nature plants

ARTICLES

DOI: 10.1038/s41477-017-0002-z

Haplotype-resolved sweet potato genome traces back its hexaploidization history

Jun Yang[1,2,4], M-Hossein Moeinzadeh[2], Heiner Kuhl[6], Johannes Helmuth[7], Peng Xiao[2], Stefan Haas[2], Guiling Liu[5], Jianli Zheng[5], Zhe Sun[5], Weijuan Fan[3], Gaifang Deng[3], Hongxia Wang[3], Fenhong Hu[3], Shanshan Zhao[1], Alisdair R. Fernie[4], Stefan Boerno[6], Bernd Timmermann[6], Peng Zhang[3*] and Martin Vingron[2*]

Here we present the 15 pseudochromosomes of sweet potato, *Ipomoea batatas*, the seventh most important crop in the world and the fourth most significant in China. By using a novel haplotyping method based on genome assembly, we have produced a half haplotype-resolved genome from ~296 Gb of paired-end sequence reads amounting to roughly 67-fold coverage. By phylogenetic tree analysis of homologous chromosomes, it was possible to estimate the time of two recent whole-genome duplication events as occurring about 0.8 and 0.5 million years ago. This half haplotype-resolved hexaploid genome represents the first successful attempt to investigate the complexity of chromosome sequence composition directly in a polyploid genome, using sequencing of the polyploid organism itself rather than any of its simplified proxy relatives. Adaptation and application of our approach should provide higher resolution in future genomic structure investigations, especially for similarly complex genomes.

With a consistent global annual production of more than 100 million tons, as recorded between 1965 and 2014 (Food and Agriculture Organization of the United Nations)[1], the sweet potato, *I. batatas*, is an important source of calories, proteins, vitamins and minerals for humanity. It is the seventh most important crop in the world and the fourth most significant crop of China. In periods of shortages of basic cereal foods, *I. batatas* frequently served as the main food source for many Chinese people. It rescued millions of lives during and up to 3 years after the Great Chinese Famine in the 1960s and was subsequently raised as a main guarantor of food security in China.

Although the sweet potato is an outstanding crop, its genome has not yet been sequenced. The reason, at least in part, is that the genome has proved very difficult to assemble, being hexaploid ($2n=6x=90$) and highly polymorphic; with a base chromosome number of 15 and a genome size of about 4.4 Gb, as estimated from its measured *C*-value (half the amount of DNA contained within a somatic cell of sweet potato)[2]. Sweet potato has a composition of two B_1 and four B_2 component genomes ($B_1B_1B_2B_2B_2B_2$), as predicted by genetic linkage studies using random amplified polymorphic DNA and amplified fragment length polymorphism markers[3,4]. The degree of homology, however, could not be estimated with accuracy since its genomic components are still poorly characterized. Recently, the genome survey sequencing of *Ipomoea trifida*, the most probable diploid wild relative of *I. batatas*, has been reported[5]. Unfortunately, the current version of the *I. trifida* genome cannot serve as an adequate reference sequence for *I. batatas* because of its low N50 length statistics and even more because of the high abundance of gaps in the assembly, estimated at more than 30%. This example also points out the problems inherent in the usual circuitous tactics employed in polyploid genome sequencing projects, which always begin with simpler diploid relatives and assume no chromosomal structure variation between polyploidy and diploid wild relatives. For example, sequencing of the autotetraploid potato (*Solanum tuberosum*) employed a homozygous doubled-monoploid potato[6]. Allotetraploid Upland cotton (*Gossypium hirsutumdi*, AADD)[7] began with diploid *Gossypium raimondii* (DD)[8] and *Gossypium arboreum* (AA)[9]. Comparably, the sequencing of allohexaploid bread wheat (*Triticum aestivum*, AABBDD) was initiated by sequencing the genomes of *Triticum urartu* (AA)[10] and *Aegilops tauschii* (DD)[11] but is still struggling with the precise sequencing of isolated chromosome arms[12,13]. All in all, a more cost-effective strategy and one that promises more efficient output for the direct sequencing of complex polyploid genomes is needed, especially for plant scientists, since polyploidy is a frequent and naturally occurring state in plants. Although the full genetic implications of polyploidization are still obscure, it is clear that this state provides an important pathway for plant evolution and specialization. For the plant genome investigator, however, decoding polyploidy genome remains a major problem.

More specifically, de novo assembly of polyploid genomes remains a critical unsolved technical problem. The high heterozygosity

[1]Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Plant Science Research Center, Chinese Academy of Sciences, Shanghai Chenshan Botanical Garden, 3888 Chenhua Road, Shanghai 201602, China. [2]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, 14195 Berlin, Germany. [3]National Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China. [4]Department of Molecular Physiology, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam, Germany. [5]Tai'an Academy of Agricultural Sciences, 16 Tailai Road, Tai'an, 271000 Shandong, China. [6]Max Planck Institute for Molecular Genetics, Sequencing Core Facility, Ihnestraße 63-73, 14195 Berlin, Germany. [7]Max Planck Institute for Molecular Genetics, Otto-Warburg-Laboratory: Computational Epigenomics Group, Ihnestraße 63-73, 14195 Berlin, Germany. Jun Yang and M-Hossein Moeinzadeh contributed equally to this work. *e-mail: zhangpeng@sibs.ac.cn; vingron@molgen.mpg.de

that comes from the presence of three or more copies of the monoploid genome will always hinder the genome assembly process, even with state-of-the-art assembly tools such as Platanus[14], MaSuRCA[15] and SOAPdenovo2[16]. The reason being that genome assembly focuses on the vast majority of bases that are invariant across homologous chromosomes; since these invariant regions are intermittent along the whole chromosome, the result is fragmentation in polyploid assembly. To address this problem, one can separately sequence each chromosome using isolated chromosome arms as reported in the wheat genome project[12,13]. The chromosome isolation technique, however, is tedious and has a long way to go before it can become routine. Nevertheless, this strategy has inspired bioinformaticians to develop new tools for assembling each pair of homologous chromosomes separately. In contrast, with genome assembly, the haplotyping process pays more attention to DNA sequence differences among homologous chromosomes. The integration of genome assembly methods and haplotyping offers us a panoramic view of all the homologous chromosomes. Haplotyping itself, however, poses its own challenges. For example, the haplotyping of the human individual genome relies mainly on fosmid-based sequencing which is costly and time consuming[17,18]. Since the distance of the adjacent variant positions between paternal and maternal chromosomes in humans is normally in the kilobase range[19], it is beyond the capacity of current cost-effective sequencing platforms to cover at least two variant positions in most cases. Nevertheless, human haplotyping studies have already indicated that genome assembly and accurate haplotyping are tightly linked[18]. Unfortunately, the computational phasing problem in polyploidy is considerably harder than for a diploid organism because in the polyploid case one cannot make inferences about the 'other' haplotype once one has seen the first.

