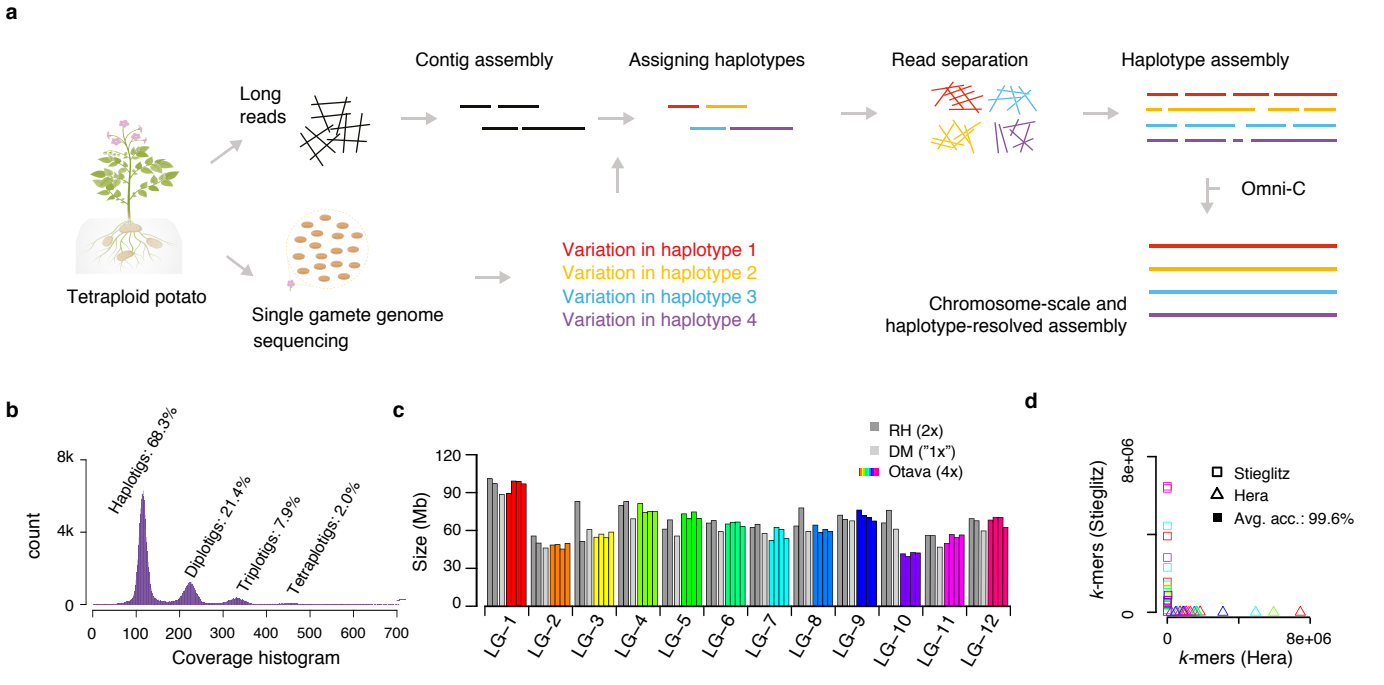
Figure legends

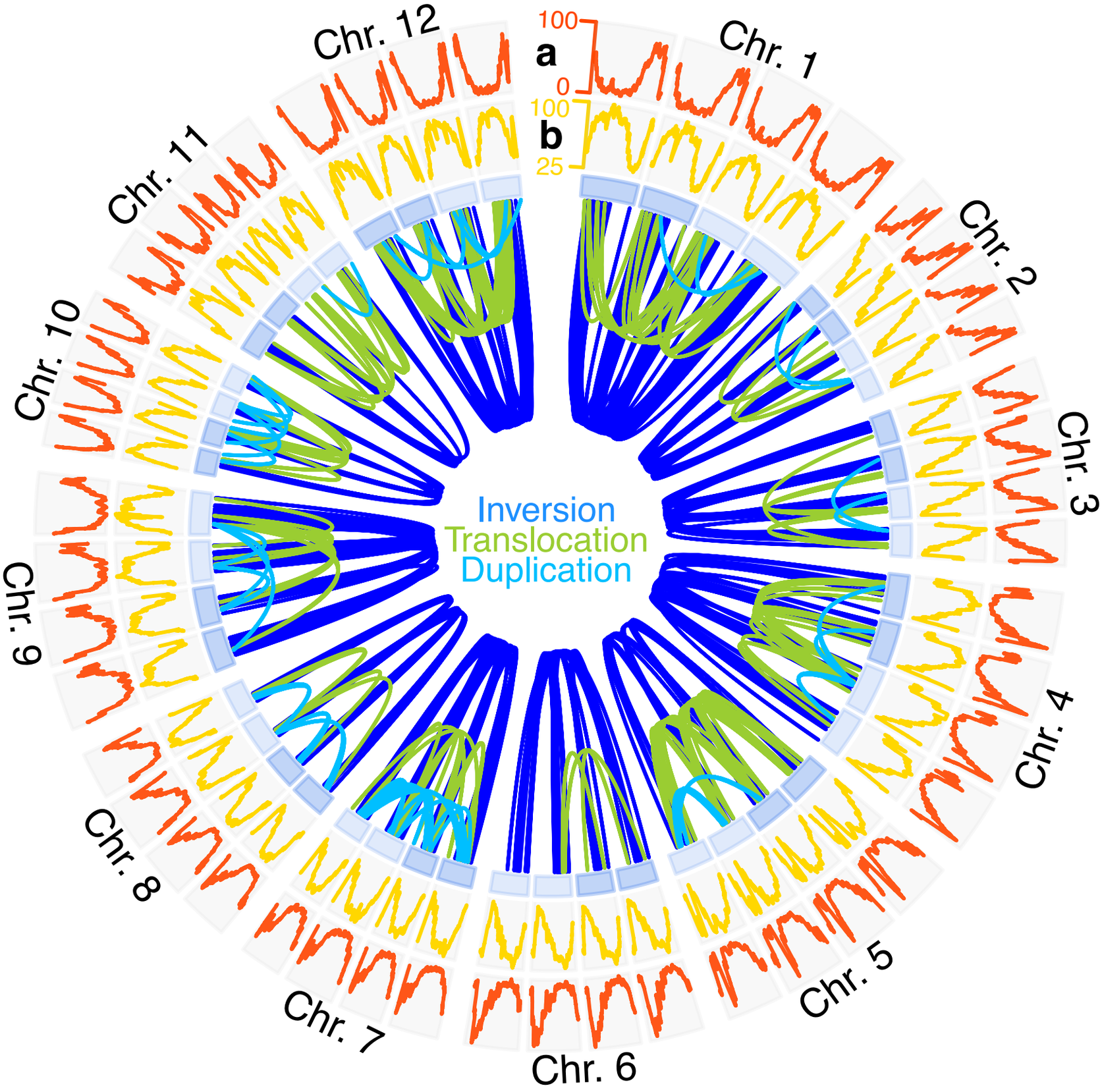
Figure 1. Haplotype-resolved assembly of an autotetraploid potato genome. a. Assembly strategy (gamete binning) for tetraploid genomes. Long reads are sequenced from somatic DNA and an initial contig-level assembly is generated. In addition, sequencing data of gamete genomes are generated. Genetic linkage enables grouping of the contig into clusters, which represent the individual haplotypes. Long reads are assigned to haplotypes based on their similarity to the contigs. Each haplotype can be assembled separately and scaffolded to chromosome-scale using Hi-C. (The figure was created with help of BioRender.com.) **b.** Histogram of sequencing depth within 10 kb windows of the initial assembly revealed existence of haplotigs (68.3%), diplotigs (21.4%), triplotigs (7.9%) and tetraplotigs (2.0%). **c**. Assembly sizes of the haplotypes were highly consistent to the *DM*[5] and *RH*[6] assemblies. **d**. *k*-mer based evaluation of the haplotyping accuracy. Each point represents one haplotype and indicates the numbers of *k*-mers specific to one of the parental genomes *Hera* or *Stieglitz*. Overall, 99.6% of the variation were correctly phased.

