A high-quality carrot genome assembly provides new insights into carotenoid accumulation and Asterid genome evolution

Iorizzo et al.

SUPPLEMENTARY NOTE

Table of Contents

Nature Genetics: doi:10.1038/ng.3565

Nature Genetics: doi:10.1038/ng.3565

Nature Genetics: doi:10.1038/ng.3565

1. Genome assembly

1.1 Plant materials and sequencing

1.1.1 Plant materials and DNA preparation

A doubled haploid orange Nantes type carrot (DH1), NCBI biosample SAMN03216637, was used for genome sequencing. Seeds from a self-pollinated DH plant were kindly provided by Rijk Zwaan seeds, Inc.. Dihaploid plants with a completely homozygous genome are known to facilitate the process of *de-novo* genome assembly. Plants were grown in a greenhouse and fresh unexpanded leaves were harvested and frozen immediately in liquid nitrogen for isolation of genomic DNA. DNA for genotyping and short paired-end (PE) libraries (insert size 170-800 nt) was extracted using DNeasy Plant Maxi Kit (QIAGEN GmbH). High molecular weight genomic DNA for mate-pair (MP) libraries was extracted using the Cetyl Trimethyl Ammonium Bromide (CTAB) method described by Murray and Thompson[1] with some modifications to limit DNA degradation. The quality and size range of the isolated DNA was confirmed following separation by gel electrophoresis through agarose gels (0.3% agarose) for 24h at 30V. DNA was quantified by PicoGreen (Life Technologies) and the DNA purity was verified by NanoDrop™ Spectrophotometers (Thermo Scientific).

1.1.2 DH1 homozygosity evaluation

To ensure the homozygous nature of the DH1 plant used in this project, its DNA was genotyped with 3,636 validated SNPs randomly distributed along the carrot genome[2]. All genotyping was performed by LGCgenomics (http://www.lgcgenomics.com/). SNPs were genotyped using the KASPar chemistry, a competitive allele-specific PCR SNP genotyping system using FRET quencher cassette oligos (http://www.lgcgenomics.com/genotyping/kasp-genotyping-chemistry/). In total, only seven (<0.01%) primer sets were polymorphic. Post-assembly examination of those seven primer sites revealed that they were all located in duplicated regions of the genome, likely producing false heterozygous signals in the KASPar

Nature Genetics: doi:10.1038/ng.3565

assay. This result provided evidence that the DH1 genotype is highly homozygous and was suitable for whole genome sequencing.

1.1.3 Whole genome sequencing

The whole genome sequencing for the DH1 *de novo* genome assembly was generated by Illumina sequencing technology at Beijing Genome Institute, Shenzhen (BGI- Shenzhen, China) (**Supplementary Table 1).** Eight paired-end libraries were prepared to sequence the DH1 genome. These included three paired-end libraries with insert sizes of 170, 280 and 800 nt and five mate-pair libraries with insert sizes of 2, 5, 10, 20 and 40 kb. All the libraries were constructed with genomic DNA >20-40 kb in length according to the manufacturer's instructions (Illumina). The quality of each library was validated using Qubit®, AGE. Whole genomic sequence (147.2 Gb) was generated solely using Illumina platforms (HiSeq 2000). Hereafter we will refer to those sequences as PE and MPE for the paired-end and mate-pair sequences, respectively. Sequences have been deposited in the NCBI Sequence Read Archive under project PRJNA268187.

1.1.4 BAC end sequences (PE BACs)

The BAC end sequences of 40,693 clones from a DH1-BAC library were kindly provided by Rijk Zwaan seeds, Inc.. The insert size estimated by pulsed-field gel electrophoresis (PFGE) was 148±70 kb. Of those, 10,818 clones had sequence information from a single end and 29,875 clones had sequences from both ends, resulting in 0.04 Gb of data with an average sequence length of 566 nt (**Supplementary Table 49**). Only sequences from BAC clones with both ends sequenced were used in the carrot assembly pipeline (**Supplementary Table 50**). Hereafter we will refer to those sequences as PE BACs (paired-end BAC sequences).

1.2 Raw data processing and estimation of genome size

1.2.1 Raw data filtering

To avoid assembly errors, the following five steps were used to filter out low-quality sequences:

1) Filter reads with "N"s that constitute more than 10 percent of nucleotides or included polyA structure reads;
2) Filter reads that have 30 nt with quality scores less than or equal to 7 for the large insert size and 40 nt for the short ones;
3) Filter reads with more than 10 nt aligned to the adapter sequences (allowing less than or equal to 3 nt mismatches);
4) Filter small insert size reads in which read1 and read2 overlapped 10 or more nt allowing 10% mismatch, where read1 and read2 are both ends of one paired end read;
5) Filter reads that are totally identical, since these reads can be considered duplicates.

The filtering was carried out using an in-house program. After filtering, approximately 8.8×10^{10}nt (186× coverage) of high-quality sequence data for *de novo* assembly were generated (**Supplementary Table 50**).

1.2.2 Genome size estimation

K-mer refers to a sequence with k nucleotides. The abundance of k-mers was quantified by dividing short-insert-size reads (170 and 800 nt PE) into sliding sequences of 17 nt, overlapping by 16 nt and by calculating the frequency of each k-mer (**Supplementary Figure 28**). The k-mer frequency follows a Poisson distribution beyond a certain quantity of data, allowing the use of this information to estimate the genome size, as well as to inspect the heterozygosity rate and repeat content. Calculations were conducted using the formula "genome size = k_num/Peak_depth," where k_num is the total number of k-mers and Peak_depth is the expected value of k-mer depth. Generally, k is 17, as is the case in the k-mer analysis of DH1. The k_num value is 28,870,387,824 and the peak depth is 61, respectively (**Supplementary Table 1**). Thus,

Nature Genetics: doi:10.1038/ng.3565

the DH1 genome size was estimated to be 473.3 Mb which is consistent with earlier flow-cytometry analysis, 473 Mb[3].

1.3 *De novo* assembly of the carrot genome

A schematic view of the carrot genome assembly pipeline is shown in **Supplementary Figure 1**. The following terms are used to describe the carrot genome assembly v2.0. A contig is a contiguous genomic sequence that does not contain unknown bases ("N"s) and that was not merged into scaffolds or superscaffolds. Scaftigs are defined as all portions of a final assembly (superscaffolds, scaffolds, and contigs) consisting of contiguous sequence, with gapped sequences split into separate scaftigs at every occurrence of unknown bases (Ns). A scaffold is a portion of the genome sequence reconstructed from end-sequenced whole genome shotgun fragments and composed of contigs with associated gaps. A superscaffold is constructed from two or more scaffolds that are connected using 1,000 "N"s. Joining of scaffolds into superscaffolds was supported by PE-BAC sequences. The final assembly is non-redundant, *i.e.* sequences or portions of sequences used in higher level constructs are removed from the lower category, *e.g.* if a contig is present in a scaffold, it is no longer present in the final assembly as a contig. Chromosome pseudomolecules were constructed from superscaffolds, scaffolds, and several contigs based on linkage maps. The nomenclature for sequences is DCARv2_Chr# for the nine chromosome pseudomolecules, DCARv2_B# for superscaffolds, DCARv2_S# for scaffolds, and DCARv2_C# for contigs, in which DCAR indicates carrot genome, and is also the registered locus tag prefix for gene annotations, v2 represents the second version of the genome assembly, and "#' is the unique numeric identifier for the sequence. The scaftig terminology was used only to describe the statistics of the assembly and not to label sequences in the final assembly.

1.3.1 Phase I – *De novo* assembly of Illumina reads

Quality filtered Illumina data were assembled using SOAPdenovo version 2.04 (ref. 4) (http://soap.genomics.org.cn). To simplify both assemblies and reduce computational complexity, SOAPdenovo employs the de Bruijn graph algorithm. The SOAPdenovo assembly

Nature Genetics: doi:10.1038/ng.3565

was divided into three steps. First, contigs are constructed by splitting the short-insert-size (170, 280 and 800 nt) library data into k-mers and constructing de Bruijn graphs. These graphs are then simplified by removing low-coverage edges and low-coverage k-mers and merging bubbles to produce contigs. Second, scaffolds are constructed by realigning all of the usable paired-end reads onto the contig sequences and calculating the shared paired-end relationships between each pair of contigs. Scaffolds are then constructed by weighing the rate of consistent and conflicting paired-ends. Finally, gaps are filled using GapCloser for SOAPdenovo[4] which uses the paired-end information to retrieve the read pairs in which one end is mapped to the unique contig and the other is located within a gap region. These read pairs are then used to perform a local assembly to fill gaps. For the DH1 genome, approximately 5.35×10^{10}nt of cleaned reads and k=45 were used to construct contigs, and 3.48×10^{10}nt were then used to construct the scaffolds. After filling the gaps, the carrot assembly v1.0 resulted in 4,182 scaffolds covering 418 Mb with an N50 (50% of the genome is in fragments of this length or longer) of 8.073×10^{5}nt kb (**Supplementary Table 51**) and 3,914 contigs covering 5.4×10^{6}nt with an N50 of 1.8×10^{3}nt. Scaftigs covered 3.906×10^{5}nt with an N50 of 3.11×10^{4}nt.

1.3.2 Phase II – Superscaffolding and correction of chimeric regions

1.3.2.1 Construction of a high quality integrated carrot linkage map

Genetic maps of three populations were used to establish a consensus map suitable for guiding the construction of superscaffolds and pseudomolecules. Each linkage map consisted of a "full" dataset including all the segregating markers mapped in each population and a "bin" data set, consisting of markers representing unique recombination events (**Supplementary Table 52**). A brief description of each map is reported below:

- The F_2 population 70349, consisting of 187 individuals, resulted from an original cross between P4201 (an inbred line with purple outer phloem and yellow xylem storage roots and purple petioles) and B6320 (an inbred with orange roots and green petioles derived from the European open-pollinated cultivars Nantes and Camberly). The 70349 genetic map included 894 co-dominant markers (482 bins) with known sequence information[5].
- The F_2 population Br1091×HM1, consisting of 138 individuals, resulted from an original cross between Brasilia, a Brazilian open pollinated variety that led to the discovery of *M.*

Nature Genetics: doi:10.1038/ng.3565

javanica resistance[6], and Homs (a Syrian open pollinated variety, with purple roots). The Br1091×HM1 genetic map included 843 co-dominant markers (367 bins) with known sequence information[7].

- The F_4 population 70796, consisting of 150 individuals, resulted from an original cross between B493 (an inbred with orange roots) and QAL (a wild carrot with white branching roots collected in Wisconsin-USA). The 70796 genetic map included 920 co-dominant markers (304 bins) with known sequence information[5].

The construction of each linkage map has been previously described[5], but it is worth mentioning that, to ensure the quality of each genetic map, the marker order was examined using CheckMatrix (http://www.atgc.org/XLinkage) to identify and remove markers with inconsistent placement due to false double recombination events. Before merging the data from each population for map integration, the co-linearity of common markers was inspected using MapChart 2.2 (ref. 8), and markers that were inconsistent were removed. In total, the three linkage maps shared 567 markers in the full dataset and 228 markers in the bin dataset. Both the "full" set and the "bin" dataset were used to generate a "full" and a "bin" integrated map using JoinMap 4.0 software[9]. Population-specific locus genotype scores were then integrated into one dataset in each Linkage Group (LG) using the Combine Groups for Mapping Integration Module, followed by locus ordering by the Regression Mapping Module of JoinMap. The following parameters were used for the calculation: Kosambi's mapping function, LOD ≥ 3.0, REC frequency ≤0.4, goodness of fit jump threshold for removal of loci = 5.0, number of added loci after which a ripple is performed = 1, and third round = yes. Markers in common were used as anchor points. The integrated maps resulted in 2,073 markers for the full dataset and 918 markers for the bin dataset (**Supplementary Table 53**), covering 622 cM and 616 cM, respectively. Following integrated map construction, marker-order correlations between composite and component map linkage groups were calculated in SAS 9.2 (SAS Institute, Cary, USA) using the PROC CORR Spearman function. Pair-wise linkage group marker-order correlations were high (greater than 0.98, $p<0.0001$; (**Supplementary Table 54)** reflecting the high co-linearity shown between common markers. Map positions of markers with known chromosome location were used to anchor LGs. After being assigned to chromosomes, LGs were oriented and numbered

Nature Genetics: doi:10.1038/ng.3565

following the chromosome orientation and classification established by Iovene *et al.* 2011 (ref. 10).

1.3.2.2 Superscaffolding and correction of chimeric scaffolds

An integrated approach was used to build superscaffolds, identify chimeric scaffolds and correct them. Three sources of sequences were aligned against the carrot assembly v1.0. They included:

- 29,875 PE BACs (**Supplementary Table 50**). Using blastn[11] with a 95% coverage, 99% similarity and minimum length 300 nt, 8,057 PE BACs unambiguously (both ends) aligned to the carrot assembly v1.0.
- All the 20 and 40 kb MPE Illumina data (**Supplementary Table 50**). Using Bowtie v2.1.0 (ref. 12) with –k 4 and –sensitive parameters, 100% coverage, over 2.0 and 8.1 M of the 20 kb and 40 kb MPE data respectively, aligned to the carrot assembly v1.0.
- Sequences of 2,075 molecular markers mapped in the carrot integrated linkage map (**Supplementary Table 53**). Using 90% coverage and 95% similarity, 1,980 markers unambiguously aligned to the carrot assembly v1.0.

For each scaffold or contig, unambiguously aligned sequences were visualized in GBrowse[13]. A custom program was used to visualize connections of PE sequences (PE BACs, 20 and 40 kb MPE) to other scaffolds or contigs. Superscaffolding was initiated with scaffolds containing sequences of mapped markers. Scaffold connections supported by at least two PE BACs were annotated and then the sequences were further connected using a custom Perl program. During this process the quality of each scaffold assembly and contiguity was verified by visually inspecting the coverage of large insert libraries (20 and 40 kb) and the consistency of marker order along the linkage map.

Possible chimeric scaffolds were identified as:

- Scaffolds containing sequences of markers mapped to different LGs or to distal locations of the same LG;
- Scaffolds with regions not covered by MPE sequences.

Nature Genetics: doi:10.1038/ng.3565

An example of a chimeric scaffold, and its correction, is reported in **Supplementary Figure 2**. Within each chimeric scaffold the chimeric region was identified as those sequences not covered by MPE or PE-BAC sequences. Those regions were than manually inspected. The mid-point between the closest unambiguously aligned PE sequences flanking the chimeric region was defined as the misassembly point. This ensured a correction of the chimeric sequences in a narrow window ranging from 20 to 40 kb. The coordinates of the misassembly points were annotated and further used to split the sequences. The corrected scaffolds were then used to progressively construct superscaffolds as described above. Adjacent scaffolds in each chromosome were separated by 1,000 "N"s.