Our initial sequencing of the *I. batatas* genome revealed that the distance between adjacent polymorphic sites is roughly one-tenth of the distance in the human genome. Based on previous statistics, there are approximately 14 million polymorphic sites in the estimated 700~800 Mb monoploid genome of *I. batatas*. This means that, on average, one read (100~150 bp length) from Illumina sequencing will cover two or three polymorphic sites. This density of such sites should permit phasing the *I. batatas* genome, employing cost-effective Illumina sequencing with paired-end libraries. The high heterozygosity of *I. batatas* makes genome assembly more challenging, but it simultaneously makes haplotyping easier.

Here we report the development of computational tools to derive a genomic sequence from a polyploid species and its genome that is phased to a large degree. The haplotype-resolved de novo assembly of the *I. batatas* genome was generated entirely based on Illumina sequencing data. The final ~836 Mb assembly has a scaffold N50 of ~201 kb (Table 1). The entire number of scaffolds was 35,919 and their lengths varied from 392 to 1,335,955 bp. Using gene and sequence synteny between *I. batatas* and *Ipomoea nil*, 75.7% of the current assembly has been anchored on 15 pseudochromosomes (Fig. 1). In total, 49,063 genes with 78,781 gene models were extracted from the genome. Furthermore, this compressed monoploid genome was phased into six haplotypes in 644,810 regions. Using phylogenetic analysis of these haplotypes, a hypothesis on the origin of modern cultivated *I. batatas* could be proposed and examined. In addition, the availability of the haplotype-resolved genome allowed to trace back its hexaploidization process and permitted the estimation of the times of two recent whole genome duplication events; these were placed at approximately 0.8 and 0.5 million years ago.

Results

Sequencing data generation. A newly bred carotenoid-rich cultivar of *I. batatas*, Taizhong6 (China national accession number 2013003), was used for genome sequencing (Supplementary Fig. 1). During the genome survey stage, three sequencing libraries were constructed and sequenced on Hiseq2500 and GS FLX + platforms (Supplementary Table 1, A500, A1kb and A454). After a preliminary genome assembly and read mapping for variant calling, the requirement for haplotyping of hexaploid *I. batatas* was estimated to be at least 40-fold monoploid genome coverage (Supplementary Fig. 2). A new library was sequenced on the Nextseq500 platform to meet the estimated data requirement (Supplementary Table 1, L500). Additional gel-free mate pair and 20k mate pair libraries were also sequenced to improve scaffolding (Supplementary Table 1, MP and AMP). The insert size distributions of these paired-end libraries are shown in Supplementary Fig. 3.

Initial assembly of consensus genome. A heterozygosity-tolerant assembly pipeline, combining de Bruijn[20] and OLC[21] (overlap layout consensus) graph strategies, was employed to carry out the hexaploid genome assembly of *I. batatas*, using error corrected Illumina reads (see Methods). A total length of ~870 Mb, mainly representing the monoploid genome, was assembled using this pipeline, with the largest scaffold being 3.7 Mb (corresponding to a completely assembled endophyte *Bacillus pumilus* genome). The second largest scaffold, which harbours 54 genes, was 581 kb (and contained 133 contigs). The N50 of all scaffolds was ~60 kb with a 5,649 bp contig N50. The total number of scaffolds was 79,089 and their length varied between 312 and 3,723,026 bp. There were 3,796 scaffolds longer than 60 kb. Besides scaffolds, there were 991,314 contigs with a total length of ~436 Mb. Among these, 92,790 contigs were longer than 1 kb harbouring a total of 175,679,534 bp. These contigs mainly reflected heterozygosity of the hexaploid genome since 97.35% contigs have been mapped back to scaffolds and one third of these contigs were found to match at full length, despite many single nucleotide mismatches or small indels, as shown in Supplementary Fig. 4.

We identified redundancy in the assembly based on exhaustive comparisons among scaffolds. When one scaffold was already covered by another longer scaffold with more than 85% sequence identity and more than 85% sequence length, the shorter one was removed. The largest scaffold in the preliminary assembly, the endophyte *Bacillus pumilus* genome, was also excluded for further analysis. After these procedures, there were 61,118 remaining scaffolds, of total length 822,598,598 bp, with 64,561 bp N50. Then these scaffolds were connected using 20 kb mate-pair library by Platanus. The N50 of all scaffolds after Platanus scaffolding was ~142 kb. The total number of scaffolds was 57,051 and their length varied between 392 and 1,152,062 bp. We call this version the 'preliminary assembly'.

Variant calling. All the Illumina raw reads were mapped back to all scaffolds of the preliminary assembly (Table 1). After removing the PCR duplicates, there were 1,725,677,696 mapped reads in the final bam file for variant calling using freebayes[22]. In total, there

Table 1 | Summary of assemblies

Assemblies	Total number	Total length (bp)	Gap (%)	N50 (bp)
Preliminary assembly scaffolds	57,051	831,919,670	11.5	142,474
Preliminary assembly unplaced contigs	991,314	435,867,107	0	708
Haplotype-improved assembly scaffolds	35,919	836,316,092	12.0	200,728
Haplotype-improved assembly unplaced contigs	41,487	13,861,310	0	480
Anchored scaffolds	7,470	633,423,954	13.1	41,911,220*

*N50 of 15 pseudochromosomes.