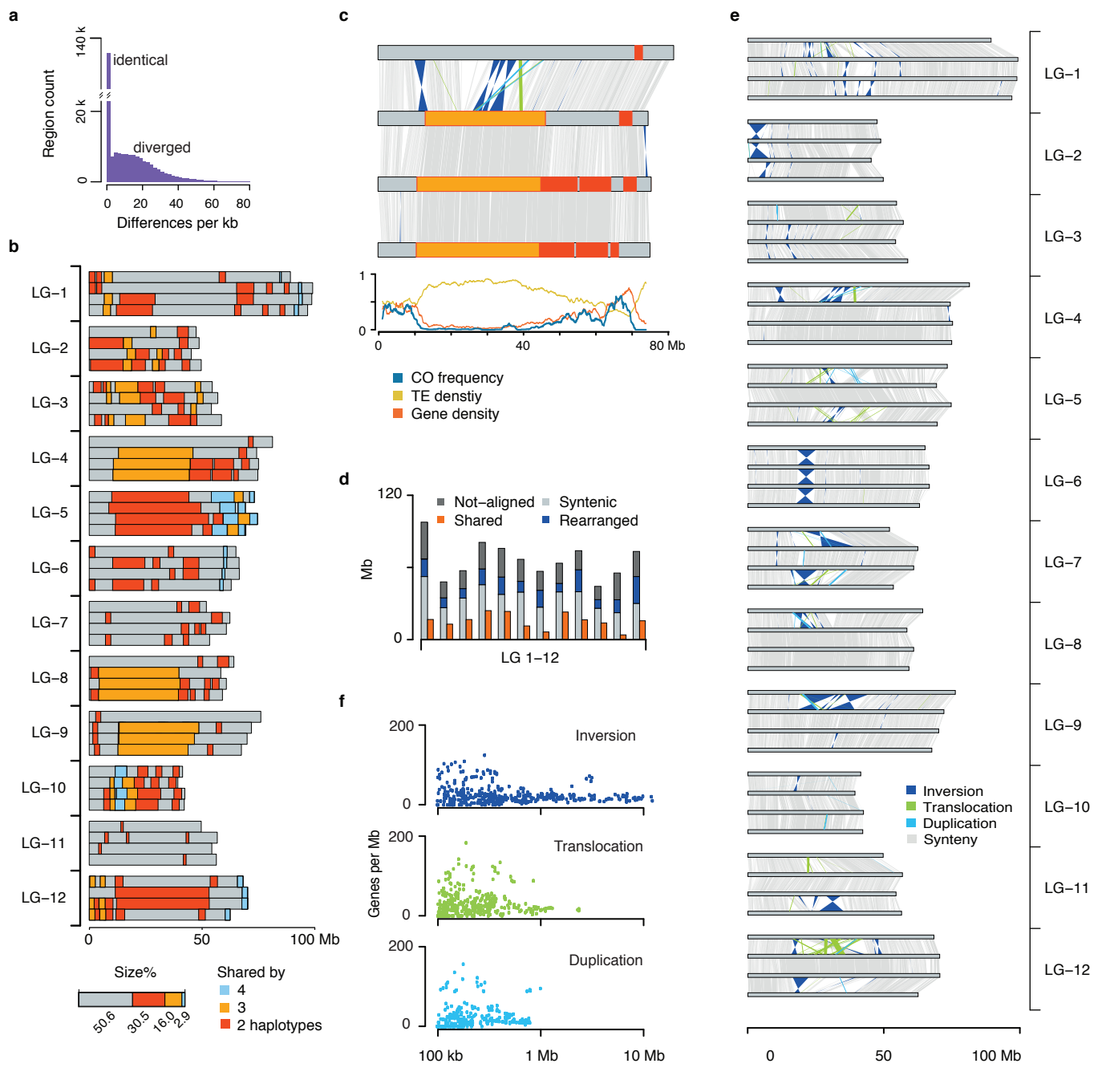
Figure 2. Haplotype-resolved and chromosome-scale assembly of the tetraploid potato cultivar *Otava*. a. Gene density (number of genes within 2 Mb windows) along the four haplotypes of each of the 12 chromosomes. **b**. Percentage of transposable element (TE) related sequence within 2 Mb windows. The links in the center show over 600 structural rearrangements larger than 100 kb found between the four haplotypes of each chromosome. Light and dark blue box refer to the maternally and paternally inherited chromosomes.