With this approach we identified and corrected 135 scaffolds with one or more chimeric regions. We then merged 881 scaffolds into 89 superscaffolds covering about 90% of the assembled genome. The average distance between pairs of PE BACs spanning scaffold-scaffold junctions was 159 kb ±33 which is highly consistent with the estimated insert size of BAC clones (150±70 kb).

The carrot assembly v1.1 covered 425.6 Mb with an N50 of 12.7 Mb, representing 90% of the estimated genome size. The assembly contains 3,853 contigs with an N50 of 1.7 kb, 3,418 scaffolds with an N50 of 65.4 kb and 89 superscaffolds with an N50 of 13.4 Mb accounting for 89.8% of the assembled genome (**Supplementary Table 55**). Scaftigs cover 390 Mb with an N50 of about 31.0 kb.

1.3.3 Phase III – Pseudomolecule construction

The integrated linkage map was used to construct pseudomolecules (chromosomes). Scaffolds and superscaffolds were assigned to a chromosome (linkage group) if they contained at least one SNP marker from the "full" set linkage map. At least three markers mapped in the "bin" set linkage map representing unique recombination events were required to orient each sequence. However, some scaffolds had just one mapped marker and therefore were not oriented. In total 60 sequences were anchored to the nine linkage groups (named Dc-Chr 1 through Dc-Chr 9 based on the LG group assignment and orientation), and 52 sequences were anchored by at least three markers, allowing for a confident determination of their orientation. Adjacent

Nature Genetics: doi:10.1038/ng.3565

superscaffolds in each chromosome were separated by 2,000 "N"s. The total length of anchored sequences was 361.2 Mb, which accounts for 84.8% of the carrot genome assembly (425.6 Mb) (**Supplementary Table 2, Supplementary Figures 29-30)**. The average ratio between genetic-to-physical distance was 576.9 kb/cM, with one recombination event every 388 kb.

To estimate the percent of unassembled reads, BWA-MEM[14] with default parameters was used to align the three short insert size PE Illumina datasets (170, 280 and 800 nt, SRA accessions SRX1135260, SRX1135259, and SRX1135261) to the carrot assembly v2.0. On average only 0.4% of the reads did not align to the genome assembly, indicating that the majority of the remaining fraction of unassembled genome (~10% considering the estimated genome size, 473 Mb) is likely duplicated sequences that did not contribute to the cumulative physical coverage of the assembly (**Supplementary Table 3**).

1.4 Organellar genomes

1.4.1 Assembly

The plastid genome was independently assembled using two methods. First, an assembly with GS De Novo Assembler version 2.7 (454 Life Sciences Corp, CT USA) was performed using all Roche 454 reads (SRA accession SRX1135252). Plastid contigs were identified using MUMmer version 3.23 (ref. 15) and connections between contigs visualized using bb.454contignet[16]. Connected contigs with appropriate coverage were concatenated, producing a complete circular plastid assembly. Homopolymer errors were resolved by mapping a 10% subset of the 280 nt Illumina library reads to the assembled sequence with GnuMap version 3.0.2 BETA[17], and selecting the most abundant length for homopolymer regions ≥5 nt.

A second plastid assembly was performed by assembling the same reads with MIRA version 3.4.1.1 (ref. 18). A single contig was present which corresponded to the plastid genome single copy regions, and one copy of the inverted repeat. Homopolymer errors were resolved as above.

Nature Genetics: doi:10.1038/ng.3565

Minor differences between the two assemblies were reconciled by mapping Illumina reads to the new assemblies with bowtie version 2.1.0 (ref. 12), and manually confirming 51 small areas of discrepancy between the two assemblies. The resulting circular assembly is 155,848 nt.

The mitochondrial genome was assembled by first performing a *de novo* assembly with MIRA version 3.4.1.1 (ref. 18) using the 280, 2k, and 5k Illumina libraries (SRA accessions SRX1135259, SRX1135263, and SRX1135266 respectively). MUMmer[15] was used to find assembled contigs which aligned with the carrot mitochondrial genome, GenBank accession NC_017855. Read pairs contributing to these contigs were extracted, and contigs linked by paired reads, and their corresponding read pairs were also extracted. A second assembly with MIRA was performed with this subset of reads. Contigs were again mapped with nucmer, and connections visualized with Circos[19]. Contigs were then assembled into scaffolds using gap5 (ref. 20) and finished manually. This assembly resulted in a single linear molecule of 244,980 nt.

1.4.2 Annotation

Manual annotation of both organellar genomes confirmed presence of all previously identified genes. In addition, the DH mitochondrial assembly included genes not previously identified in the carrot mitochondrial reference sequence, GenBank accession NC_017855.1. These are *Rpo*, a DNA-directed RNA polymerase (locus tag DCAR_032461), *Dpo*, a DNA-directed DNA polymerase, previously described by Robison and Wolyn[21], GenBank accession AY521591.2 (DCAR_032462), and *Orf320*, a hypothetical protein 2-like, a portion of which was described previously in GenBank accession AY061991.1 (DCAR_032463).

The region of the mitochondrial genome which includes these additional genes, from 234,740 to 244,979, has a much higher read coverage, and this region may actually be a mitochondrial plasmid as described by Robison and Wolyn[21]. Other sequence differences include a 12 nt insertion in *Rpl2* exon 2. These data demonstrated that a large amount of genetic variation exists at the organelle genome level even between samples sharing a very close genetic relationship.

Nature Genetics: doi:10.1038/ng.3565

1.5 Assembly quality verification

The reliability of reference sequence data is crucial for the interpretation of downstream structural and functional genomic analysis. Thus, after the manual correction of the chimeric scaffolds, a comprehensive analysis was carried out to evaluate the quality of the final carrot genome assembly.

1.5.1 Analysis of sequence depth, GC content and sequence contamination

The read depth distribution was estimated by aligning Illumina GA reads onto the assembled sequence of the carrot assembly v1.0. Mapping was carried out using SOAPaligner (http://soap.genomics.org.cn/soapaligner.html) by allowing at most two mismatches. The number of aligned reads was then calculated for each position (**Supplementary Figure 3 panel A**).

Sequencing bias and contamination can influence the median GC content across the genome. Usually the genomic regions with high or low GC content will possess a low sequencing depth compared to the median GC content region. A custom program was used to calculate the GC content and depth average using the assembly v1.0 as reference (**Supplementary Figure 3 panel B**). The average GC content of the carrot genome was estimated around 35%, similar to that of other species (**Supplementary Figure 3 panel C**). The relationship between the average depth of coverage and the % GC frequency indicated there were no obvious sequence biases or contaminations.

In addition to the GC content analysis, presence of possible sequence contamination was evaluated using DeconSeq[22](http://deconseq.sourceforge.net/), a database of non-plant genomes. Scaftigs from the assembly V2.0 were split and used as input sequence into DeconSeq. The analysis indicated no sequence contamination.

1.5.2 Evaluation of sequence assembly consistency

1.5.2.1 Evaluation of sequence correctness using PE data

Two independent sets of paired-end data were used to evaluate the correctness of the assembled sequences. They included:

Nature Genetics: doi:10.1038/ng.3565

- PE data from a 454 library of DH1with an insert size of 8 kb (SRA accession SRX1135252). The library was prepared at the Biotechnology Center, UW-Madison (WI, USA) according the manufacturer's protocol (http://lifescience.roche.com/). Sequencing was performed with a GS-FLX platform, generating about 0.23 Gb (0.5M reads) of data. After sequencing, to accurately estimate the insert size, the PE reads were aligned to the carrot plastid genome and only PE reads that aligned with both ends to a unique location with coverage of 90% or more and an identity of 97% or better, and 90 nt minimum length, were used to calculate the distance between the two ends of a paired-end sequence. The average distance was estimated to be 8.3 ±2.3 (standard deviation) kb (**Supplementary Table 7**). The same procedure was used to align the 8 kb PE reads to the genome assembly v2.0 and estimate the distance between the PE pairs.
- 4,717 PE BACs that unambiguously aligned (filtering parameters described in 1.3.2.2) with both ends to the carrot genome assembly that were not used to join scaffolds into superscaffolds during phase II of the assembly process.

Assuming that the distance between two ends of a paired-end sequence represents their true physical distance, we estimated the observed distance between the two end sequences of the PE data based on their alignment against the carrot genome assembly. The fraction of PE data that aligned within the expected library insert size should reflect the fraction of assembled sequences that are consistently contiguous and correctly assembled. The results of this analysis were expressed as percent of PE reads that aligned within the average estimated insert size of the PE library (PE-BACs and 454), plus/minus twice the standard deviation. Overall, 99.4% and 95.6% of the 454 PE reads and PE BACs, respectively, unambiguously aligned within the estimated library insert size (**Supplementary Table 7**).

1.5.2.2 Linkage map construction and alignment to pseudomolecules.

To independently verify the order of superscaffolds along the nine pseudomolecules, F_2 population 85036, consisting of 84 individuals (unpublished data), was used to generate a linkage map including GBS SNP markers. Sequencing and library preparation was carried out at the Biotechnology Center, UW-Madison (WI, USA). DNA was quantified using Quantus PicoGreen ds DNA Kit (Life Technologies, Grand Island, NY) and normalized to 10ng/µl. 50 ng of DNA was used for each GBS reaction. GBS libraries were prepared as described by Elshire *et*

al.[23], with minimal modification and half-sized reactions. Briefly, DNA samples were digested with ApeKI, adapters ligated, and all samples pooled for sequencing and run on a single Illumina HiSeq 2000 lane, using paired end, 100 nt reads and v3 SBS reagents (Illumina, San Diego, CA). Primary analysis was performed with CASAVA 1.8.2.

The resulting reads were analyzed using TASSEL version 4.3.11 (ref. 24) with paired-end data preprocessed for TASSEL compatibility with bb.tassel (https://github.com/dsenalik/bb). SNPs were called using documented GBS pipeline procedures[25], with non-default parameter of mintagoccurrence=2. A total of 85,178 SNPs were obtained. Only sequences containing SNPs that unambiguously aligned to the carrot genome assembly were kept (18,007 SNPs). Finally, SNPs scored as heterozygous but with an allele ratio A:B far from 1:1 were eliminated if the ratio was < 0.3 or > 3.0, where A and B were the two alleles for a given SNP, leaving 516 high quality markers for linkage mapping.

Mapping was carried out with JoinMap 4.0 using the same parameters described above (see paragraph 1.3.2.1) and the order of markers across the linkage map was verified using CheckMatrix (http://www.atgc.org/XLinkage). Markers that were inconsistently placed due to false double recombination events were removed. The resulting map covered 450 cM and included 394 markers (**Supplementary Figure 31**). At LOD > 10, with less than 10% missing data for marker and genotype, 394 markers were grouped into nine linkage groups.

The 394 markers aligned to 36 superscaffolds covering 343.5 Mb (81%) of the assembled genome. Overall, 82.2% of adjacent markers matched the orientation of scaffolds and superscaffolds (**Supplementary Figure 4**). The markers in discordant alignment were on average 0.08 cM or 0.36 Mb apart. Using from one to three intervening markers, pairwise comparisons of the percent of concordant markers ranged from 91 to 98%, and the median genetic distance between discordant markers (2 to 9% of the total) ranged from 0.14 to 0.18 cM (**Supplementary Figure 5**). This strongly indicated that the majority of discordant markers likely reflect genotyping errors, missing data, insufficient marker density or low numbers of genotypes used to generate the linkage map.

Nature Genetics: doi:10.1038/ng.3565

1.5.2.3 Fluorescence *In Situ* Hybridization (FISH)

FISH experiments were carried out to evaluate the consistency and the coverage of the carrot genome assembly into telomeric regions. BAC clones which contained sequences that unambiguously aligned near the ends of pseudomolecules corresponding to Chr 1, 2, 4, 5, 6, 8 and 9 were used as probes.

To obtain meiotic preparations, immature umbels were collected from flowering plants of a DH1 line and fixed in Carnoy's solution (ethanol:glacial acetic acid, 3:1). Prior to slide preparations, umbels were washed clean of fixative solutions in distilled water (3×, 5 min each washing). Anthers isolated under a stereomicroscope were macerated in the enzyme mixture consisting of 4% (w/v) cellulose Onozuka R10 (Duchefa Biochemie) and 2% (w/v) pectolyase Y23 (Duchefa), in distilled water, pH 4.8 for 30 min at 37 °C. After digestion, one anther was transferred to a glass slide and preparation was performed as described previously[10].

Hybridization was carried out using three types of probes: (1) BAC probes specific for subtelomeric regions on the short (1S-2S-4S-5S-6S-8S-9S,) and long (1L-2L-4L-5L-6L-8L-9L) arms of each chromosome, (2) carrot chromosome-specific BAC probes[10] and (3) the telomeric probe (telo). Additional probes to test the location of a specific repetitive sequence were hybridized to Chr 1. A probe corresponding to CL80 repetitive sequence (see repetitive sequence analysis) was prepared by PCR amplification from DH genomic DNA using primers designed according to the consensus sequence of the monomer (**Supplementary Table 56)**. The PCR product was checked on a 1.5% agarose gel and a single band/fragment of expected size was cut, purified from the agarose gel and used as probe. The plasmid K11 containing the putative centromeric repeat of carrot (Cent-Dc) was also used as a FISH probe to detect the putative carrot centromeres[10]. A list of DH1 BAC clones used in this analysis is reported in **Supplementary Table 56**.

All probes except the telomeric were labeled with either biotin-16-dUTP or digoxygenin-11-dUTP using nick translation mix (Roche Diagnostic) following the manufacturer's protocol until the length of the probe fragments averaged about 100–500 nt. Labeled DNA was purified with Quick Spin G-50 Sephadex Columns (Roche Diagnostics) following the manufacturer's protocol. To identify telomere regions, a synthetic seven nt 'telo' probe (Cy5-TTTAGGG) was

Nature Genetics: doi:10.1038/ng.3565

used. The FISH procedure was carried out according to published protocols[26,27]. Carrot genomic DNA sheared up to 500 nt fragments was used as a blocking DNA in the hybridization mixture. Most BAC probes required an excess of 500× blocking DNA to reduce background signal while several probes an excess of 1,000× was necessary. Biotin- and digoxygenin-labeled probes were immuno-detected with 10 μg/ml of Alexa Fluor 488-conjugated streptavidin antibody (Life Technologies) and two μg/ml rhodamine-conjugated anti-digoxygenin antibody (Roche Diagnostics), respectively. Chromosomes were counterstained with one μg/ml of 4',6-diamidino-2-phenylindole (DAPI) in Prolong Gold antifade solution (Invitrogen). Slides were examined with AxioImager M2 Zeiss microscope. All images were captured digitally using BV MV System (Applied Spectral Imaging) and Case Data Manager 4.0 software (ASI). Results of FISH experiments are reported in **Figure 1** and **Supplementary Figure 6**.