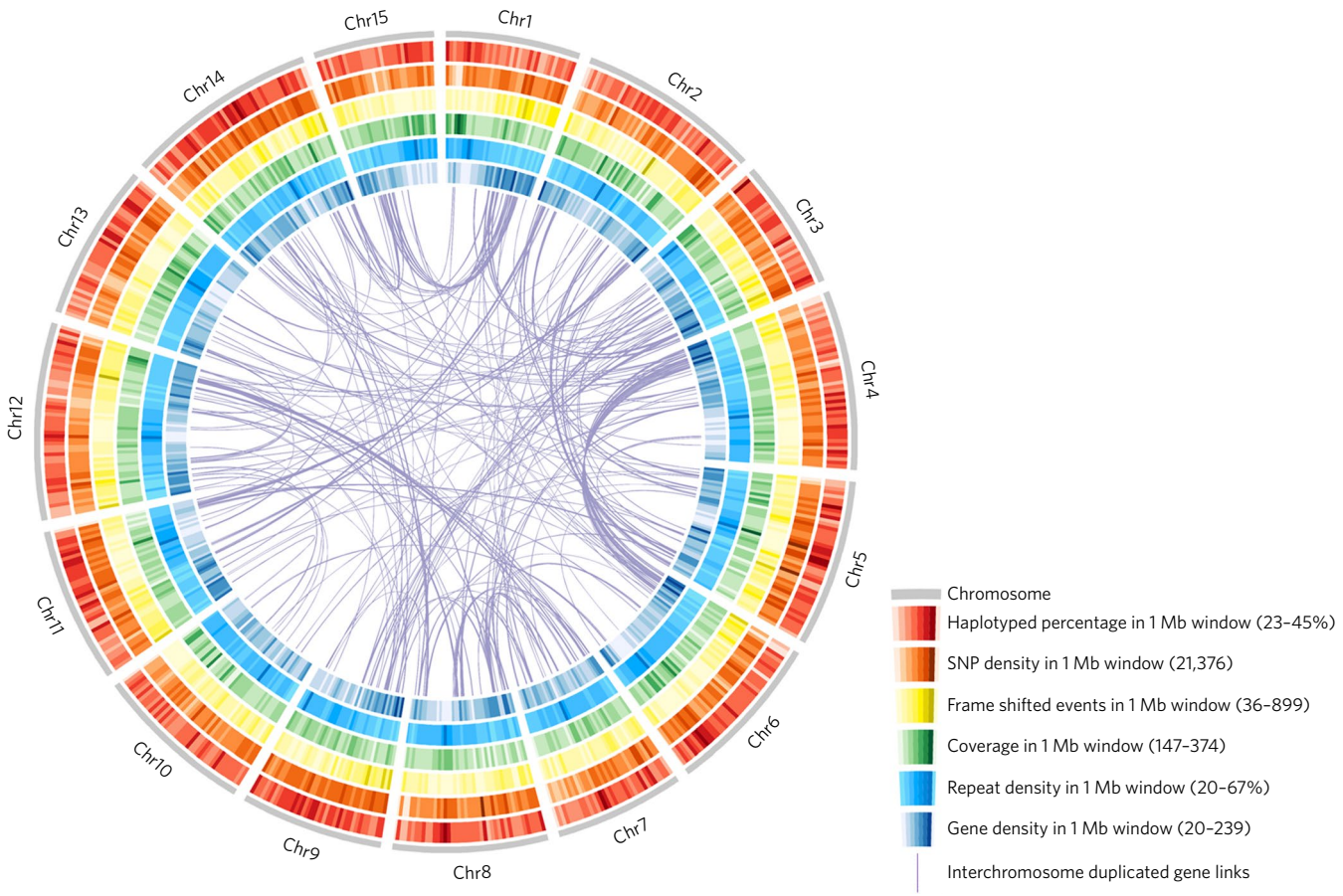

Fig. 1 | Outline of current sweet potato genome assembly. Summary of haplotyped percentage (red), single-nucleotide polymorphism (SNP) density (orange), frame-shifted events because of variants in haplotypes (yellow), sequencing coverage (green), repeat density with putative centromere positions (light blue), gene density (dark blue). Interchromosome duplicated genes are shown as links.

were 14,342,083 variations, consisting mainly of single nucleotide polymorphisms but also indels, across the 831,919,670 bp assembly. Most of the variant positions harboured two possible alleles (Fig. 2a). We observed, on average, one polymorphic site every ~58 bp and a median distance of 20 bp between polymorphic sites. The distance distribution peaked at 6 bp and only 7% (1,004,917/14,285,053) of observed distances are longer than 150 bp (Fig. 2b). These findings confirmed our earlier conclusion that the *I. batatas* genome is very heterozygous and formed the basis for phasing haplotypes using 100–150 bp Illumina reads. A high correlation ($r = 0.975$) was found between number of variations and scaffold length (Supplementary Fig. 5), which increases our chances for phasing the scaffolds.

Phasing of haplotypes. For phasing, we developed an algorithm that assigns reads in a seed region of polymorphic sites to six presumptive haplotypes. For example, three polymorphic sites with two alleles each would give rise to eight combinations. Sequencing errors, however, could artificially inflate the number of hypothesized haplotypes. Thus, we looked for six haplotypes that have the most support in terms of sequencing reads (see Methods and Fig. 3). Haplotypes can be identified from combinations of alleles over many polymorphic sites that are connected in a read. Our algorithm looks for combinations of two, three and four polymorphic sites. These phased regions were then extended continuously by searching for elongation of individual haplotypes. Finally, 542,361 regions were found and ~30% of the genome was phased into six haplotypes. These haplotypes could be further extended by connecting paired-end reads. Some of the paired-end reads map to haplotypes within one scaffold, and other paired-end reads connect haplotypes from different scaffolds because of better matching to the phased sequence. These connections are used for the haplotype-improved assembly.

Haplotype-improved assembly. All the Illumina raw reads were subsequently mapped against haplotype sequences generated by the phasing step. Only perfectly matched paired-end reads were considered as haplotype connections. The interscaffold and intrascaffold connections were separated for haplotype-based scaffolding and haplotype elongation, respectively. Gap sizes between connected scaffolds were estimated and super scaffolds were generated by SSPACE[23] resulting in haplotype-improved assembly. The final assembly N50 of this consensus genome was ~201 kb (Table 1). The total number of scaffolds was 35,919 and lengths varied between 392 and 1,335,955 bp. When the contigs in the preliminary assembly were mapped against the haplotype-improved assembly, one-third of these contigs could be mapped back at full length though allowing for a fairly large number of mismatches. These contigs probably comprise several particular haplotypes (Supplementary Fig. 6). Using a similar scenario for removing redundancy in scaffolds, there were 41,487 remaining contigs, with a total length of ~14 Mb (Table 1). There were 7,470 scaffolds in the haplotype-improved assembly with five or more genes, accounting for 75.7% of the total length of all scaffolds, which could be anchored to 15 pseudochromosomes according to the gene and sequence synteny between our assembly and the newly released genome of *I. nil*[24]. The 15 pseudochromosomes with the remaining scaffolds served as our final

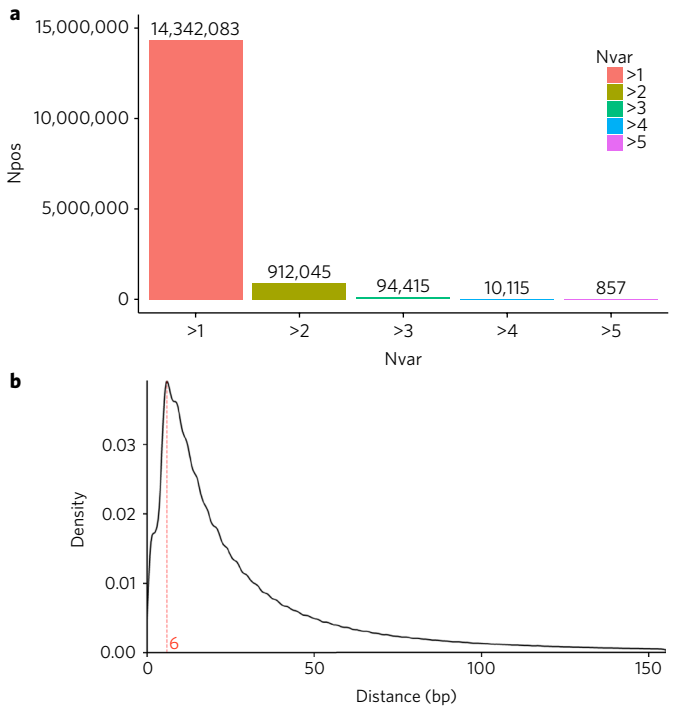

Fig. 2 | Summary of variations. a, The total numbers of variations with different number of variant alleles. On average 831,919,670/14,342,083 = 58 bp with one variant, which means short reads (100 bp and 150 bp) are informative for haplotyping. Npos, number of positions; Nvar, number of variants. **b**, Adjacent variation distance distribution peaked at 6 bp (red dashed line). Only 1,004,917/14,285,053 = 7% observed distances are longer than 150 bp.