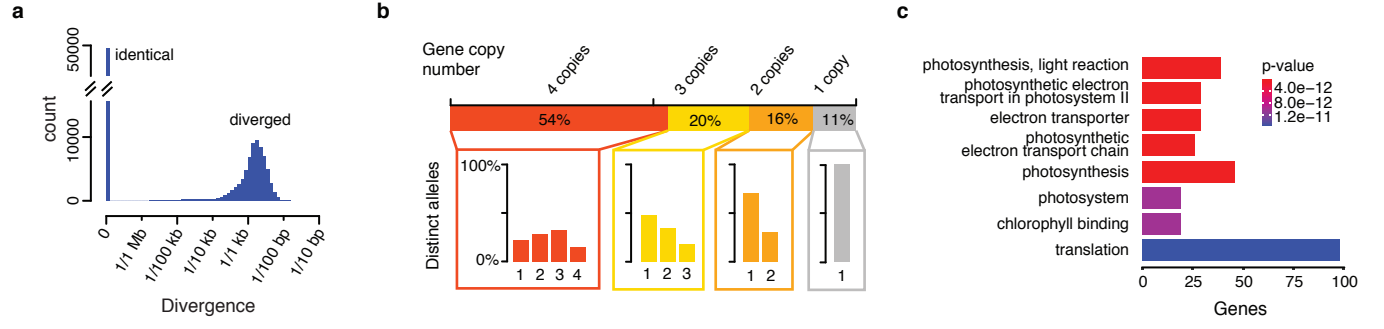
Figure 3. Haplotype analysis of the tetraploid genome. a. SNP density as observed in pairwise comparisons between the haplotypes revealed two separated peaks. The high abundance of highly similar regions suggested the existence of identical-by-descent (IBD) blocks. **b.** IBD blocks across the genome. Regions shared by two, three or four haplotypes are colored in red, orange or blue. **c.** A zoom-in on the IBD regions and structural rearrangements of LG-4. Large IBD regions were more likely to occur in peri-centromeric regions with low gene, but high TE content and suppressed meiotic recombination. (Colors as defined in **b** and **e**) **d.** Average alignment statistics and structural rearrangements in each chromosome. **e.** Structural rearrangements between the four haplotypes of each LG. **f.** Correlation of the size of 220 duplications, 207 translocations and 234 inversions with gene density.

Figure 4. Impact of haplotype divergence on genes. a. Pair-wise allelic divergence of genes. **b.** Presence/absence variations of genes. Overall, 53.6%, 20.0%, 15.9% or 10.5% of the genes showed four, three, two or one allele(s) within the tetraploid genome with an average of 3.2 allelic copies per gene. Within the genes with four allelic copies, one (22.3%), two (29.8%), three (32.5%) or four (15.4%) divergent allele/s were observed. **c.** GO enrichment analysis of genes with four identical alleles.


a
Contig assembly
Assigning haplotypes
Read separation
Haplotype assembly
Long reads
Tetraploid potato
Single gamete genome sequencing
Variation in haplotype 1
Variation in haplotype 2
Variation in haplotype 3
Variation in haplotype 4
Omni-C
Chromosome-scale and haplotype-resolved assembly
b
Haplotigs: 68.3%
Diplotigs: 21.4%
Triplotigs: 7.9%
Tetraplotigs: 2.0%
count
Coverage histogram
c
RH (2x)
DM ("1x")
Otava (4x)
Size (Mb)
LG–1
LG–2
LG–3
LG–4
LG–5
LG–6
LG–7
LG–8
LG–9
LG–10
LG–11
LG–12
d
Stieglitz
Hera
Avg. acc.: 99.6%
k-mers (Stieglitz)
k-mers (Hera)


Chr. 1
Chr. 2
Chr. 3
Chr. 4
Chr. 5
Chr. 6
Chr. 7
Chr. 8
Chr. 9
Chr. 10
Chr. 11
Chr. 12
100
a
0
100
b
25
Inversion
Translocation
Duplication


a
Region count
140 k
20 k
0
identical
diverged
0
20
40
60
80
Differences per kb
b
LG–1
LG–2
LG–3
LG–4
LG–5
LG–6
LG–7
LG–8
LG–9
LG–10
LG–11
LG–12
0
50
100 Mb
Size%
50.6
30.5
16.0
2.9
Shared by
4
3
2 haplotypes
c
1
0
0
40
80 Mb
CO frequency
TE denstiy
Gene density
d
120
0
Mb
Not–aligned
Shared
Syntenic
Rearranged
LG 1–12
f
200
0
Inversion
Translocation
Duplication
Genes per Mb
100 kb
1 Mb
10 Mb
e
LG–1
LG–2
LG–3
LG–4
LG–5
LG–6
LG–7
LG–8
LG–9
LG–10
LG–11
LG–12
Inversion
Translocation
Duplication
Synteny
0
50
100 Mb


a
50000
10000
0
count
identical
diverged
0
1/1 Mb
1/100 kb
1/10 kb
1/1 kb
1/100 bp
1/10 bp
Divergence
b
Gene copy number
4 copies
3 copies
2 copies
1 copy
54%
20%
16%
11%
Distinct alleles
100%
0%
1 2 3 4
1 2 3
1 2
1
c
photosynthesis, light reaction
photosynthetic electron transport in photosystem II
electron transporter
photosynthetic electron transport chain
photosynthesis
photosystem
chlorophyll binding
translation
p-value
4.0e−12
8.0e−12
1.2e−11
0
25
50
75
100
Genes