1.5.2.4 Gene space coverage

Three analyses were used to assess gene space coverage of the carrot genome assembly. They were:

- 58,751 consensus carrot expressed sequence tags (ESTs) from Iorizzo *et al.*[28] were aligned to the genome using Blastn[11]. Only ESTs with alignment of identity ≥90% and coverage ≥50% were included.
- RNA-Seq assembly and alignment. 512.9M PE RNA-Seq reads from 20 sequencing libraries representing expressed sequences from 20 different DH1 carrot tissues, developmental conditions or developmental stages (NCBI BioSamples SAMN03965304–SAMN03965323) were assembled with Trinity r2013_08_14 (ref. 29). To estimate the library insert size, raw reads were mapped to the Trinity assembly with Bowtie v2.1.0 (ref. 12). Sequences were mapped to the carrot genome assembly using TopHat v2.0.11 (ref. 30) with the following non-default parameters: 1) insert size: as estimated with Bowtie; 2) min-intron-length 20; 3) max-intron-length 10000;
- Scaftigs from the carrot assembly v2.0 were aligned to 258 ultra-conserved genes from the Core Eukaryotic Genes Dataset using CEGMA v2.4 (ref. 31).

Nature Genetics: doi:10.1038/ng.3565

These analyses indicated that about 94% of the ESTs, 98% of RNA-Seq data and 99.9% of Core Eukaryotic genes aligned to the carrot genome assembly, providing evidence that the assembly covers the majority of gene space (**Supplementary Tables 4-6**).

Together, the assembly statistics and verification provided evidence that the assembly is of high quality. Compared to other genomes that used primarily NGS data, the carrot genome assembly is among the most complete published plant genomes, in terms of genome coverage and sequence contiguity length (**Supplementary Table 8**).

Nature Genetics: doi:10.1038/ng.3565

2. Genome characterization and annotation

2.1 Repetitive sequences

Mobile elements (MEs) were identified in the genome by a combination of *de novo* and homology based approaches. Tandem repeats were detected using Tandem Repeats Finder v4.07b[32]. To study the mode of amplification of DNA mobile elements in the carrot genome, a detailed characterization of two families of MITES was carried out by identification of terminal inverted repeats (TIRs) motifs.

In addition, to identify and characterize large clusters of tandem repetitive sequences such as telomeric and possible centromeric repeats across different *Daucus* species, a graph-based sequence clustering analysis was performed using RepeatExplorer[33].

2.1.1 Homology based prediction and *de novo* identification of repeats

Mobile elements in the genome assembly were identified at both the DNA and protein level. RepeatMasker v3.2.9 (http://www.repeatmasker.org) was applied to screen the genome assembly for low complexity DNA sequences and interspersed repeated elements using a custom library (a combination of Repbase v16.02 and plant repeat database). RepeatProteinMask (an extension of RepeatMasker) was used to perform RMBlast against the ME protein database to find known repeat sequences at the protein level.

Ab initio prediction program RepeatModeler version 1.0.4 (http://www.repeatmasker.org/RepeatModeler.html) was employed to build the *de novo* repeat library from the assembled genome, refined by removing the contaminated sequences possibly derived from bacterial and redundant duplicated sequences in the library. Using this library as a database, RepeatMasker was implemented to identify and classify homologous repeat elements in the genome. In addition, LTR_FINDER version 1.1.0.5 (ref. 34) was used to search the whole genome for the characteristic structure of the full-length long terminal repeat (LTR) retrotransposons. Subsequently, a custom program was used to merge all the predictions and generate a combined repetitive sequence annotation to mask the carrot genome.

Nature Genetics: doi:10.1038/ng.3565

ME accounted for 44.9% (190 Mb) of the assembled carrot genome (**Table 1**, **Supplementary Table 9**). This value is larger than those observed in other sequenced genomes of similar size, for example, grape[35] (41.4%, for 487 Mb) and melon[36] (20%, for 375 Mb). With 57.4 Mb, the fraction of class II transposable elements in the carrot genome is higher than in most other plant genomes including rice (48 Mb)[37]. A large fraction of MEs are of relatively recent origin, with a sequence divergence rate of less than 10% (**Supplementary Figure 32)**.

2.1.2 Characterization of *Krak* and *DcSto* elements

Methods

Miniature inverted repeat transposable elements (MITEs) belonging to two previously described groups, *Tourist*-like *Krak*[38] and *Stowaway*-like *DcSto*[39], carrying complete terminal inverted repeats (TIRs) were identified from the assembled genomic sequence using TIRfinder[40]. The following search parameters were defined on the basis of the previous reports: tirMask: GKGYCTGTTTGG and CTCCCTYYSKYMC, tsdMask: TWA and TA, tirSeqMismatches: 5 and 1, tsdSeqMismatches: 0 and 0, tirMaskMismatches: 2 and 0, tsdMaskMismatches: 0 and 0, for *Krak* and *DcSto*, respectively. The same parameters were used to mine for *Tourist*-like and *Stowaway*-like elements in kiwifruit, pepper, tomato and potato. Output multifasta files were manually curated to remove redundant hits and false positives, and group the remaining MITE copies into families fulfilling the 80-80-80 criterion[41].

Consensus sequences were used to investigate intra and interspecific relationships among families with Circoletto[42] a tool allowing visualization of similarity calculated by blastn[11] with Circos[19]. Prior to analysis, each family of *Stowaway*-like elements was manually inspected to identify internally rearranged copies carrying insertions >10 nt which were removed from subsequent steps. Within-family similarity was calculated from a Kimura 2-parameter pairwise distance matrix. Evolutionary history of related *DcSto* elements was investigated by aligning individual copies with ClustalW followed by clustering by NJ method[43] based on genetic distances computed using Kimura 2-parameter algorithm[44] using MEGA6 (ref. 45), with 1,000 bootstrap replications. The relative divergence time was presented as a time tree generated using RelTime method[46] with MEGA6.

Nature Genetics: doi:10.1038/ng.3565

Results

Krak and *DcSto* were represented by 404 and 4028 copies carrying intact terminal inverted repeats, respectively (**Supplementary Table 57**). Both groups included copies that were too divergent to meet the 80-80-80 criterion[41] and were divided into nine and 14 families, respectively. Grouping was confirmed by phylogenetic analysis and very well supported clades (**Supplementary Figure 33**). The most numerous family of *Krak*, *DcKrak1*, was represented by 152 copies, while *DcSto6* comprised 1008 copies. In general, insertions in both groups with respect to their distance from genes were localized in a very similar manner.

More than half of insertion sites (55.2 and 54.0% for *Krak* and *DcSto*, respectively) were inside or within a range of 2 kb from the nearest gene (data not shown). This distribution has prompted the hypothesis that *DcSto* and *Krak* elements are preferentially associated with genes.

To test this hypothesis we compared the distribution of a simulated random insertion sites with the observed distribution of *DcSto* and *Krak* elements (**Supplementary Figure 8**). The density of insertions overlapping with a gene prediction in the simulation was 32.6%, higher than the *DcSto* (29%) and *Krak* (28.4%) density, suggesting that there is no tendency to insert within a gene. The high percent of these two families of MITEs near genes might be due to their abundance. Comparison of the frequency of *DcSto* and *Krak* elements and the simulated insertions at different distance from genes indicated that the two curves are highly correlated (*DcSto*: r^2=0.94, $p<0.05$; *Krak*: r^2=0.61, $p<0.05$) and that there is no statistical difference between the two curves (*DcSto* p=0.99; *Krak* p=0.94), further supporting the hypotheses that *DcSto* and *Krak* elements are randomly distributed and not preferentially inserted near or within genes.

Compared to other Asterid species, carrot included more copies of *Stowaway*-like MITEs elements, consistent with our annotation indicating the expansion of DNA transposons in the carrot genome relative to other species genomes (**Supplementary Table 58**). Among other Asterid species, in tomato and kiwifruit only 29 and 17 copies were identified, respectively. Search for additional copies in tomato using a more relaxed TIRfinder mask only marginally increased the number of hits, indicating that *Stowaway*-like MITEs did not proliferate extensively in that species. *Stowaway*-like families in Solanaceae were interrelated

Nature Genetics: doi:10.1038/ng.3565

(**Supplementary Table 59**, **Supplementary Figure 34**) , at both intra- and inter-species level, indicating that the expansion of MITE families in the Solanaceae started before the divergence of the three species ca. 36 Mya[47]. Carrot *DcSto* families did not have any apparent relationship to those present in Solanaceae. In addition carrot MITEs had an evident lower level of intraspecific similarity (**Supplementary Figure 34**). The only interrelated group of *DcSto* families comprised *DcSto1*, *DcSto2*, *DcSto4*, *DcSto5* and *DcSto8*. Of those *DcSto2*, *DcSto5* and *DcSto8* were more closely related while *DcSto1* and *DcSto4* formed a more distant group (**Supplementary Figure 33**). This pattern of inter- and intra-specific relationship of carrot *DcSto* MITEs, point to a lineage specific evolution and a parallel expansion of multiple families followed by family specific diversification.

To investigate the mechanisms of the MITE amplification in the carrot genome, pairwise distances were calculated for each of the characterized *DcSto* families (**Supplementary Figure 7 a1-c1**). The histograms of most MITE families are wave-like curves with a sharp modal distribution. Four of them have a unimodal distribution, six have a bimodal distribution and two have a multimodal distribution. This pattern indicated that each family has experienced rapid population expansion (burst) during carrot genome evolution. Estimates of relative divergence times among and within interrelated *DcSto* families show differing branch lengths indicating that the amplification bursts occurred at different times (**Supplementary Figure 7 a2-c2**). The RelTime tree of unimodal histogram for *DcSto*7b (**Supplementary Figure 7 a2**) indicated uniform relative divergence time, except for a few ancestral elements, suggesting rapid amplification from few master members. Therefore we conclude that some MITE families resulted from one amplification burst, whereas other families have experienced multiple rounds of amplification during the carrot genome evolution.

2.1.3 Characterization of the main tandem repetitive sequences

Material and Methods

RepeatExplorer analysis[33] was performed using a set of one million randomly distributed paired-end Illumina reads from DH1. More detailed investigation of cluster graphs was performed using the SeqGrapheR program[48]. Clusters containing tandem repetitive sequences are

Nature Genetics: doi:10.1038/ng.3565

characterized by a globular cluster layout due to the high similarity of the repetitive units. To select for potential tandem repetitive sequences, a custom program was developed to calculate the node to edge ratio (number of nodes/no. of edges) among aligned sequences in the cluster. Clusters with a ratio >0.09, representing more than 0.05% of the genome, were selected for further analysis. Tandem repeats were identified using Tandem Repeats Finder v4.07b[32] with parameters 2 7 7 80 10 50 500, and filtered for ≥70% similarity.

After the analysis of DH1 we extended our analyses to five *Daucus* species to search for the presence of the candidate tandem repeat families that we had identified in DH1. These five species were representatives of the two clades recognized in the genus *Daucus*[49] and included: 1) *D. syrticus* (Ames 29108, 2n=2x=18) and 2) *D. aureus* (PI 319403, 2n=2x=22) from *Daucus* clade I; and 3) *D. guttatus* (PI 286611, 2n=2x=20); 4), *D. littoralis* (PI 295857, 2n=2x=20), and 5) *D. pusillus* (PI 349267, 2n=2x=22), from *Daucus* clade II. One million paired-end reads from each species were subjected to a pairwise comparative analysis using the RepeatExplorer pipeline with the comparative analysis settings (http://repeatexplorer.umbr.cas.cz/static/html/help/manual.html#example-history-2-comparative-analysis-of-repeats-between-two-genomes). A custom Perl program was used to extract the reads from each pairwise comparative analysis and select the clusters containing the sequences that had similarity hits to the tandem repeats of the DH1 genome. In addition, we also searched for the presence of tandem repeat (TR) clusters specific of each species.

The abundance and localization of selected repetitive sequences in DH1 and other *Daucus* species was also investigated by Fluorescence in Situ Hybridization (FISH). The probe for the CL80 repeat was prepared by PCR amplification from DH1 genomic DNA as described above (see section 1.4). Chromosome and probe preparation, FISH and re-probing of the same slides were performed according to published protocols[50,10]. Plasmid K11 containing several Cent-Dc monomers was used to highlight the centromeres of DH1. Slides were examined under a Leica DM6000B epifluorescence microscope. Images were captured with a DFC365 FX CCD camera and LAS AF software (Leica Microsystems). Final contrast of the images was adjusted in Adobe Photoshop v. 6.0.

Results

Nature Genetics: doi:10.1038/ng.3565

The analyses performed provided a unique opportunity to study the evolutionary dynamics of TRs across a set of related species with different phylogenetic relationships. The six species included in this analysis are representatives of the two clades recognized in the genus *Daucus*[49,51]. Species relationships in this genus are not well resolved and there are limited data for the estimations of the divergence times among *Daucus* species. According to Spalik *et al.*[52], the *Daucus* genus started diversifying about 20 Mya. The species used in this study also differed for their chromosome numbers, including 2n=18, 2n=20 and 2n=22.

Analysis of the DH1 sequences led us to identify four families of satellite DNA (**Supplementary Tables 10**, **60**). These repeats accounted for about 7% of the genome, which represented a larger fraction of TR than that estimated in the assembled genome (1.4%). Chromosomal regions rich in satellite DNA create a technical challenge to the assembly efforts and, indeed, these regions are often incomplete or absent even within extensively sequenced genomes. Our results confirmed the power of this approach to capture underrepresented portions of the assembled carrot genome. The length of the satellite repeat units ranged from 159 nt (CL1) to 170 nt (CL8). The CL81 repeat was specific to the DH1 genome and it was not detected in any other *Daucus* species analyzed. In the present work we focused on the two most abundant tandem repeat families of the genus *Daucus*, which were identified through our comparative analysis and were represented by DH1-CL1 and DH1-CL80 (**Supplementary Figure 9**).