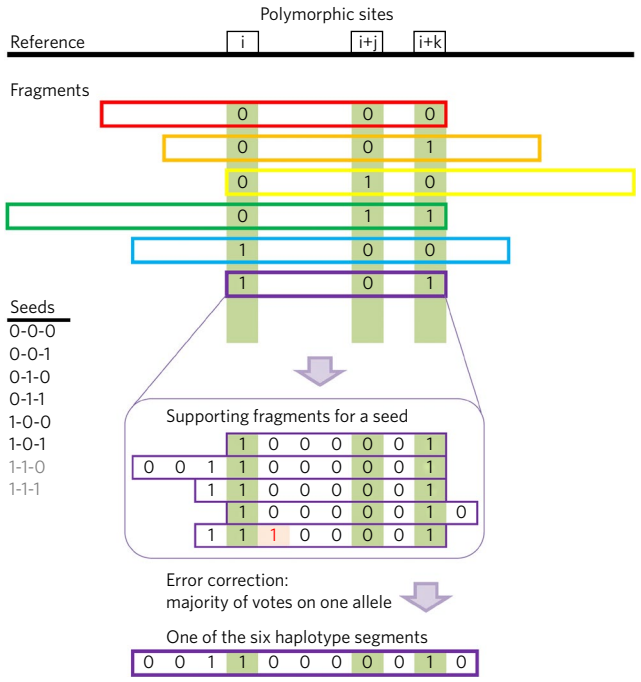

Fig. 3 | Illustration of seed-finding algorithm. Step 1, searching all possible seeds using different seed lengths. As an example, seed length 3 is shown here. Step 2, selecting the six highest supported sequence patterns in the seed; grey indicates less supported sequence patterns. Step 3, extension of particular seed based on supporting fragments. Reference, a sequence of variants in the scaffolds with each variant is coded as 0. Fragment, the sequence of variants from one chromosome which is covered by one read. Variants are coded 0–5. Paired-end information is not used here. Seed, a number of polymorphic sites together representing six or more unique sequence patterns. Haplotype, a certain region in one chromosome.

assembly, which we annotated, used for final phasing of haplotypes, and from which we traced the hexaploidization of the *I. batatas* genome (Fig. 1).

Genome annotation and cDNA validation. After six transcriptome data sets were mapped to the final assembly, 78,781 gene models from 49,063 gene loci were extracted and revised by StringTie[25] and TACO[26], respectively. To validate the predicted gene models from transcriptome data, 10,063 expressed sequence tags (ESTs) were generated by Sanger sequencing. These sequences were then mapped to the genome. Of these, 9,425 were located correctly and found to have corresponding predicted gene models. The functional annotations of predicted genes were obtained by homologous protein sequence searching in the public databases Uniprot (http://www.uniprot.org/) and NCBI protein database (https://www.ncbi.nlm.nih.gov/). RepeatModeler[27] was employed to identify repeat sequences in the present genome. The repeat sequence classifications are summarized in Table 2. The horizontally transferred T-DNAs reported by Kyndt et al.[28] were also investigated in the current genome. We found multiple copies of these T-DNAs suggesting that the horizontal gene transfer event has happened before hexaploidization of *I. batatas* (T-DNA1, NCBI accession KM052616, hit scaffold14997; and T-DNA2, NCBI accession KM052617 have been located in pusedochromosome 7).

Gene cluster identification. The discovery of operon-like gene clusters in plant genomes has prompted attempts to decode the regulatory mechanism in plant specialized compound biosynthesis[29]. Gene clusters of paralogous genes have been identified in a wide range of species including maize, lotus, cassava, sorghum, poppy, tomato, potato, rice, oat and *Arabidopsis*[30,31]. In the *I. batatas* genome, we identify four clusters of six or seven genes for alkaloids (Fig. 4, GC002, GC006, GC014 and GC024), three six-to-eight gene clusters for terpenes (Fig. 4, GC001, GC005 and GC031), and a six-gene cluster for cellulose (Fig. 4, GC028) by searching orthologous genes using Exonerate[32]. In addition to the gene clusters shown in Fig. 4, there

Table 2 | Summary of repeat sequence identification

Type of elements	Number of elements	Length occupied (bp)	Percentage of genome*
LTR	213,439	92,066,503	10.987
DNA elements	256,260	50,863,135	6.070
LINE	40,146	20,814,265	2.484
Simple repeat	324,390	14,578,880	1.740
RC/Helitron	12,067	5,610,834	0.670
Low complexity	45,912	2,324,352	0.277
SINE	4,168	618,723	0.074
rRNA	322	97,450	0.012
Satellite	456	33,443	0.004
snRNA	219	29,584	0.004
Unknown	985,978	195,232,724	23.299
Total	1,883,357	382,269,893	45.619

*Scaffolds and unplaced contigs were taken as input sequences of RepeatModeler. LTR, long terminal repeat; LINE, long interspersed nuclear element; RC, rolling-circle transposable element; SINE, short interspersed element; snRNA, small nuclear RNA.

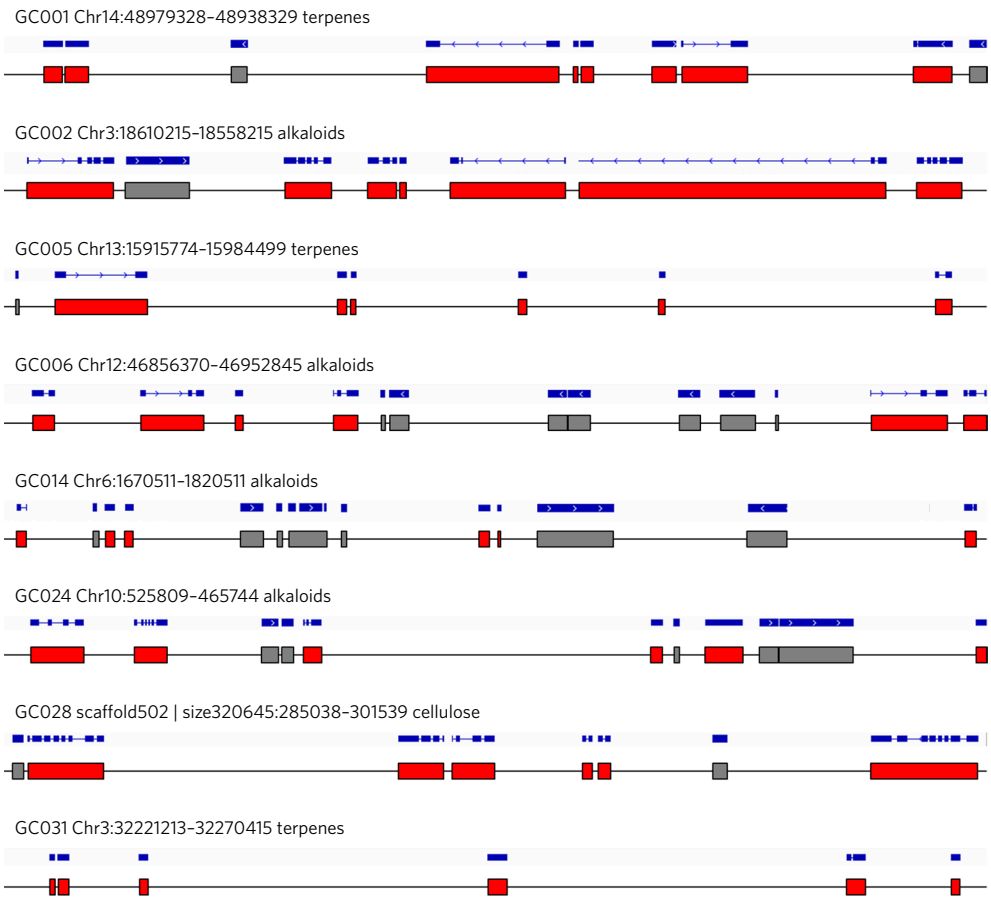

Fig. 4 | Identified gene clusters in present *I. batatas* genome. Blue bars indicate the gene structure, the exon as box and the intron as line. The red boxes indicate genes from a cluster and the grey boxes indicate unrelated genes. The best hit genes of GC001 are *ZmBx4*, *AtMRO*, *AtMRO*, *PsCYP82×2*, *SbCYP71E*, *AtMRO*, *SlP450-1* and *SbCYP71E*; the best hit gene of GC002 is *SlGAME7*; the best hit genes for GC005 are *ZmBx5*, *PsCYP82Y1*, *OsCYP76M5*, *LjCYP71D11*, *PsCYP82×2* and *LjCYP736A2*; the best hit genes for GC006 are *PsCYP82×1*, *PsCYP82×2*, *OsCYP71Z6*, *ZmBx4*, *LjCYP79D4* and *PsCYP82×2*; the best hit genes for GC014 are *PsCXE1*, *PsCXE1*, *PsCXE1*, *SlGAME3*, *StSGT1* and *MeUGT85K4*; the best hit genes for GC024 are *SlGAME11*, *SlGAME7*, *SlGAME7*, *SlGAME3*, *SlGAME3* and *SlGAME3*; the best hit gene of GC028 is *StCeSy*; the best hit genes for GC031 are *MeCYP71E*, *OsCYP71Z7*, *OsCYP99A2*, *SlP450-1*, *OsCYP71Z6* and *SlP450-1*. All the detailed gene cluster information can be found in Supplementary Table 2.

were 68 more gene clusters found in the current genome assembly (Supplementary Table 2). The results indicate that pathway regulation via clustered genes is commonly used in the genus *Ipomoea*. Although all the orthologous genes found here are based on protein sequence similarity, their biological functions are not necessarily the same as those reported for similar gene clusters in other species. Nevertheless, the identified gene clusters in *I. batatas* open up possibilities for investigating metabolic regulatory mechanisms in this plant.

Updating and validation of haplotypes. Based on the final assembly, we applied our phasing procedure a second time and phased 644,810 regions into 1,928,359,400 bp haplotype sequences. Thus, the new assembly leads to extended haplotypes in 18.9% more regions and a 29.5% improvement of total length. We validated these final haplotypes using a set of Roche 454 reads (Supplementary Table 1, A454) that had been produced earlier but which had not been utilized for assembly or phasing. The reads are on the order of 1000 bp long and can thus serve to identify errors in the haplotype reconstruction. A large fraction of these 454 reads display haplotypes that have been correctly reconstructed by our short-read based methodology. More than 60% of overlaps between haplotypes and 454 reads are identical at variant loci. The longest reconstructed haplotype contained 92 polymorphic sites without any mismatch indicated by 454 reads. There may be many longer perfectly reconstructed haplotypes that remain undetected because of the limited 454 read length (Supplementary Fig. 7).

Variant updating in phased regions. Variants from freebayes have been updated according to phased haplotypes. When all six haplotypes covered the specific variant position, our method distinguishes two possible situations: (1) the six haplotypes indicate the same variant, and (2) some haplotypes indicate different variants. Situation (1) indicates an error in variant calling and we remove the variant position from the original VCF file. For situation (2), our method ranks alleles based on the number of supporting haplotypes. Then the first rank is picked as the reference allele and the rests are considered as the alternative ones. Using this strategy, we removed 115,228 variant positions and updated 1,514,176 variant positions to generate the final VCF file.

The phased haplotypes in protein coding region also allowed us to evaluate the impact of variants. We focused on the frame shift events, identifying 172,830 events across 34,691 putative transcripts (Fig. 1, yellow circle). Furthermore, when all haplotype sequences were mapped back to the diploid *I. trifida* consensus genome, we could identify the conserved ancient variants and haplotypes that are shared between this hexaploid organism and its diploid progenitors. This also led us to check the phylogenetic relationship of phased haplotypes.