DH1-CL1, the most abundant tandem repeat of the DH1 genome, corresponded to the putative carrot centromeric satellite Cent-Dc[10]. As reported, the typical Cent-Dc unit of 159 nt was made up of shorter monomers of 39-40 nt, suggesting that Cent-Dc-159 nt could represent a higher-order repeat (HOR) structure (**Supplementary Figure 9**). The large amount of sequence data generated by this study, along with the BAC end sequences (BES) of the DH1 BAC library, allowed us to perform a genome-wide analysis of the Cent-Dc sequences, including the detection of the most frequent HOR structure(s) and the main variants. Our sequence analysis demonstrated that Cent-Dc units of 159 nt indeed represented a HOR made up of four monomers (A, B, C, D) of 40 nt, 40 nt, 40 nt and 39 nt, respectively. Each 39-40 nt monomer in DH1 had accumulated private polymorphisms (for a total of 15 private polymorphisms) (**Supplementary Figure 9 panel II**) similar to the Cent-Dc sequences contained in the plasmid K11. The average pairwise similarity among sequences of individual 39-40 nt monomer (A versus B, C, and D; B

Nature Genetics: doi:10.1038/ng.3565

versus C, and so on) was lower than the similarity between adjacent 159 nt monomers **Supplementary Table 61**), which is a typical feature of HORs. A blastn[11] search against DH1 BES resulted in the identification of 385 BES containing Cent-Dc sequences. More than 93% of these BES contained one to four Cent-Dc-159 nt, each consisting of the typical A-B-C-D HOR structure. Although multimers with different monomer combinations (for example "BACD") and/or missing one of the four monomers (for example "ACD" missing B) were observed, it was not possible to identify an alternative predominant multimer combination, suggesting a relatively uniform distribution of the "ABCD" HOR structure in the carrot genome. Phylogenetic analysis of individual 39-40 nt monomers (A, B, C, D) extracted from 17 BES indicated that monomers located at equivalent positions in the 159 nt (multimeric) units are highly homologous (**Supplementary Figure 9 panel III**). Comparative analysis with other *Daucus* species revealed that Cent-Dc represented the most abundant TR sequence also in *D. syrticus* (2n=18; named Ds-CL1) and *D. aureus* (2n=22; Da-CL1), both belonging to the clade I of *Daucus*. However, while in *D. syrticus* this repeat was organized in a HOR structure that was similar to the DH1 Cent-Dc, no such structure was found in *D. aureus.* Indeed in *D. aureus*, Cent-Dc homolog sequences (Da-CL1) have not accumulated private polymorphisms indicating that the repeat is present as a single monomer of 40 nt (**Supplementary Figure 9 panel II, Supplementary Table 61**). Interestingly, the most abundant tandem repeat (2.2% of the genome) in *D. pusillus,* Dp-CL5, (2n=22, belonging to *Daucus* clade II) had only a weak similarity with DH1 Cent-Dc. However, the initial 40 nt of the Dp-CL5 monomer of *D. pusillus* shared >82% similarity with DH1 40 nt Cent-Dc monomer A (**Supplementary Figure 9 panel IV**). The remaining portion of the Dp-CL5 monomer was different from Cent-Dc, suggesting the loss of the Cent-Dc-like motif in the flanking sequences and/or insertions during its evolutionary history. Alternatively, we might speculate that Dp-CL5 may represent a highly degenerated HOR structure, and that the ancestral 39-40 nt monomers have acquired such a massive amount of polymorphisms that they cannot be recognized as individual monomers anymore. Overall these data suggested a common origin for the Cent-Dc, Ds/Da-CL1 and Dp-CL5 satellite families, which predated the divergence of the *Daucus* species. However, Cent-Dc was not detected in *D. guttatus* and *D. littoralis* resequencing data. FISH using the plasmid K11 (which contains Cent-Dc sequences and was previously Sanger-sequenced) did not generate any signal in either *D. guttatus* and *D. littoralis* (**Figure 1**), confirming our *in silico* analysis.

Nature Genetics: doi:10.1038/ng.3565

The CL80 repeat was the other most abundant satellite of the genus *Daucus*. CL80 was detected in species of both clade I and II, suggesting that its origin predated the divergence of the two clades. In each species carrying the CL80 sequence, the length of the consensus TR was 169 nt with a pairwise average similarity (between any two species analyzed) of 96.5%. However, the abundance of CL80 differed among the species analyzed (**Supplementary Table 10**). *In silico* analysis of the distribution of CL80 across the nine assembled pseudomolecules revealed that over 58% of the clustered sequences localized on chromosome 1 at the junction between superscaffold 7 and 8 (**Figure 1**). The remaining sequences localized in short non-anchored contigs and scaffolds. FISH analysis of CL80 confirmed its location on carrot chromosome 1 and its organization in a tandem array (**Figure 1**). The CL80 repeat was associated with a previously reported knob of the long arm of carrot chromosome 1 (ref. 10), between carrot BACs 20G08 (superscaffold 7) and 20P12 (superscaffold 8). The CL80 TR accounted for 2.4% and 3.9% of the genomes of *D. guttatus* and *D. littoralis* (both 2n=20), respectively, and represented the most abundant tandem repeat in these species. It should be noted, however, that due to the read length and the parameters we used in the analysis, the presence of other potential (and longer) tandem repeats could not be ruled out. CL80 was not detected in the genomic sequences of *D. aureus* (2n=22) and *D. syrticus* (2n=18). FISH analysis confirmed that CL80 was amplified in the genomes of *D. guttatus* and *D. littoralis* (**Supplementary Figure 10**). In addition, it revealed that CL80 distribution was not conserved among these species. The CL80 repeat was detected on all chromosomes of *D. littoralis*, at both distal (subtelomeric) and intercalary regions (**Supplementary Figure 10, A-I**). Intercalary signals likely spanned the centromeric regions of *D. littoralis* chromosomes (**Supplementary Figure 10, D-F)**. Thus, CL80 could be a candidate centromeric satellite repeat in this species. Each chromosome had either two CL80 hybridization sites (that is, one intercalary signal and a distal one; four chromosomes), or three sites (that is, signals at both ends plus an intercalary signal; 16 chromosomes) (**Supplementary Figure 10, A-C)**. The subtelomeric locations of CL80 were confirmed by co-hybridization with an end-labeled $(TTTAGGG)_4$ oligo-probe on meiotic pachytene chromosomes of *D. littoralis* (**Supplementary Figure 10, G-I**). CL80 hybridized at most chromosomal ends of *D. guttatus* (**Supplementary Figure 10, J-O**). Intercalary signals were detected on four chromosomes, however these signals appeared to localize to the pericentromeric regions rather than the centromeres. The sizes and intensities of the FISH signals varied among different chromosomes, likely reflecting differences

Nature Genetics: doi:10.1038/ng.3565

in copy numbers of the repeats at different sites. Several FISH signals were weak. We counted up to 28 subtelomeric and four intercalary signals (that is, eight chromosomes with signals at both ends; six chromosomes with one subtelomeric signal; two chromosomes with one intercalary and one subtelomeric signal; two chromosomes with signals at both ends plus an intercalary signal; two chromosomes with no detectable signal). CL80 resembled, in several aspects, the satellite repeat CentO-C2/TrsC identified in *Oryza rhizomatis*. Similar to CL80, CentO-C2/TrsC localized at both subtelomeric regions and functional centromeres of several *O. rhizomatis* chromosomes[53]. In addition, CentO-C2/TrsC hybridized exclusively to the subtelomeric regions of the chromosomes of the related species *O. officinalis*[54]. Lee *et al.*[53] hypothesized that this repeat had originated outside the functional centromere also based on its higher copy number at the subtelomeres. It will be interesting to analyze additional *Daucus* species to better understand the evolutionary dynamics of the CL80 repeat. If future research confirms the association of CL80 with the functional centromeres of *D. littoralis*, it will also be interesting to investigate what sequences occupy the centromeres of *D. guttatus*.

2.2 Gene prediction and annotation

Prior to gene prediction, all ME-related repeats were masked using RepeatMasker. Gene prediction was based on the integration of *de novo* gene prediction and evidence-based predictions. Evidence-based prediction included protein-based homology searches from closely related or model species and ESTs/RNA-Seq experiment data aided prediction.

2.2.1 Gene prediction and functional database annotation

De novo prediction was carried out using AUGUSTUS v2.5.5 (ref. 55), GENSCAN v.1.1.0 (ref. 56) and GlimmerHMM-3.0.1 (ref. 57), All programs were trained on the model species *A. thaliana* and *S. lycopersicum* training sets.

The protein sequences of *S. lycoperisum* (v2.3, ftp://ftp.solgenomics.net/genomes/Solanum_lycopersicum/annotation/ITAG2.3_release/), *S. tuberosum* (v3.4, ftp://ftp.solgenomics.net/genomes/Vitis_vinifera/annotation/Genoscope_12X_2010_03_19/), *A.*

Nature Genetics: doi:10.1038/ng.3565

thaliana (TAIR10.0, ftp://ftp.solgenomics.net/genomes/Arabidopsis_thaliana/), *B. rapa* (v1.0, https://phytozome.jgi.doe.gov/) and *O. sativa* (IRGSP build 5, http://rapdb.dna.affrc.go.jp/download/build5.html) were mapped to the carrot genome using tblastn[11] (Blastall 2.2.23) to generate protein-based gene models. Only hits with an E-value <1e-5 were retained and further analyzed with GeneWise 2.2.0 (ref. 58) to search the most accurate spliced alignments.

Carrot ESTs[28] were aligned to the genome using BLAT[59] with the following parameters: identity ≥0.90, coverage ≥0.90. This generated putative spliced sequence alignments. The output of BLAT was then analyzed with PASA[60] to detect the spliced gene models.

Besides carrot ESTs, we generated 513 million RNA-Seq reads from 20 libraries (**Supplementary Table 11**). To identify the accurate splice junctions between exons, we first used the fast gap-alignment RNA-Seq mapper TopHat v2.0.9 (ref. 30) to align the RNA-Seq reads to the carrot genome. The alignment results were then used as input for Cufflinks[61] to obtain a set of assembled transcripts, which can be taken as the candidate gene models integrated with other evidence.

All gene models produced by *de novo* prediction, protein-homology searches, and prediction and transcript based evidence were integrated using GLEAN v1.1 (ref. 62). The final gene set yielded 32,113 genes. The majority (98.7%) of the gene predictions had cDNA-EST expression evidence (**Supplementary Table 13**), demonstrating the high accuracy of gene prediction. The mean coding sequence size was 1,183 nt (**Supplementary Table 11**), similar to other annotated genomes, with an average of 4.99 exons per gene. About 89% of the genes have either known homologs or can be functionally classified (**Supplementary Table 12**).

Putative gene functions were assigned according to the best match of the alignments using blastp[11] (E-value≤10^{-5}) to SwissProt and TrEMBL databases. The motifs and domains of genes were determined by InterProScan version 4.7 (ref. 63) against protein databases including ProDom, PRINTS, Pfam, SMART, PANTHER and PROSITE. Gene Ontology IDs for each gene were obtained from the corresponding InterPro entries. All genes were aligned against KEGG (Release 58) proteins, and the pathway in which the gene might be involved was derived from

Nature Genetics: doi:10.1038/ng.3565

the matched genes in KEGG. Gene models with no match in these databases were labeled "hypothetical proteins."

To identify over- and under-represented carrot InterPro domains, gene annotation of the carrot genome was compared with kiwi, potato, tomato and grape. The input datasets from carrot and the other sequenced genomes are described in section 3.1. For each species, the cumulative number of unique IPR domains were used to apply Fisher's exact test ($p<0.01$). Carrot was found to be enriched for genes involved in a wide range of molecular functions including selective molecule interactions (binding), oxidoreductase activity, secondary metabolism and diverse cellular functions (**Supplementary Table 14**).

2.2.2 Non-coding RNA prediction and annotation

We searched candidate microRNAs and snRNAs in the assembled carrot genome using INFERNAL[64] against Rfam database (Release 9.1). We identified tRNA using tRNAscan-SE v1.1.23 (ref. 65). Although highly repetitive and likely not fully assembled, several ribosomal RNA sequences present in the assembled nuclear genome sequence were detected by homologous blastn[11] searches using the closest available species with complete sequences, *Panax ginseng, P.quinquefolius*, and *Thapsia garganica* (accessions KM036295.1, KM036296.1, KM036297.1, and AJ007917.1). We detected 27 intact 5S rRNA motifs, and 4 nearly intact 18S-5.8S-26S motifs. A summary of the results is reported in **Supplementary Table 15.**

Nature Genetics: doi:10.1038/ng.3565

3. Resequencing

3.1 Materials and Methods

3.1.1 Plant materials for resequencing

To evaluate a broader range of carrot genetic diversity and determine population structure, resequencing was conducted in 35 accessions covering a wide range of genetically and phylogenetically diverse materials, NCBI BioProject PRJNA291976 (BioSamples SAMN03766317–SAMN03766351). Included were 18 cultivated accessions (*D. carota* subsp. *sativus*), thirteen wild accessions (eight *D. carota* subsp. *carota*, four *D. carota* subsp. *gummifer*, one *D. carota* subsp. *capillifolius)* and four other *Daucus* species (**Supplementary Table 16, Supplementary Figure 11**). We defined eastern carrot as accessions from the Middle East, Central Asia and Eastern Asia and western carrot as those from the Middle East, Europe, Africa, and North and South America. Cultivated carrots included 14 open-pollinated cultivars and local land races and four inbred lines. White, yellow, purple and orange root types were represented in this collection of samples. Wild carrots (*D. carota* subsp. *carota* and subsp. *gummifer*), other *Daucus* species and subspecies represent a collection of accessions that have been previously characterized morphologically and genetically by Arbizu *et al.*[49,51] to ensure accurate species designation.

DNA from single plants was extracted as described by Murray and Thompson[1] and quantified using Quantus PicoGreen ds DNA Kit (Life Technologies, Grand Island, NY). Paired-end libraries with insert sizes ranging from 250 to 350 nt were constructed according to the manufacturer's instruction (Illumina, San Diego, CA, USA). Sequencing was performed by Illumina sequencing technology at Beijing Genome Institute, Shenzhen (BGI-Shenzhen, China). Whole genome resequencing of 35 *Daucus* plants generated from 5.2×10^9 to 2.9×10^{10} nucleotides of sequence with an average of 1.06×10^{10} at a median depth of 14×.

3.1.2 Mapping, SNP detection, and Validation

We used BWA-MEM version 0.7.10 (ref. 14) to map the resequencing reads from all carrot genotypes to the carrot reference genome using the following parameters -a -M –t 42. Alignments were filtered using SAMtools version 0.1.19 (ref. 66) for only primary alignments with quality of at least 30, i.e. parameters -q 30 -F 256. Duplicate reads were marked using MarkDuplicates from Picard tools version 1.119 (http://broadinstitute.github.io/picard/). The GATK version 3.3-0 (ref. 67) was used to identify SNP variants for each genotype using the GATK best practices method using RealignerTargetCreator, IndelRealigner, HaplotypeCaller, and GenotypeGVCFs. Then SelectVariants was used to separate SNPs, indels, and other variants[68,69]. Reads used to construct the doubled haploid reference genome were also analyzed as a control, and variants that were also present here were filtered out with a custom Perl program. Variants were then filtered using VCFTools v0.1.12a[70] with parameters --maf 0.1, --min-meanDP 5, and --max-missing 1.

After filtering and variant detection with GATK from 39,695,937 SNP variants we generated 1,393,431 filtered SNPs. From this SNP set, 49,365 biallelic SNPs were randomly selected and used for both population structure and phylogenic analyses. After selection, the selected SNPs were plotted across the genome to ensure even coverage (data not shown).