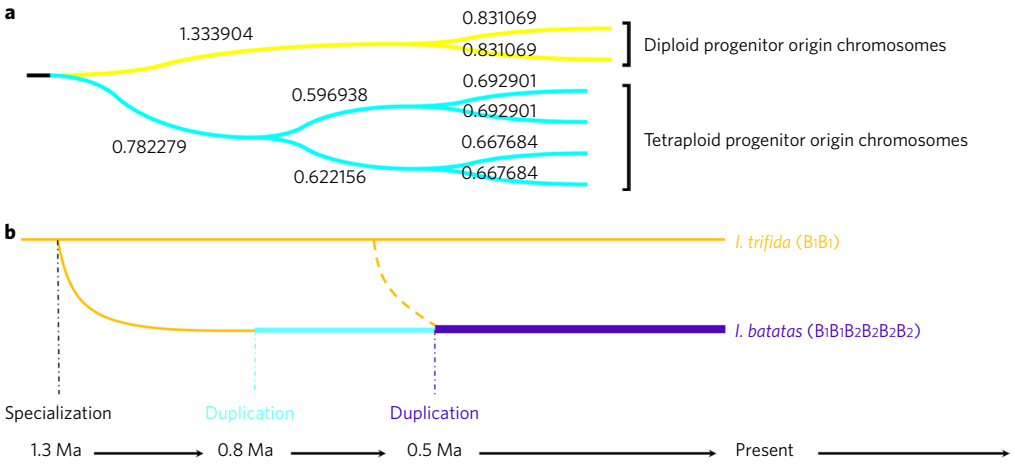

Fig. 5 | Evolutionary history of cultivated *I. batatas* revealed by phylogenetic analysis of homologous chromosome regions. a, The dominant topology structure of all phylogenetic trees. Numbers indicate the average branch length of trees in this structure. **b**, The time points of B_2 subgenome specialization and two whole-genome duplication events were estimated as 1.3, 0.8 and 0.5 million years ago (Ma). The estimation is based on 0.7% mutation rate per million years. The dashed curve indicates the crossing between diploid and tetraploid progenitors.

Phylogenetic analysis of homologous chromosome regions. Phylogenetic analysis was applied to the final haplotype-resolved genome. All the regions phased into six haplotypes were used for constructing UPGMA (unweighted pair group method with arithmetic mean) trees with the MEGA-Computing Core[33]. We generated 644,360 phylogenetic trees and determined the distribution of their topologies over the six possibilities, explaining the evolution of six haplotypes. Note that between any two such trees we do not identify the sequences, but rather focus on the branching pattern alone. We denote the topologies as annotated in Supplementary Fig. 8. The branching pattern of two haplotypes versus four haplotypes is the stable majority class with or without consideration of repetitive sequences. Furthermore, among the four-haplotype subgroup, the division two versus two dominates (Supplementary Fig. 8). The average branch length of these 2:4–2:2 trees were obtained and a consensus tree was constructed (Fig. 5a). These results suggest two whole-genome duplication (WGD) events in *I. batatas* hexaploidization history together with a further diversification of the B_2 subgenome (Fig. 5b). Assuming a 0.7% mutation rate (base pairs per million years[34]) the tetraploid progenitor of *I. batatas* was produced by the first WGD event estimated at 0.8 million years ago. The origin of modern cultivated *I. batatas* would then have been the result of an initial crossing between this tetraploid progenitor and a diploid progenitor, followed by the second WGD event occurring about 0.5 million years ago. The most probable diploid progenitor of *I. batatas* is also the most likely ancestor of modern *Ipomoea trifida*, although the tetraploid progenitor is still unknown. It might be identified, however, by a genome survey of the genomes of modern tetraploid species in *Ipomoea* genus.

Discussion

Here, we have reported the generation of sequencing data and the development of the necessary algorithms to compute long haplotype segments for the *I. batatas* genome. Based on an initial assembly and phasing, the computational process was repeated, yielding an improved assembly. In principle, haplotype phasing and haplotype-aided assembly can be iterated further to improve both results. In the present work, we only updated haplotypes based on the final assembly and observed remarkable improvements in terms of more phased regions and longer haplotype length. Paired-end reads have been employed to extend haplotypes both within and across assembly scaffolds. Overall, high-quality assembly and phasing can now be achieved with a comparatively low investment in sequencing (two runs on a HiSeq2500, two runs on a NextSeq500 and one run on a Hiseq4000; see Supplementary Table 1) but increased computational effort, both in terms of algorithm development and sheer computational power.

Based on the gene synteny between the haplotype-improved assembly and the *I. nil* genome[24], we generated 15 pseudochromosomes of *I. batatas*. Although the precise reconstruction of entire chromosomes is not possible using this methodology and is reserved for alternative approaches including long-read sequencing technologies, many evolutionary questions regarding the ancestry of polyploid organisms can now be tackled using the methodologies we have developed. For the *I. batatas* genome, we have shown, based on the available haplotype alignments, that the majority of computed phylogenetic trees groups two haplotypes versus four, the latter being in turn symmetrically grouped into two and two. The presence of the other, less supported phylogenetic groupings may indicate a high recombination rate between B_1 and B_2 subgenomes. Although the different haplotype alignments obtained by our data and analysis do not allow identification of the full set of chromosomal connections among the segments, this phylogenetic pattern dominates. It is in agreement with a scenario in which modern cultivated *I. batatas* originated from a cross between a diploid progenitor and a tetraploid progenitor, followed by a whole-genome duplication event. The precise timing of these evolutionary events, however, relies on an estimation of the average mutation rate which remains somewhat uncertain. Nevertheless, we have demonstrated the power of phylogenetic analysis of the haplotypes derived by our approach in the reconstruction of the hexaploidization history of *I. batatas*. When applied to the polyploidization of plants in general, phylogenetic analysis based on haplotype reconstruction could prove to be the most reliable way to study the origin of each set of chromosomes in complex polyploid organisms.

Our seed-based computational approach has thus proven successful on this very heterozygous genome, even with only short sequencing reads. This half haplotype-resolved hexaploid genome represents the first successful attempt to investigate the complexity of chromosome sequence composition directly in a polyploid genome using sequencing of the polyploid organism itself rather than any of its simplified proxy relatives. Adaptation and application of our approach should provide higher resolution in future genomic structure investigations, especially for similarly complex genomes. The approach presented here has a high degree of flexibility, which can be employed with many kinds of sequencing technologies and

is ready to use for haplotyping tasks in a wide range of research programmes. With the availability of longer reads in the future, the same computational philosophy should be applicable for phasing of genome segments in other polyploid organisms, even when the density of polymorphic sites is lower. Given the importance of *I. batatas* as a crop species, a resolved genome will be of prime importance for improving yield security and nutritional quality via gene or genome engineering. Such avenues are now open for, on the one hand, rendering the plant more stress tolerant and disease resistant, classical goals of plant breeding, but, on the other hand, for better understanding the metabolic capacities of this species in allowing effective biofortification as has recently been demonstrated in cassava[35]. We believe, therefore, that the methodology presented in this paper additionally presents an important advance in the genomic analysis of polyploid organisms. Meanwhile, the haplotype-resolved genome sequence of sweet potato is the beginning of the precise genome manipulation era for this untapped human food resource.

Methods

Genome sequencing and assembly. A newly bred carotenoid rich cultivar of sweet potato (*I. batatas*), Taizhong6, whose China national accession number is 2013003, was selected as the target cultivar. Total genomic DNA was isolated from in vitro cultured plants following the method described by Kim and Hamada[36]. In total, six sequencing libraries were constructed and sequenced on Hiseq2500, Nextseq500, Hiseq4000 and GS FLX + platforms (Supplementary Table 1, A500, A1kb, L500, MP, AMP and A454) according to the manufacturer's instructions (Illumina, Inc. and Roche Applied Science), respectively.

The main steps of consensus genome assembly were as follows:

a) Read-correction of all Illumina data using the BFC package (https://github.com/lh3/bfc).
b) Assembling all short reads by IDBA-UD with –mink 100 –maxk 151 –step 23.
c) Further assembling the IDBA outputs by a long read assembler (Newbler 3.0).
d) Two scaffolding runs using the Platanus scaffolder on Newbler 3.0 output.
e) Gap closing using all corrected Illumina reads (Platanus GapCloser).
f) Remove redundancy in preliminary assembly scaffolds.
g) Platanus scaffolding using 20 kb insertion library.

Detailed information on the pipeline presented in this paper can be found in the Supplementary Information.

Haplotyping algorithm and pipeline. All the Illumina raw reads were mapped back to all scaffolds using BWA (Version 0.7.12-r1039). PCR amplification duplicates were removed via Samtools (version 0.1.19-44428 cd). Freebayes (version 0.9.14-19-g8a407cf) was employed for variant calling in the hexaploid genome.

For haplotyping, our method takes FASTA, BAM and VCF formatted files as inputs. We based the reconstruction on seed regions, which are small sets of polymorphic sites. For example, three polymorphic sites with two alleles each would allow up to $2\times2\times2=8$ haplotypes. Only a subset of those will be supported by reads. The algorithm searches for all possible seed regions containing six or more sequence patterns. Seeds, however, can be interleaved. Haplotypes in a seed region are sorted by the number of their supporting reads. Different seed regions are sorted according to the number of supporting reads for the sixth-most strongly supported haplotype, because we expect six haplotypes.

The following steps were done iteratively on each seed region from the sorted seed list. The six most supported sequence patterns in each seed region were considered to be the six haplotype cores. Then, each haplotype was built up on each core according to the supporting reads for that sequence pattern (Fig. 3). The haplotypes were extended via uniquely matched reads. A uniquely matched read is a read that matches exactly one of the six haplotypes while being distinct from the other haplotypes in the seed region. Chaining haplotypes through uniquely matched reads produced six haplotypes in an extended seed region (Supplementary Fig. 9a,b).

Then the adjacent overlapping seed regions were merged (Supplementary Fig. 9c). One haplotype in a seed region was merged to a haplotype in the adjacent seed region if they shared a uniquely matching read with each other. Thereafter, we utilized paired-end reads to connect the haplotypes obtained so far. All the Illumina raw reads were mapped back to the phased haplotypes. Only perfectly matched paired-end reads were considered as haplotype connections. The innerscaffold haplotyping exploited the paired-end reads within one scaffold. The algorithm started with the highest supported connection and merged it as a new haplotype. Possible conflicts were checked and avoided in this step. A conflict occurred when two haplotypes in one seed region were connected via a path through the haplotypes in other seed regions. Paired-end reads were further utilized for interscaffold connection to elongate and merge the haplotype from different scaffolds. The haplotyping method utilizes parallel computation in order to speed up the analysis.

Haplotype evaluation using 454 reads. To evaluate the haplotyping accuracy, we used a set of 454 (Supplementary Table 1, A454) reads that had been produced earlier but were not utilized for assembly or phasing. Each 454 read can be considered as a short DNA fragment from one chromosome, except for some chimeric reads in rare cases. Roche 454-trimmed reads were mapped against genome assembly. Only the polymorphic sites indicated by variant calling were extracted and their overlaps with haplotypes were evaluated. The 'match' and 'mismatch' sites of each overlap were recorded for further analysis and visualization.

Haplotype-improved assembly and variant correction. A python script was employed to convert all interscaffold connection information into SSPACE tab format. Exhaustive comparisons among scaffolds using blast were employed to identify and remove redundancy in scaffolds. If one scaffold was already covered by another longer scaffold with more than 85% sequence identity and more than 85% sequence overlap, the shorter one was removed. Additional manual checking of long candidates (>10 kb) via Circos visualization was also included. Finally, the non-redundant scaffold sequences without the endophyte *Bacillus pumilus* genome were employed as input for scaffolding.

We refined variants according to phased haplotypes. There are three possibilities for one phased variant: (1) all six haplotypes covered the variant, (2) some of the haplotypes covered the variant and (3) the variant was located outside of haplotype blocks. There was insufficient information to update alleles in the second and third categories. Our method distinguished the first group from the other two and categorizes its members in the following subgroups: (1) all alleles are the same and (2) the alleles are different. The first subgroup was put aside as an error in variant calling, which meant that there was insufficient support to consider this position as a variant position. We therefore removed the variant calling result of this position from the original VCF file. For the second group, our method ranked alleles firstly based on the number of supporting haplotypes, secondly according to the number of supporting reads. Then the first allele was picked as a reference allele and the rest were considered as alternative alleles.

Genome annotation. Six transcriptome data sets from four previous studies[37–40] and two additional RNA-seq experiments were mapped on the current genome using HISAT2 and HISAT[41]. These transcriptome data represented different plant tissues such as leaves, petioles, stems and roots from different developmental stages. All gene models were extracted by StringTie[25]. A consensus transcriptome was generated by TACO[26]. The obtained transcripts were further annotated using homologous protein searching in the public databases NCBI and Uniprot.

RepeatMasker (Version open-4.0.5) was used to mask and classify repeat sequences in the genome. RepeatModeler (Version 1.0.8)[27] was employed to identify novel repeat sequences.

Phylogenetic analysis. A python script was developed to extract alignments from all phased regions. MEGA-Computing Core[33] was utilized to compute all the phylogenetic trees of extracted alignments on a computer farm. Trees were classified into groups based on their topology structures and average haplotype lengths in these groups were compared. The dominant tree structure was selected and a consensus tree with average branch lengths was obtained.

Data and code accessibility. The *I. batatas* genome sequence, including pseudochromosomes and unanchored scaffolds, haplotype-resolved regions, and unplaced contigs, are publicly available at the sweet potato genome browser http://public-genomes-ngs.molgen.mpg.de/SweetPotato/ and Ipomoea Genome Hub http://www.ipomoea-genome.org/. Two transcriptome data sets and one sequenced cDNA library are also available for download. cDNA clone delivery is also possible upon request. The assemblies and the WGS raw data have been deposited with European Nucleotide Archive (ENA) under project number PRJEB14638 and National Center for Biotechnology Information (NCBI) under project number PRJNA301667. The Ranbow program, the implementation of the described algorithm, is available from https://www.molgen.mpg.de/ranbow.