The accuracy of SNP calls was evaluated using a set of 3,202 previously characterized SNPs[2]. Since this set was previously evaluated across a wide range of wild and cultivated accessions, including many accessions used in this study, we consider this a high-quality germplasm diversity set. Using a custom Perl script, we evaluated the fraction of these high quality SNPs that were detected in the resequencing set. In total 3,056 (95.2%) SNPs matched previous SNPs and allele calls, 114 (3.56%) SNPs matched the polymorphic site but were not the previously identified allele. Finally, only 32 (<1%) out of 3,202 SNPs were not detected. This analysis demonstrated the accuracy of detected SNPs, which will provide a valuable resource for biological discovery and germplasm improvement in carrot.

Nature Genetics: doi:10.1038/ng.3565

3.1.3 Population Structure and Phylogenetic Analysis

To determine the population structure of the 35 resequenced diverse genotypes, 49,365 biallelic SNPs were randomly selected from the 1,393,431 original filtered SNPs using the GNU/Linux shuf program, and custom Perl scripts. VCFTools[70] was used to convert the VCF data to PLINK PED and MAP files (PLINK v1.070)[71]. PGDSpider (v2.0.7.2)[72] was then used to convert the data from PLINK format to STRUCTURE[73] format. STRUCTURE (v2.3.4) analysis was replicated ten times with 20,000 burn-ins and 10,000 Monte Carlo iterations on these randomly selected SNPs for each of the estimated population sizes (K values) 1 to 8, using the admixture model with no previous population information. Inferalpha and computeprobs were set to 1, otherwise all other parameters were set to default values. The most accurate population structure was determined by the method discussed in Evanno *et al.*[74] using StructureHarvester (v0.6.94)[75]. Population structure was visualized using Distruct software (v1.1)[76].

Phylogenic analysis was completed using PHYLIP (v3.5)[77] with the subset of SNPs used for STRUCTURE analysis. The dataset for PHYLIP was created using the frequency function of VCFTools[70] and custom Perl scripts. Seqboot was used for bootstrapping with 1,000 replicates, and genetic distances were calculated using gendist. A neighbor-joining tree was created using the neighbor function and a consensus tree was generated using consense. All PHYLIP functions were performed using default parameters. The neighbor-joining tree was visualized using FigTree (v1.4.2) (http://tree.bio.ed.ac.uk/software/figtree/) (**Figure 2a**).

STRUCTURE analysis of 35 genotypes of cultivated carrot, wild *Daucus carota* and other related *Daucus* species identified seven population clusters ($K = 7$), as determined by the ΔK method[74] is presented in **Figure 2b**.

Nature Genetics: doi:10.1038/ng.3565

4. Genome evolution

To extend our knowledge about the carrot genome, and the association between genes and phenotypes, concepts should be considered at a level higher than that of the single genome, rather at a community genome scale. The available sequence of a growing number of plant genomes provides the means to extract biological knowledge through the detection of similarities and differences within and between genomes of closely or more distantly related species. Indeed, the best method we have to reconstruct the evolutionary past of any species is by comparison with its living relatives. Using such comparative approaches, (1) knowledge can be transferred from model to non-model organisms, (2) insights can be gained into the evolution of specific genes or entire metabolic and signaling pathways, (3) genes of importance for niche-specific plant adaptations can be identified and, (4) large-scale genomic events, such as whole-genome duplications (WGDs), can be unveiled. In this context, the density of the "genome community" will exponentially improve our ability to characterize a genome and associate genes with functions. Carrot belongs to the Euasterid II clade, a member of the Asterid clade that encompasses about 32,000 species including other important crops such as lettuce, sunflower, and the more closely related members of the Apiaceae such as celery, parsley and cilantro[78]. Currently, only two genomes, horseweed (*Conyza canadensis*)[79] and artichoke[80], are available for species in the Euasterid II lineage whereas the Euasterid I lineage has several sequenced genomes, including tomato[81], potato[82], pepper[47], coffee[83] and oil sesame[84]. Here we carried out the first comparative analysis including a member of the Euasterid II clade to establish, at the genome-wide level, the phylogenetic relationships of the carrot genome with its relatives, estimate the temporal divergence of the Euasterid clade, study the mode of evolution of the carrot genome and identify genes that may have contributed to carrot adaptation and its biological characteristics.

4.1 Orthologous gene clusters and comparative analysis

Identification of orthologous genes represents the first step to study the evolution of a genome and to further characterize lineage specific gene families. Gene clusters were identified using OrthoMCL v2.0.2 (ref. 85). The peptide sequences used were from thirteen species,

Nature Genetics: doi:10.1038/ng.3565

including *Arabidopsis thaliana*, *Arabidopsis lyrata, Actinidia chinensis*, *Brassica rapa*, *Carica papaya*, *Coffea canephora, Daucus carota*, *Vitis vinifera*, *Prunus persica*, *Solanum lycopersicum*, *Solanum tuberosum*, *Oryza sativa* and *Sorgum bicolor*. The input datasets from carrot and the other sequenced genomes are listed in (**Supplementary Table 19)**. *Lactuca sativa* (lettuce) only had Expressed Sequence Tags (ESTs) available[86]. To obtain translated protein sequences, for lettuce, assembled EST/unigenes were converted to protein sequences starting at the six-frame translation. The following six steps were used to filter out low-quality sequences:

1) Remove the genes which have internal stop codons in the Coding DNA Sequence (CDS);
2) Retain the genes which have the longest alternative splicing sites;
3) Remove the genes which align (blastn[11]) against a database of repetitive DNA elements (Repbase) (E-value <1e-5, identity >50% and coverage >80%);
4) Remove the genes with length ≤150 nt;
5) Remove the chondriosome and plastome genes.
6) For genes with mixed bases, change codons into NNN, and the corresponding amino acid into X.

After filtering, the pairwise sequence similarities between all input protein sequences were calculated using all-by-all blastp[11] with an E-value of 1e-05 and minimum match length of 50%. Evaluation of clustering of genes was performed by OrthoMCL with inflation value (-I) of 1.5.

Given the limited number of genes available for lettuce, OrthoMCL output including lettuce was limited to the identification of single copy genes to establish phylogenetic relationships and divergence time. To avoid bias due to the limited number of genes and sequence information available, lettuce genes were removed from analyses that were carried out to identify lineage-specific gene families in the study of their mode of evolution.

In total 309,314 predicted genes were clustered in 37,811 gene families. Across the thirteen species, the number of genes in families ranged from 32,643 (*B. rapa*) to 8,152 (*L. sativa*) (**Supplementary Table 62)**.

Nature Genetics: doi:10.1038/ng.3565

4.2 Phylogenetic analysis and divergence time estimation

Methods

The peptide sequence from 312 single copy orthologous gene clusters was extracted and used to construct the phylogenetic relationships among the species included in this study and estimate their divergence time. Protein sequences for each cluster were aligned using MUSCLE[87]. Aligned blocks were then converted (back-translation) in coding sequences (CDS). Fourfold degenerate sites (4DTv) were extracted from each alignment and concatenated to one super gene for each species. The sequences were then used to estimate neutral substitution rate per year and divergence time. Then, PhyML[88] was used to construct the phylogenetic tree.

The Bayesian Relaxed Molecular Clock (BRMC) approach was used to estimate the species divergence time using the program MCMCTREE v4.0, which is part of the PAML package[89]. The "Correlated molecular clock" and "JC69" models were used. The MCMC process of PAML MCMC TREE program was run to sample 100,000 times, with sample frequency set to 2, after a burn-in of 10,000 iterations. "fine tune" parameters were set to make acceptance proportions fall in interval (0.15, 0.7). Other parameters were set at default. Two independent runs were performed to check convergence. Published divergence time between sorghum-rice (<55Mya, >35Mya)[90,91,92], tomato-potato divergence (<4Mya,>2Mya)[81], and grape-rice (<130Mya,>240Mya)[35] were used to calibrate the divergence time.

Results

Whole genomes provide information to identify orthologous genes that did not undergo genome/gene duplications and losses, and consequently, to allow accurate reconstruction of phylogenetic relationships. Therefore, despite the fact that smaller sets of plastid or nuclear genome markers have established phylogenetic relationships of plants and angiosperms, phylogenomic analyses using large datasets from the nuclear genome are more powerful in establishing robust phylogenetic relationships[93,94].

In this study, phylogenetic analysis carried out with 312 single copy orthologous genes produced a well-supported tree that confirmed previously established phylogenetic relationships

Nature Genetics: doi:10.1038/ng.3565

based on a smaller set of DNA markers[66], and on plant genomes deposited in databases (http://www.phytozome.net/). Tomato-potato and carrot-lettuce, all members of the Euasterid clade, separated in two well supported groups corresponding to the Euasterid I and Euasterid II clades (**Supplementary Figure 12 panel A**), respectively. Kiwi, a member of the Ericales family, was placed at the base of the Euasterid clade, as expected.

The estimation of the divergence time among members of the Asterid clade indicated that the Asterids diverged from their sister clade, the Rosids, in the early cretaceous period about 113 Mya. Carrot diverged from kiwi about ~101 Mya and from potato and tomato ~90.5 Mya (**Supplementary Figure 12 panel B**). These splits likely represent the diversification of the three major Asterid branches, basal Asterid (Ericales) from the Euasterids clade, and Euasterids I from Euasterids II clade, respectively. These estimates are consistent with the hypothesis that the Asterids clade diverged from the Rosids in the late early cretaceous, and further radiated in its major subgroups Ericales, Lamids (Euasterids I) and Campanulids (Euasterids II) in the early cretaceous period (http://www.mobot.org/MOBOT/research/APweb/welcome.html). Further divergence between carrot and lettuce, both members of the Euasterid II clade, probably occurred ~72 Mya (**Supplementary Figure 12 panel B**).

4.3 Genome synteny and genome duplication

Methods

To study the evolution of the carrot genome we used two methods: 1) classical ks-based (synonymous substitution rate) data analysis between paralogous genes; 2) a paleo-genomic approach to reconstruct the paleopolyploid history of the carrot genome through a comparative analysis with grape, kiwi, tomato and coffee.

Chromosome collinearity within carrot, and between carrot and the tomato, grape and kiwi genomes was carried out with MCscan[95] (http://chibba.agtec.uga.edu/duplication/mcscan). The following parameters were used to detect syntenic blocks: alignment similarity e $\leq 10^{-05}$; average intergenic distance (u) = 40; number of genes required to call synteny, (s) = 5; gap penalty (g) = 2. Carrot gene pairs with blastp[11] hits and differences of gene rank along the chromosomes = 1,

Nature Genetics: doi:10.1038/ng.3565

were classified as tandem duplications and removed from for further synteny analysis. This accounted for 5,745 genes.

All duplicated genes in the syntenic blocks were extracted and used to calculate the ks and 4DTv value according to the HKY[96] model.

To reconstruct the paleopolyploid history of carrot genome evolution we used the method described by Salse[97]. Briefly, grape-grape syntenic blocks were detected using the same parameters described above and classified into seven ancestral chromosomes (**Figure 3c, Supplementary Figures 35, 36)**. This information was then used to detect grape-carrot syntenic blocks descending from the seven ancestral chromosomes and further investigate the expansions of these blocks in carrot compared with four other genomes including grape, kiwi, tomato and coffee.

To estimate the divergence and WGD time point in the carrot and tomato genomes we used a method described by Vanneste *et al.*[98]. Briefly, after removing gene families that were not consistent with the 13 species phylogeny, 3,743 genes were used to calculate divergence or WGD time by the mcmctree (MCMCTREE in paml version 4.4) method. Root time was set between 1.45 and 2.06, which is the divergence time of monocot and dicots[99].

Tree nodes with only two genes from one species were regarded as alpha event duplications while nodes with more than two genes from the same species were regarded as beta event duplications. To estimate the alpha triplication in tomato, nodes with ≤ 3 genes were selected.

Results

A summary of the syntenic blocks detected is reported in **Supplementary Table 63**, **Supplementary Figure 35-37**. Although carrot and tomato share a more recent evolutionary ancestry, the kiwi genome shared the highest number of syntenic blocks (1,860) and orthologous gene pairs (23,518) (**Supplementary Table 63)** with carrot. This suggests that the tomato genome experienced more extensive lineage specific gene loss or genome rearrangements that fragmented syntenic blocks inherited from the carrot-tomato Euasterid genome ancestor.

Nature Genetics: doi:10.1038/ng.3565

On the basis of transversions at fourfold degenerate sites (4DTv) obtained from the 8,239 paralogous gene pairs, we observed three peaks at 4DTv values of 0.2-0.3, 0.31-0.6 and 0.61-1 (**Figure 3**). We named the three peaks Dc-α, Dc-β and Dc-γ, respectively. The oldest peak overlaps with the carrot-Arabidopsis speciation and with the ancestral hexaploidization (γ) event previously detected in Arabidopsis[100], supporting the hypothesis that the γ triplication is shared with all eudicots and was associated with the radiation of this large group of plants[101]. The two recent peaks corresponding to the Dc-α and Dc-β WGD appeared about 43 Mya and 70 Mya, respectively (**Supplementary Table 64)**.

To study the level of expansion and the evolution of the carrot WGDs, carrot paralogs corresponding to the Dc-α, Dc- β and Dc- γ peaks were used to estimate the depth of their corresponding paralogous blocks. Out of 351 duplicated blocks identified in carrot, 341 (97%) blocks harbored paralogous genes detected within the Dc-α, Dc-β and Dc-γ peaks. Analysis of the block depth indicated that the majority (52%) of blocks associated with the Dc-α peak had a depth of two, whereas the majority (39%) of blocks associated with the Dc-β WGD had a depth of three, and blocks associated with the Dc-γ mainly had a depth of two (**Figure 3e**, **Supplementary Table 65**).

Overall the integration of syntenic block depth and paralogous divergence rate indicated that the Dc-α and the Dc-β WGD were likely a duplication and a triplication, respectively.

To confirm our hypothesis we reconstructed the paleopolyploid-history of the carrot genome. Grape-grape syntenic blocks descending from the seven ancestral proto-chromosomes were identified and named according to Jaillon *et al.*[35]: A16=g1-g14-g17; A1=g2-g15-g16; A4=g4-g9-g11; A7=g5-g7-g14; A10=g6-g8-g13; A13=g3-g4-g7; A16=g1-g4-g17; A19=g10-g12-g19 (**Figure 3c, Supplementary Figure 35**). In total 949 grape-carrot blocks were detected (**Supplementary Figure 35**). Blocks overlapping with larger and contiguous carrot-carrot paralogous blocks were merged. After this step, 315 segments were identified and classified into the seven ancestral chromosomes (**Supplementary Figure 36**)(**Figure 3c**). Except for chromosomes 7, 8 and 9 which harbor segments from six out of seven ancestral proto-chromosomes, all of the other chromosomes harbor at least one segment from each of the seven ancestral proto-chromosomes (**Figure 3c**). Based on the distribution of the seven ancestral blocks, we estimated that at least 60 fusions or translocations occurred during the evolutionary

Nature Genetics: doi:10.1038/ng.3565

history of the carrot genome to account for the structure of the nine chromosomes. These regions cover ~298 Mb of the carrot genome and include 26,473 genes.

4DTv values for each of the carrot-carrot pairwise paralogs descending from the seven ancestral chromosomes were used to study their WGD evolutionary history (**Supplementary Figure 14**). For each proto-chromosome, the majority of pairwise comparisons had a 4DTv value ranging from 0.2-0.3 and 0.31-0.6 (**Figure 3b**), supporting our previous hypothesis for two carrot lineage specific WGD, Dc-α and Dc-β. Gene pairs with 4DTv values ranging between 0.61-1.0 and corresponding to the ancestral γ eudicot triplication were evident but largely lost (**Figure 3b**).

Ancestral blocks were then aligned to the tomato, kiwi and coffee genomes (**Supplementary Figure 37)**. One to five and one to six ratios were predominant for all comparisons, confirming our hypothesis that the two lineage specific, whole-genome multiplications, were probably a triplication and a duplication (**Figure 3d-e**).

Of the 36,513 gene families generated from 13 species, 7,854 containing at least one carrot, one Arabidopsis, one tomato and one grass species (rice or sorghum) gene were selected to build phylogenies using PhyML. Phylogenic trees were rooted by treebest (Version 1.9.2).

4.4 Comparative analysis with horseweed (*Conyza canadensis*)

Methods and results

Based on divergence time estimates and WGD duplication analysis, the Dc-β WGD could be shared with members of the Asterales order, which includes other important crops including artichoke, lettuce and sunflower. To address this evolutionary question we carried out a comparative analysis with the horseweed genome[79], a member of the Asterales order. Since gene predictions were not available for the horseweed genome the same gene prediction pipeline used for carrot (Supplementary Note 2.2) was used to predict and annotate horseweed coding sequences. In total 38,199 genes were predicted.

Predicted coding sequences were clustered using OrthoMCL[85] to find single copy gene families across 14 species. A Maximum-likelihood tree was reconstructed based on the fourfold

Nature Genetics: doi:10.1038/ng.3565

degenerate sites from the 963 single copy gene families. The phylogenetic tree indicated that as expected horseweed grouped in the Asterids branch (data not shown).

Since the horseweed genome assembly is highly fragmented (N50 20kb) no obvious syntenic/collinear blocks were detected between carrot and horseweed, or intra-horseweed genomes. Reciprocal best blastn[11] hits of intra-horseweed or horseweed-vs-other species were used to calculate the paralog/ortholog gene divergence. The 4DTv plot (**Supplementary Figure 13)** highlighted the presence of a peak between 0.3-0.4 in horseweed which could correspond to a WGD. The peak preceded the lettuce-horseweed divergence peak (0.2-0.3) suggesting that this WGD may be shared with lettuce. This hypothesis was confirmed by the Ks analysis; a peak located in the 1.2-1.6 range from intra-Horseweed Ks distribution preceded the horseweed-lettuce divergence peak (0.8) (**Supplementary Figure 13**).

4.5 Genome fractionation

Methods

Despite the identification of two lineage specific WGDs, the number of predicted genes in carrot (32,113) is similar to tomato (33,585) which experienced only one WGD after the divergence with carrot, indicating extensive gene fractionation following the Dc-α and Dc-β WDGs. To study the mode of retention of duplicated blocks we carried out a detailed characterization of these genomic regions. Although all genes were duplicated after the WGD, we excluded genes that could not be assigned onto syntenic blocks, as their evolutionary status could not be clearly inferred. To distinguish retained and lost genes we collected all syntenic blocks containing genes associated with the Dc-α, Dc-β and Dc-γ WGD events. Duplicate genes with multiple copies present in syntenic blocks were classified as retained genes, while those without any corresponding duplicate copy in syntenic blocks were classified as lost genes. The depth of duplicated retained genes was then determined.

To estimate the enrichment of specific gene ontology classes, GO[102] categories with more than 50 genes were tested using the program package FUNC[103]. To avoid bias due to gene loss caused by multiple WGDs this analysis was carried out only with the subset of genes associated with the most recent Dc-α WGD event. In the FUNC package, a hypergeometric test was used to

identify GO categories with overrepresentation or underrepresentation of Dc-α WGD retained genes and tandemly duplicated genes. To avoid over representing significant GO categories due to the hierarchical structure of GO annotation, each GO category was tested by subtracting genes belonging to all of its child categories. The test outputs with $p<0.01$ and a false discovery rate <0.01 were considered as significant.

Results

Blocks associated with the Dc-γ peak retained the least number of genes (1,213), had the shortest average length (583 kb) and included the least number of blocks (57). Blocks associated with the Dc-α were the largest on average (809 kb) and the densest, in terms of retained genes (17 paralogs/block). Blocks associated with the Dc-β were the most numerous (196) and retained the largest number of paralogs (4,794). Although the higher depth of the Dc-β event (3x) could bias these estimates, only 644 retained genes had a depth of three, indicating that the rate of retention of genes duplicated at the Dc-β WGD was indeed higher (**Supplementary Table 20-21**)

The ontology analysis of duplicated genes associated with the Dc-α event revealed that these genes made a major contribution to central biological processes such as regulation of transcription (GO:0006355), cell cycle (GO:0007049) and various forms of transport (GO:0006886, GO:0046907, GO:0044765) (**Supplementary Table 22**). They also comprised genes involved in more specific functions such as those with protein domains involved in selective molecule interactions (binding) and protein dimerization activity.

Among genes duplicated in tandem, those involved in disease resistance were under-represented. Considering that resistance (R) genes typically expand through tandem duplications this result suggests that the carrot genome did not experience a large expansion of R genes.

Nature Genetics: doi:10.1038/ng.3565

5. Regulatory and resistance genes – Gene family analysis

5.1 Identification of carrot specific gene families

To identify carrot-specific gene families and genes shared by phylogenetically close groups, genes were classified as: carrot families (orthoMCL clusters containing only carrot genes); Asterid specific families (orthoMCL clusters containing genes from carrot, kiwi, tomato or potato); and Euasterid specific families (orthoMCL clusters containing genes from carrot, tomato or potato). Gene annotations were retrieved to evaluate InterPro and Gene Ontology (GO) of lineage specific gene families.

Overall, 200,030 genes clustered in 11,375 families that were shared among the core angiosperm plants. Additionally, 31,086 genes clustered into 3,765 families that were shared among core dicots (**Supplementary Figure 15**) and 456 genes, clustered in 77 families were specific for the core Asterid clade. A core set of 261 clusters, comprising 1,070 genes, was specific of the Euasterid clade.

In carrot, 26,320 genes clustered in 13,881 families, with 4,470 genes clustered in 1,166 families specific to carrot (**Supplementary Figure 15**). Of these carrot specific genes, 1,864 contain an InterPro domain and were assigned gene ontology category (GO). We also found 6,060 genes that did not cluster with any of the genes from the thirteen species. This number is similar to the number of non-clustered genes from the other genomes included in this analysis (**Supplementary Table 62**). Of these genes, 5,119 contained an InterPro domain and 3,378 genes had assignments to GO categories. In these two subsets of carrot-specific genes, protein domains involved in selective molecule interactions (binding) and signaling pathways (protein kinase) were abundant and perhaps contributed to rapid adaption and diversification of the carrot genome (**Supplementary Tables 23, 24**).

5.2 Annotation of regulatory genes

Considering that a large fraction of expanded and lineage specific gene families identified in the carrot genome were related to proteins involved in regulatory functions (binding), we carried

Nature Genetics: doi:10.1038/ng.3565

out an extensive comparative genome-wide characterization of genes potentially involved in the carrot genome regulatory network. Genes involved in regulatory mechanisms have been recognized as an important source of the diversity and change that underlies the evolution of plants[104].

Methods

Plants regulate gene expression through transcription factors (TFs), transcription regulators (TRs) and chromatin regulators (CRs). To establish a solid genomic framework for carrot genetic and genomic studies and to understand how diverse evolutionary mechanisms contributed to the expansion-contraction of the carrot gene regulatory network we used PlantTFcat (http://plantgrn.noble.org/PlantTFcat/)[105] to annotate all possible candidate TFs, TRs and CRs in the carrot genome. This program couples the identification by InterProScan[106] and comprehensive prediction logic, based on relationships between gene families and conserved domains enabling the classification of plant TFs, TRs, and CRs with high coverage and sensitivity. To simplify the terminology in this paper we will refer to TFs, TRs, and CRs as regulatory genes (RGs).

For comparative analysis, PlantTFcat was used to screen the predicted proteomes from 11 genomes including *D. carota, S. lycopersicum, S. tuberosum, Coffea canephora, A. chinensis, A. thaliana, B. rapa, V. vinifera, Prunus persica, Carica papaya* and *O. sativa*. To increase the stringency of the analysis, we adopted a customized filter to remove possible false predictions. After scanning the proteome sequences with PlantTFcat, only predicted RGs with blastp[11] hits with E-value $\leq 10^{-5}$ to InterPro domains specific for each RG family were retained and used for further analysis.

To study the mode of evolution of specific RG families, predicted RG classes from all the species used in this analysis were grouped with OrthoMCL[85] as described above (see section 3.1). Based on the results from OrthoMCL, the whole genome duplication analysis, and their physical location (rank position of each gene along the assembled carrot genome), the mode of duplication of each RG was classified as: Unclustered (U, RGs that did not cluster with any other RGs); Singletons (S, orthologs that were single copy in carrot); Tandem (T, paralog RGs that had a difference of gene rank ≤2); Proximal (P, paralogs that had a difference of gene rank >2 to

≤20); Dispersed (D, paralogs that had a difference of gene rank >20); and Whole genome duplication (W, paralogs detected by the whole genome duplication analysis as genes retained after the Dc-α, Dc-β or γ WGDs). Genes designated to multiple categories were assigned to a single category using the following hierarchy: T>W>U>S>P>D. For a subset of RG classes, a phylogenetic analysis was carried out to establish if the duplications occurred before or after the speciation as compared to its closest member of each OrthoMCL cluster. Carrot-specific clusters were considered as derived from recent lineage specific duplications. For phylogenetic analysis, multiple alignments with complete protein sequence were conducted using ClustalW[107] with default parameters. Phylogenetic trees were constructed by using the neighbor-joining method, with pairwise deletion, using MEGA version 6 (ref. 45).

Results

Based on the PlantTFcat pipeline, across 103 RG families we predicted 3,267 unique RG candidates in the carrot genome, whereas 3,209 (tomato), 4,435 (potato), 2,256 (coffee), 3,486 (Kiwi), 4,028 (Arabidopsis), 4,743 (*B. rapa*), 2,297 (grape), 2,728 (peach), 2,310 (papaya) and 3,203 (rice) unique RGs were detected in others genomes (**Supplementary Table 25**). The largest proportion of RGs in carrot (2,700 genes) and the other species were classified as TFs, consistent with previous findings in other plant genomes[108] (**Supplementary Table 26**). Considering the similarity among certain domains of some RG families, classification of the same genes into multiple families was expected, and was limited to only 236 genes (7.7%). This involved RG families that share highly similar protein domains or that contain the same conserved domains. For example, 15 RG were annotated into three families, JmjN, JmjC and Jumonji. From previous studies it is known that these three RG families are closely related and share the same protein domain[109]. Further characterization confirmed, for all predicted genes, the presence of the expected conserved protein domain characteristic of these RG families and indicated the mis-assignment for six of these 15 genes in the JmjN family, since they contained additional domains characteristic of the JmjC family. These results indicated that the annotation of a limited number of genes to multiple families does not affect the sensitivity of the analysis and occurs in very closely related RG families which normally require further manual annotation.

Nature Genetics: doi:10.1038/ng.3565

In order to test the sensitivity of our analysis, for 18 RG families we compared our results with selected publications or the TF database of *A. thaliana*, for which detailed phylogenetic analyses has previously been carried out (**Supplementary Table 25**). A score for each family was calculated as the known Arabidopsis genes/number of genes annotated in our study. A score of 1 indicated that all previously identified genes were present in our analysis, whereas a score <1 indicated that not all the previously identified genes were identified. Across the 18 RG families evaluated, 15 families had a score of 1, and three families had a score <1 ranging from 0.92 (B3-domain/RAV) to 0.95 (MADS-MIKC).

Overall these results supported the accuracy of our analysis at a genome-wide level, and provided the opportunity to study the group of candidate RGs in plant genomes and their possible implication on the evolution of the carrot genome.

Across all the families we noticed a large expansion of the BTB-POZ-MATH RG family in rice, which is consistent with the previous finding of the large expansion of this gene family in grass genomes[110]. We also detected the expansion of the MADS-Type 1 genes in papaya.

The total number of candidate RGs in the carrot genome was similar to other genomes (**Supplementary Table 25**). Considering only RG families with at least a total of 100 predicted genes across all the species, 27 RG families in the carrot genome were over-represented relative to all species, six RG families were overrepresented relative to species encompassing the Euasterid clade, and 23 RG families were overrepresented relative to species encompassing the Asterid clade. Interestingly, MADS type 1 and MADS-MIKC subfamilies were among the most under-represented RG families in carrot, relative to all the other species analyzed in this study. Considering this unexpected finding, an additional genome-wide analysis of RGs across all the 11 genomes was carried out using iTAK, another publically available tool to perform genome-wide identification of candidate RGs (http://bioinfo.bti.cornell.edu/cgi-bin/itak/index.cgi). The analysis detected the same set of MADS genes in the carrot genome, making carrot the angiosperm genome with the lowest number of MADS TFs (http://planttfdb.cbi.pku.edu.cn/search_result.php). Considering the importance of this TF family involved in key developmental processes in plants, in particular the development of reproductive organs[111], carrot represents a unique genetic model to further investigate how its genome

compensated for the low diversity of MADS genes during its evolution, and its implications on the development of the reproductive system.

Overall, genomes that did not experience additional WGD after the γ paleo-hexaploidization event shared among eudicots, such as coffee, grape, peach and papaya, harbor a significantly lower number of RGs. In contrast, the *B. rapa* genome which experienced the highest number of WGDs (three)[112] among sequenced plant genomes, harbors the largest set of predicted RGs (5,128). In carrot, about 33% (1,070) of RGs were retained after the Dc-α and Dc-β WGDs. This fraction accounts for the approximate fraction of extra RGs (30%) that the carrot genome contains as compared to those genomes that did not experience a WGD event after the γ paleo-hexaploidization.