Received: 25 August 2016; Accepted: 11 July 2017;
Published online: 21 August 2017

References

1. *Crops* (FAO, accessed 1 August 2017); http://www.fao.org/faostat/en/#data/QC
2. Ozias-Akins, P. & Jarret, R. L. Nuclear DNA content and ploidy levels in the genus *Ipomoea*. *J. Am. Soc. Hortic. Sci.* **119**, 110–115 (1994).
3. Ukoskit, K. & Thompson, P. G. Autopolyploidy versus allopolyoloidy and low-density randomly amplified polymorphic DNA linkage maps of sweetpotato. *J. Am. Soc. Hortic. Sci.* **122**, 822–828 (1997).
4. Kriegner, A., Cervantes, J. C., Burg, K., Mwanga, R. O. M. & Zhang, D. A genetic linkage map of sweetpotato (*Ipomoea batatas* (L.) Lam.) based on AFLP markers. *Mol. Breeding* **11**, 169–185 (2003).
5. Hirakawa, H. et al. Survey of genome sequences in a wild sweet potato, *Ipomoea trifida* (H. B. K.) G. Don. *DNA Res.* **22**, 171–179 (2015).

6. The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189–195 (2011).
7. Li, F. et al. Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* **33**, 524–530 (2015).
8. Wang, K. et al. The draft genome of a diploid cotton *Gossypium raimondii*. *Nat. Genet.* **44**, 1098–1103 (2012).
9. Li, F. et al. Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat. Genet.* **46**, 567–572 (2014).
10. Ling, H. et al. Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* **496**, 87–90 (2013).
11. Jia, J. et al. *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**, 91–95 (2013).
12. The International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**, 1251788 (2014).
13. Choulet, F. et al. Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345**, 1249721 (2014).
14. Kajitani, R. et al. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).
15. Zimin, A. V. et al. The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
16. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**, 18 (2012).
17. Duitama, J. et al. Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of single individual haplotyping techniques. *Nucleic Acids Res.* **40**, 2041–2053 (2012).
18. Cao, H. et al. *De novo* assembly of a haplotype-resolved human genome. *Nat. Biotechnol.* **33**, 617–622 (2015).
19. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
20. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2002).
21. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
22. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at http://arxiv.org/abs/1207.3907v2 (2012).
23. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
24. Hoshino, A., Jayakumar, V., Nitasaka, E., Toyoda, A., Noguchi, H., Itoh, T., Shin-I, T., Minakuchi, Y., Koda, Y. & Nagano, A. J. et al. Genome sequence and analysis of the Japanese morning glory *Ipomoea nil*. *Nat. Commun.* **7**, 13295 (2016).
25. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
26. Niknafs, Y. S., Pandian, B., Iyer, H. K., Chinnaiyan, A. M. & Iyer, M. K. TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat. Methods* **14**, 68–70 (2017).
27. Smit, A. & Hubley, R. RepeatModeler - 1.0.8 (Institute for Systems Biology, 2015); https://sourceforge.net/u/djinnome/jamg/ci/47152a01077445af52726d76270e60bb360bb2f2/tree/3rd_party/RepeatModeler-open-1.0.8/
28. Kyndt, T. et al. The genome of cultivated sweet potato contains *Agrobacterium* T-DNAs with expressed genes: an example of a naturally transgenic food crop. *Proc. Natl Acad. Sci. USA* **112**(18), 5844–5849 (2015).
29. Nützmann, H. W. & Osbourn, A. Regulation of metabolic gene clusters in *Arabidopsis thaliana*. *New Phytol.* **205**, 503–510 (2015).
30. Fernie, A. R. & Tohge, T. Location, location, location – no more! The unravelling of chromatin remodeling regulatory aspects of plant metabolic gene clusters. *New Phytol.* **205**, 458–460 (2015).
31. Boycheva, S., Daviet, L., Wolfender, J. L. & Fitzpatrick, T. B. The rise of operon-like gene clusters in plants. *Trends Plant Sci.* **19**, 447–459 (2014).
32. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
33. Kumar, S., Stecher, G., Peterson, D. & Tamura, K. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* **28**, 2685–2686 (2012).
34. Ossowski, S. et al. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
35. Li, K. T. et al. Increased bioavailable vitamin B_6 in field-grown transgenic cassava for dietary sufficiency. *Nat. Biotechnol.* **33**, 1029–1032 (2015).
36. Kim, S. H. & Hamada, T. Rapid and reliable method of extracting DNA and RNA from sweetpotato, *Ipomoea batatas* (L). Lam. *Biotechnol. Lett.* **27**, 1841–1845 (2005).
37. Firon, N. et al. Transcriptional profiling of sweetpotato (*Ipomoea batatas*) roots indicates down-regulation of lignin biosynthesis and up-regulation of starch biosynthesis at an early stage of storage root formation. *BMC Genomics* **14**, 460 (2013).
38. Wang, Z. et al. *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). *BMC Genomics* **11**, 726 (2010).
39. Xie, F. et al. De novo sequencing and a comprehensive analysis of purple sweet potato (*Ipomoea batatas* L.) transcriptome. *Planta* **236**, 101–113 (2012).
40. Tao, X. et al. Digital gene expression analysis based on integrated *de novo* transcriptome assembly of sweet potato [*Ipomoea batatas* (L.) Lam.]. *PLoS ONE* **7**, e36234 (2012).
41. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

Acknowledgements

We thank J. Dai and Z. Nikoloski for helpful discussions during the haplotyping. We thank J. Zhu and Y. Jiang from Purdue University, J. Yu from Iowa State University and G. Gheysen from Ghent University for their invaluable comments during proofreading. J. Yang acknowledges support from the Alexander von Humboldt Foundation (Forschungsstipendium für erfahrene Wissenschaftler). M-Hossein Moeinzadeh acknowledges support from IMPRS-CBSC doctoral programme. This project was funded by the International Science & Technology Cooperation Program of China (2015DFG32370), the National Natural Science Foundation of China (31201254, 31361140366, 31501353), the National High Technology Research and Development Program of China (2011AA100607-4, 2012AA101204-3), the Chinese Academy of Sciences (2012KIP518), the China Postdoctoral Science Foundation (2012M520945), the Shanghai Municipal Afforestation & City Appearance and Environmental Sanitation Administration (G102410, F122422, F132427, G142434, G152429) and the Science and Technology Commission of Shanghai Municipality (14DZ2260400, 14ZR1414100).

Author contributions

J.Y., M-H.M., H.K., A.R.F., B.T., P.Z. and M.V. planned and coordinated the project and wrote the manuscript. G.-L.L., J.-L.Z. and Z.S. supplied the newly bred cultivar, Taizhong6. W.-J. F., G.-F.D. H.-X.W. and S.-S.Z. prepared genomic DNA. H.K. conducted the primary genome assembly and repeat sequence identification. J.Y. and M-H.M. conducted haplotyping and genome evolution analysis. S.B. managed part of sequencing work. J.H., P.X., S.H. and F.-H.H. supported and inspired a part of the analysis.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at doi:10.1038/s41477-017-0002-z.

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

Correspondence and requests for materials should be addressed to P.Z. or M.V.