In the carrot genome, large scale duplications and dispersed duplications account for the majority of RG duplications. Only two families, Nin-like and TCP, experienced major expansions through tandem duplications (**Supplementary Table 27**). Among all RG families, the average number of RGs retained after WGDs was 36%. These estimates are likely underestimated since genes retained after the ancestral WGDs like the Dc-γ and the Dc-β WGDs were likely not fully detected due to the high genome fractionation and chromosome rearrangements that the carrot genome experienced during its evolution. For 21 RG families, the fraction of RGs retained after WGDs was larger than the average. Many of those gene families like ARF, C3H-WRC/GRF, C2C2-dof have been shown to be involved in complex regulatory mechanisms controlling plant development and contribute to the wide range of phenotypic diversity existing in plants[113,114]. Interestingly, eight of the top 10 RG families retained after WGDs in the carrot genome were found highly correlated with post-WGD retention in plant genomes by Lang *et al.*[108], which associated this correlation to the function of these RG families and the complexity of their role in plant development. Similarly, Omidbakhshfard *et al.*[114] estimated that genes belonging to the C3H-WRC/GRF family in plants had a high post-WGD retention rate (>50%). These findings clearly support previous predictions regarding the importance of WGDs accounting for the accumulation of RGs in plant genomes[108]. In addition these results open interesting evolutionary questions about the role that the structural or functional characteristics of the different RG families have on post-WGD retention and their implication on plant diversity.

Nature Genetics: doi:10.1038/ng.3565

To further study the evolutionary history and the expansion of the carrot genome regulatory network we carried out a detailed analysis of expanded carrot RG families.

zf-GRF family

Several classes of zinc-finger motifs are present in transcription factors (TF) and function as part of the DNA-binding and protein-protein interaction domains that have been implicated in the regulation of important biological processes that are unique to plants, such as flower development, light-regulated morphogenesis and pathogen response[115]. In the carrot genome the most expanded RG family is represented by a set of genes harboring the zf-GRF domain. This set of genes accounts for the expansion of the zinc finger (IPR010666) detected in the set of carrot specific genes (Supplementary Table 37). None of the currently available transcription factor databases include this gene family as a RG member. The PlantTFcat database includes this gene family as a possible member of the RG network since the zinc binding domain (zf-GRF) is found in a variety of DNA-binding proteins. It seems likely that this domain is involved in nucleic acid binding. It is named GRF after three conserved residues in the center of the alignment of the domain. A recent study in *Medicago truncatula* indicated that genes belonging to this TF family may play an important role initiating the symbiotic interaction with *Rhizobium*[116]. To our knowledge, there have been no other studies characterizing this gene family in additional plant genomes. The number of zf-GRF genes ranged from 62 in carrot to one in kiwi and none in papaya. In carrot, of the 62 GRF annotated genes, 27 shared ancestry with other genomes included in our analysis, and four OrthoMCL clusters were carrot specific (**Supplementary Table 66**). Characterization of the structure of this gene indicated that its length ranged from 240 to 550 nt and contained only the zf-GRF domain (**Supplementary Table 67**), (**Supplementary Figure 38**). Reconstruction of the evolutionary history of this RG family indicated that its expansion in the carrot genome occurred after the divergence with tomato, potato and coffee, the species with which carrot shares its most recent evolutionary ancestry (**Supplementary Table 68**) (**Supplementary Figure 38**).

JmjC family

Protein sequences containing Jumonji C (JmjC) domains have been shown to be involved in chromatin remodeling, acting as histone demethylases[117]. This class of CRs is evolutionarily

Nature Genetics: doi:10.1038/ng.3565

conserved in species spanning yeast to human. In the model plant *A. thaliana*, 21 JmjC CRs have been identified, and few have been characterized. They have been shown to be involved in gametophyte development[118], brassinosteroid response[119] and RNA silencing[120]. Recent studies indicated that the JmjC gene *AtJMJ30* (also known as JMJD5) regulates the pace of the circadian clock in Arabidopsis, influencing flowering time, a trait that plays an important role in plant domestication and adaption[121].

Based on phylogenetic relationships and their structure, plant JmjC proteins could be divided into five groups[109], one of which contains proteins with the JmjC domain only, and the other four groups contain the JmjC domain in conjunction with other domains, including the WRC domain (IPR014977), JmjN domains (IPR003349), zf-C5CH2 (IPR004198), FYRN (IPR003888), FYRC (IPR003889).

The number of JmjC genes detected in our study ranged from 42 in carrot to 19 in papaya (**Supplementary Tables 25 and 69**). The majority of JmjC genes (94% of genes) shared ancestry with other species. In carrot, 30 JmjC genes shared ancestry with at least one species included in this study (**Supplementary Table 66**). Among all the OrthoMCL clusters, none were Asterid or Euasterid-specific, indicating no broad lineage specific expansion of JmjC families (**Supplementary Table 70**) Based on the orthologous and phylogenetic relationships, carrot JmjC were assigned to the five subfamilies. One carrot JmjC was classified as JARID group I, 16 as JMJD2 group II, four as group III, 13 as JHDM2 group IV and five as JmjC domain only (**Supplementary Table 69**).

Expanded JmjC subgroups in carrot included carrot-specific JmjC paralogs/homologs to group JMJD2 II (carrot cluster 3, 4 and 18), JmjC proteins homologous to JHDM2 group IV (carrot cluster 6 and 20) (**Supplementary Figure 16**). Carrot JmjC cluster 4 includes the Arabidopsis *RELATIVE OF EARLY FLOWERING 6* (*REF6*) orthologs[121]. Consistent with previous studies, *REF6* shared a close phylogenetic relationship with *EARLY FLOWERING 6* (*ELF6*). These two genes have demonstrated a fundamental role in regulation of Arabidopsis flowering acting as a *FLOWERING LOCUS C* (*FLC*) repressor (*REF6*) or repressing (*ELF6*) the photoperiod pathway[122]. Interestingly, two carrot *REF6* orthologs from the expanded cluster 4, DCAR_016424 and DCAR_026201, were located in duplicated genomic blocks associated with

Nature Genetics: doi:10.1038/ng.3565

the ancestral γ paleo-hexaploidization (**Supplementary Table 69**). In accordance with this observation, the phylogenetic analysis indicated that DCAR_016424 is more closely related to Arabidopsis REF6, and that it diverged from its onholog DCAR_026201 before the divergence from other eudicots (**Supplementary Figure 16**). Considering the support from the evolutionary history of these two genes in the carrot genome and their phylogenetic relationships we hypothesize that after the eudicot paleo-hexaploidization event (γ WGD), all eudicot species included in this study, besides carrot, retained only one copy of the *REF6* paleopolyploid onholog. The carrot genome retained two copies DCAR_016424 and DCAR_02601, with the latest gene further experiencing multiple events of lineage specific duplications. Whether these genes play a similar role regulating flowering time in carrot remains to be determined, but these results will be helpful for future functional analyses to unravel their divergent roles as related to flowering time and perhaps may play an important role during the evolution and the domestication of this species.

Overall, the analysis of the JmjC family indicated that most of the JmjC gene duplications occurred after the speciation with members of the Euasterid I sister clade (tomato, potato and coffee) included in this study (**Supplementary Table 28**).

TCP family

TCP proteins (TCPs) are plant TFs involved in cell growth and proliferation, regulating plant morphology and architecture. They constitute a means through which evolution shapes plant diversity. Functions associated with TCPs include plant branching, gametophyte development, flower development, leaf development, regulation of hormone pathways, mitochondrial biogenesis, seed germination and regulation of the circadian clock[123]. Functional studies across different angiosperms indicated that some TCP families conserved their original functions. For example, studies in sorghum, rice and Arabidopsis indicated that *TB1* TCP homologs, promoting the axillary branches in maize[124], may represent a crucial point for the determination of the fate of axillary meristems and regulation of branching in angiosperms.

The number of candidate TCP TFs detected here ranged from 50 in carrot to 15 in grape making carrot TCPs the largest set of TCP genes detected in this study (**Supplementary Table 25, 71**). On the basis of sequence identities, 88% of the carrot TCP genes had at least one

ortholog in another species and 74% of TCPs had at least one paralog in the carrot genome (**Supplementary Table 66**) indicating that multiple duplication events contributed to the diversification of this TF family in angiosperms. Among all the OrthoMCL clusters, none were Asterid or Euasterid specific, indicating no broad lineage specific expansion of TCP families (**Supplementary Table 72**). Three carrot specific clusters including 18 TCPs and one cluster sharing ancestry with other species contributed to the major expansion of this TF family in carrot.

Based on the homology of the TCP domains, TCP proteins can be divided into two major classes, class I and class II[123]. It is suspected that TCP classes I and II act antagonistically by competing for common targets or partners. Among the 50 carrot TCPs, 38 were classified as class I and 12 were classified as class II, making class I carrot TCPs largely over-represented (**Supplementary Table 71**, **Supplementary Figure 17**). Expanded carrot TCP clusters were all classified as class I TCPs and mainly derived from tandem duplications. Among the expanded clusters, OrthoMCL11 includes orthologs to Arabidopsis *AtTCP11*, a gene that influences the growth of leaves, stems and petioles, and pollen development[125].

The integration of phylogenetic analysis with the mode of gene duplication in the carrot genome indicated that most of the TCP genes in carrot duplicated after the speciation with members of the Euasterid I sister clade (tomato, potato and coffee) included in this study (**Supplementary Table 28**)

GeBP family

The *GLABRA1 ENHANCER BINDING PROTEIN* (*GeBP*) is a novel plant specific TF family that contains non-canonical Leu-Zipper motifs[126]. GeBP TFs are predicted to play an important role in hormonal pathways, in particular, cytokinin regulation. Hormones such as cytokinin regulate diverse aspects of plant growth and development, including the function of meristems, chloroplast development, vascular differentiation, leaf senescence, modulation of sink-source relationships, nutrient acquisition, nodulation, and the response to biotic and abiotic stresses[127].

Nature Genetics: doi:10.1038/ng.3565

To date, except for Arabidopsis, no genome-wide scale studies have been conducted across multiple plant genomes to identify and characterize GeBP TFs. In the carrot genome, a total of 15 genes were predicted to encode the GeBP proteins (**Supplementary Tables 25, 73**). Although the carrot GeBP genes were expanded relative to other genomes included in this study, the largest set of GeBP genes was predicted in the Arabidopsis (23 GeBPs) and *B. rapa* (18 GeBPs) genomes. Surprisingly, grape and coffee genomes harbor only one and two genes of the GeBP TF family, respectively.

Carrot GeBPs shared a conserved orthologous relationship with dicot and monocot GeBPs, except for four GeBPs derived from lineage specific recent duplications, and one cluster including six carrot GeBPs and two kiwi GeBPs (**Supplementary Tables 62, 74**). Based on the phylogenetic analysis, most of the Arabidopsis and *B. rapa* predicted GeBPs clustered in the same clades indicating lineage specific expansion of this TF family in these two species (**Supplementary Figure 18**). In carrot, expanded GeBP subgroups included genes sharing closer phylogenetic relationships with Arabidopsis *GeBP*, *GPL1*, *GPL2* and *GPL3*, the first group of characterized GeBP genes in *Arabidopsis thaliana*. Recent studies demonstrated that these three genes in Arabidopsis inhibit the induction of the type-A ARRs and thus may antagonize the negative feedback regulation of cytokinin signaling[126] influencing cell expansion[128].

Overall the present study indicated that multiple duplication events likely occurred after the speciation with members of the Euasterid I, and these duplications have contributed to the expansion of this gene family in the carrot genome (**Supplementary Table 28**). In addition, this analysis represents a foundation to carry out further functional characterization of these genes and evaluate their effects on carrot morphology.

Two-component signal transduction system

Protein sequences annotated in this RG family harbor the conserved protein domain IPR001789, a domain functioning as a receiver component in the two-component signal transduction system (TCST). The TCST system is an ancient and evolutionary conserved signaling mechanism in prokaryotes and eukaryotes. Functional characterization of this gene family in Arabidopsis indicated that the two-component elements are involved in plant hormone, stress, and light signaling[129]. The system is usually composed of a membrane-localized His

Nature Genetics: doi:10.1038/ng.3565

protein kinase that serves as an input signal, and a response regulator (a RG), that mediates the output and plays a key role in the TCST modulation. In the Arabidopsis genome, 54 genes were previously annotated to belong to the TCST system, of which 42 genes contain the receiver domain acting as regulatory genes[129]. This included members of the AHK (six genes), ethylene receptor family (ETR, EIN) (three genes) and the response regulators (RR) (33 genes). The RR RG family is the largest group of putative response regulators, and based on their structure and phylogenetic relationships these genes are classified into four subfamilies: A-type ARRs that contain only the receiver domain; B-type ARRs with the receiver domain fused to the DNA-binding domain; C-type ARRs; and pseudo-response regulators (APRRs) that contain only the receiver domain which has diverged from the ARR receiver sequence domain. A and B type ARR receivers are involved in cytokinin signaling pathway and APRR genes are involved in the regulation of the circadian rhythms[130]. The role of the C-type ARRs is still unknown.

Using our pipeline, a total of 576 RR RGs were identified across the 11 genomes, with carrot harboring the largest set of 71 candidate type-A RR genes (**Supplementary Table 25**). Further characterization of this RG family indicated that this group of proteins included not only the type-A RRs but all possible genes involved in the TCST system and containing a receiver domain (AHKs, ETRs, EINs, A-B-C type ARRs, APRRs). Cluster analysis indicated that 64% of carrot genes annotated in these subfamilies shared ancestry with other species (**Supplementary Table 66**). Based on their orthologous and phylogenetic relationships, in carrot a total of 14 genes were classified as AHK (10 genes) and EIN-ETR (four genes). A total of 46 genes were classified as type-A RRs (eight genes), type-B RRs (14 genes), type-C RRs (15 genes), and APRRs (nine genes) (**Supplementary Table 75**)**,** (**Supplementary Figure 19**). Three clusters including type-B RGs and APRR RGs were specific of Euasterid and Asterid species, a sign of broad lineage specific divergence of members of these RG subfamilies (**Supplementary Table 76**). With the exception of the ethylene response regulators genes (ETRs and EINs) all the other subfamilies were expanded in carrot relative to other genomes and perhaps suggest a parallel expansion since previous studies in model species indicated that these proteins function in a coordinated interactive manner. Expanded APRR clusters 7 and 16 share orthologous relationships with Arabidopsis PRR7 and PRR5 (**Supplementary Figure 19**), respectively, a pair of genes which are functionally involved in late flowering response[131], a trait that may have

Nature Genetics: doi:10.1038/ng.3565

played an important role in carrot domestication. Multiple rounds of duplications contributed to the expansion of this family in the carrot genome and this expansion occurred after the divergence between members of the Euasterid I and II clades (**Supplementary Table 28**). Type-C RR genes mainly expanded by tandem and proximal duplications, while all other subfamilies largely expanded by segmental duplications. Overall these results indicated that the carrot genome has experienced a large diversification of the TCST system involved in the cytokinin signaling pathway and circadian clock. These results establish a foundation to further determine the functional association between the evolutionary forces that shaped the carrot genome and their consequences on carrot physiology.

B3 domain superfamily

The plant B3 superfamily encompasses four RG families, including the well characterized auxin response factor (ARF) family and the LAV family, and less well characterized families, such as RAV and REM[132]. A structurally common characteristic of these families is the presence of a ~110 amino acid protein domain called the B3 domain, initially identified in maize as a domain with binding activity in the *VIVIPAROUS1* gene (*VP1*)[133] and considered a plant specific protein domain. A number of RGs belonging to the B3 superfamily have been shown to regulate a multitude of biological processes in plants, controlling or influencing both vegetative and reproductive development. Members of the LAV family are known to regulate seed development and storage reserve accumulation. The ARFs are involved in various auxin-mediated physiological processes, including apical dominance, tropic response, lateral root formation, vascular differentiation and shoot elongation. Over-expression and under-expression of *RAV1* (AT1G13260) in Arabidopsis resulted in lateral root retardation and earlier flowering, respectively[134]. Finally, members of the REM family have demonstrated involvement in the vernalization response. Considering the expansion of this superfamily in carrot and their important role in several aspects of plant biology we carried out a detailed genome-wide comparative analysis for this RG superfamily.

Among the members of this superfamily, PlantTFcat identifies the B3-domain, REV and ARF families as three distinct groups of RGs. Further characterization of these three families indicated that a fraction of RAV genes were annotated in the B3-domain family since some RAV genes containing only B3 protein domains are structurally similar to the B3 domain family. For

Nature Genetics: doi:10.1038/ng.3565

this reason we grouped genes annotated to the B3-domain and RAV family into one large family we refer to as B3-domain/RAV (**Supplementary Table 25**)

Across the 11 genomes we annotated 757 genes in the B3-domain/RAV family and 275 genes in the ARF family (**Supplementary Table 25**). Although the carrot B3-domain/RAV group was expanded relative to the average number of RGs in this family, *B. rapa* and potato genomes harbor the largest set of this RG group. The expansion of this RG family has previously been reported for *B. rapa*[135] but not for potato. No major expansion of the ARF family was observed in any of the plant genomes analyzed in this study. Consistent with our previous observations about the impact of WGDs and RG retention, the coffee, grape, peach and papaya genomes harbor the least number of ARFs, likely as a result of no additional WGDs after the eudicot paleo-hexaploidization event. In carrot over 60% of the ARF genes were retained after the two lineage specific WGDs (**Supplementary Table 77**). In total, 82 carrot B3-domain/RAV genes and 27 ARF genes grouped into 25 and seven OrthoMCL sub-groups, respectively, sharing ancestry with at least one eudicot genome (**Supplementary Table 66**). Based on homology and phylogenetic relationships among the B3-domain/RAV RGs, we differentiated 21 RAV genes, seven LAV genes, and 58 REM genes (**Supplementary Table 78**). Twelve orthoMCL clusters including members of all three B3-domain/RAV sub-families, LAV, REV and REM were expanded in the carrot genome (**Supplementary Table 79**). Major expansions included sub-groups 1 and 4, including orthologs of the *VERNALIZATION1* (*VRN1*), a well characterized gene in Arabidopsis that directly regulates flowering time by interaction with the floral repressor *FLOWERING LOCUS* (*FLC*) (**Supplementary Figure 20**).

5.3 R-Gene characterization

Methods

MATRIX-R pipeline[136] was used to automatically retrieve, annotate and classify plant resistance (R) genes. The pipeline uses signature domain information to systematically ascribe proteins to different R gene families depending on the presence or absence of multiple domains in a single protein, which are based on published R gene functional characterization. Processing about 1,500 proteins/minute, this is a powerful tool to discover new putative R genes and to

Nature Genetics: doi:10.1038/ng.3565

perform genomic and evolutionary studies of plant resistance genes with a discovery rate of 100% and correct classification 95% of the time (tested on *A. thaliana* and *S. lycopersicum*). Protein sequences of 91 cloned R genes falling into the four major R classes (http://prgdb.crg.eu/) were used as a starting point of the pipeline. Protein sequences belonging to single R classes were aligned using MUSCLE 3.6 (ref. 87) followed by manual editing. The resulting alignments for each group were used as a base for the creation of an aligned subset of conserved regions and of a set of hidden Markov models (HMMs) using the HMMER v3 (ref. 137). A total of 60 HMMs were built, 15 for the CNL class, 24 for the TNL class, eight for the RLK class and 13 for the RLP class. For each protein, R-FINDER calculates:

- the coil potential, with COILS[138] to detect CC domains
- the putative transmembrane domains with TMHMM
- inferred putative protein localization (http://www.cbs.dtu.dk/services/TMHMM/)
- the matching score with the HMM modules previously created

According to user defined thresholds, all proteins with a significant match with the HMM modules were stored and then assigned to R-classes on the basis of the type of HMM, the presence/absence of coils and/or transmembrane domains.

This workflow was used to screen the predicted proteomes from nine species including *D. carota*, *S. lycopersicum*, *S. tuberosum, C. canephora, C. annum*, *A. chinensis*, *A. thaliana*, *V. vinifera* and *O. sativa*. To understand the pattern of evolution of R genes in carrot, the list of genomes included most of the sequenced genomes from species belonging to the Asterid clade. The set of predicted proteins identified via HMM profiling was further analyzed using InterProScan version 5.0 (ref. 106) to verify the presence of conserved domains and motifs characteristic of R-proteins (NBS; LRR; TIR; KINASE; SERINE/THREONINE). R genes were further classified based on the presence of the different domains. CNL and TNL classes harbor the three principal domains, TIR or CC, NBS and LRR. RLP and RLK classes contain the serine-threonine- kinase-like domain or kinase domain, respectively, in combination with the LRR domain. RPW8-NL and RLK-GNK2 classes include other genes which have been described as conferring resistance through various molecular mechanisms. It is important to mention that while genes containing NBS domains are specialized in resistance to pathogens, the RLP, LRR and RLK classes are also associated with other molecular functions in plants which encode a RLK and a RLP protein, such as the Arabidopsis *CLAVATA1* and *CLAVATA2* genes,

Nature Genetics: doi:10.1038/ng.3565

respectively[139], and regulate both meristem and organ development. To capture the broad collection of carrot genes possibly involved in resistance to biotic stress, these two R gene classes were included in this analysis.

The characterization of the orthologous and evolutionary history of genes containing the NBS domain was carried out as described above (see section 4.1). The distribution of R genes along the nine carrot chromosomes was analyzed to identify R-gene clusters or arrays. A cluster was defined as a group of at least four R genes from any orthologous group and class predicted in a region spanning <200 kb[140]. An array was defined as genes belonging to the same orthologous R gene class that grouped in the same phylogenetic group supported by a bootstrap support value >65.

Results

Based on our R-gene pipeline, we predicted 634 putative R genes in carrot, whereas 736 (tomato), 1,448 (potato), 998 (pepper), 1,040 (coffee), 648 (Arabidopsis) 855 (grape) and 1,204 (rice) R genes were detected in other species (**Supplementary Table 29**). In carrot, 295 R genes may exert their disease resistance function as cytoplasmic proteins through canonical resistance domains, such as the NBS, LRR and TIR domains. In addition, 339 genes were classified as transmembrane receptors, including 242 receptor-like kinases (RLK), and 97 receptor-like proteins. Most of the cytoplasmic R-gene classes were underrepresented in the carrot genome relative to other Euasterid genomes (tomato, potato, pepper and coffee). The CNL class was the only overrepresented class of R genes in carrot. To compare NBS R genes, a total of 214 CNL, 226 TNL, 1,069 NL and 1,185 N R genes from all the species used in this analysis were grouped with OrthoMCL. The number of OrthoMCL clusters ranged from six (TNL) to 22 (CNL) (**Supplementary Table 30**). Expanded CNL and NL subgroups were identified in carrot compared to other species (**Supplementary Tables 31, 32**). In contrast, N and TNL subgroups were largely underrepresented in carrot relative to the members of the Euasterid I clade (tomato, potato and pepper) (**Supplementary Tables 33, 34**) likely reflecting the lineage specific expansion of NBS genes previously reported in Solanaceae species[141].

Analysis of gene duplications of carrot CNL genes indicated that multiple duplication events, and in particular tandem duplications, have contributed to their expansion after the

Nature Genetics: doi:10.1038/ng.3565

speciation with the Euasterid I sister clade (**Supplementary Table 35**). Phylogenetic analysis of all the CNL genes detected here across the nine genomes, highlights the lineage specific expansion of this R gene family (**Supplementary Figure 21**)**.** Although the NL class was not overrepresented in carrot, rounds of duplication contributed to the lineage specific diversification of this class of genes in carrot (**Supplementary Table 35**).

Overall, 98% of predicted R genes were located in pseudomolecules, thus providing the opportunity to characterize the distribution of these gene families (**Supplementary Figure 22**). The remaining genes were located in unanchored scaffolds. Unlike poplar[117] and watermelon[36], all chromosomes contain NBS-encoding genes in carrot. In total 206 (32.5%) R genes were located within 27 genomic clusters and 14 arrays (**Supplementary Table 36**)**,** (**Supplementary Figure 22**). The largest cluster included 12 predicted R genes spanning 554 kb and is located on the long arm of Chr 7 (CL20). Clusters containing CNL R genes on Chr 3 and Chr 7, and clusters containing NL genes on Chr 2 contributed to the expansion of these two R gene classes in carrot. One cluster containing carrot- specific RLK (three genes) and LRR (one gene) genes and spanning a region of only 50 kb co-localized in the same region as the carrot *Mj-1* locus controlling resistance to *Meloydogine javanica*[6] (**Supplementary Figure 22**). Cytoplasmic classes have the highest percent (48%) of genes in clusters or arrays. Thirteen clusters harboring 89 R genes comprised a mix of cytoplasmic and transmembrane genes (**Supplementary Table 37**). It is worth noting that 200 kb is smaller than the genome-wide average distance between recombination points in carrot (388 kb), contributing to a tendency for R gene clusters to be inherited intact. Overall this analysis confirmed the important role that carrot lineage-specific tandem duplications played in the rapid evolution of resistance genes. In addition, R gene clusters may provide a reservoir of genetic diversity from which new plant-pathogen specific interactions can evolve.

Nature Genetics: doi:10.1038/ng.3565

6. A candidate gene controlling carotenoid accumulation

6.1 Introduction

Carotenoids were first discovered in *Daucus carota* (carrot) and named accordingly, however very little is known about the genetic control of carotenoid accumulation in yellow, red, and orange carrot storage roots. In plants, carotenoids play an essential role in light capture, photoprotection, as well as providing precursors to important downstream compounds including norisoprenoids, strigolactone and abscisic acid[142,143,144]. In humans, provitamin A carotenoids, such as beta-carotene, are converted to vitamin A, which is critical for maintaining normal vision, a healthy immune system, and effective cellular communication and differentiation[145,146]. Several genetic loci in carrot have been shown to be associated with carotenoid accumulation including *Y* (which blocks synthesis of all carotenoids), Y_1 and Y_2 (which block the synthesis of carotenes but not xanthophylls), *L* and L_2 (which block the synthesis of lycopene), and *Rp* (which results in reduction of pigmentation)[147,148,149,150,151]. However, genes underlying these genetic loci have not been characterized in carrot, and the only gene proven to alter carotenoid profile, specifically increased levels of alpha-carotene, is a defective carotene hydroxylase CYP97A3, which is a homolog of Arabidopsis *lut5* (ref. 152). Additionally, although homologs of all known carotenoid biosynthetic genes have been identified in carrot, and phytoene synthase transcript quantities are somewhat higher in orange carrot than in yellow or white carrot roots[153,154] none of these genes have been found to be responsible for the large accumulation of carotenoids in the carrot taproot.

A two-gene model, including the *Y* and Y_2 genes, has been proposed to explain the phenotypic differences between white and orange carrots[149,155,156]. In this model $Y\_Y_2\_$ conditions white roots, $yyY_2\_$ yellow, $YY\ y_2y_2$ pale orange, and yyy_2y_2 orange. Previous research has identified several QTL associated with carotenoid accumulation, specifically, the *Y* and Y_2 genes have been mapped to linkage groups 2 (Chr 5) and 5 (Chr 7), respectively[151,157].

A better understanding of carotenoid accumulation in carrots will contribute to improved nutritional content in this crop and may provide novel targets to pursue increased carotenoid

Nature Genetics: doi:10.1038/ng.3565

accumulation in other species. To this end, we evaluated the carrot genome to identify candidate genes with fine-mapping and explore the genetic control of carotenoid accumulation in carrot with transcriptome analysis in two independent mapping populations, both segregating at the *Y* locus. Additionally, we used resequencing data to associate polymorphisms with phenotypes in the genomic region that includes *Y*.

6.2 Materials and Methods

6.2.1 Plant Materials

The F_2 mapping population, 97837, was grown the winter of 2013-2014 at the UC Desert Research and Extension Center. This population was derived from an intercross between the yellow-rooted selection BCVTHT and the white-rooted variety, White Belgian (**Supplementary Figure 39**). We selected 39 white ($Y\_Y_2Y_2$) and 49 yellow ($yyY_2\ Y_2$) carrot roots from the population for genotyping and phenotyping (**Supplementary Figure 40**). An additional 165 samples, from the 97837 population were grown in the UW Madison Walnut street greenhouse and used for fine mapping. Five F_3 populations derived from self-pollination of 97837 plants were grown in the summer of 2014 at the UW Madison Hancock Research Station and an additional four similarly derived populations were grown during the winter of 2014-2015 at the UC Desert Research and Extension Center.

Another mapping (F_4) population, 70796, was grown during the winter of 2010-2011 at the UC Desert Research and Extension Center. 70796 was derived from a cross between B493, a dark orange USDA inbred carrot, and QAL, a wild white-rooted carrot (*D. carota* var. *carota*)(**Supplementary Figure 39**). Pedigree and preliminary data[158] indicated that this population segregates at the *Y* locus for dark orange (yyy_2y_2) and pale orange ($Y\_y_2y_2$), phenotypes and genotypes. A total of 285 B493 × QAL F_4 roots were used for phenotyping and genotyping. In addition, testcrosses were made using the recessive orange inbred, B493.

Nature Genetics: doi:10.1038/ng.3565

6.2.2 HPLC and phenotypic evaluation

Carotenoid content was quantified using lyophilized root tissue for HPLC analysis as described by Simon and Wolff[159] and Simon et al.[160]. Briefly, 0.1g of lyophilized and ground carrot root tissue was soaked in 2ml of hexane at 4°C. After 15 hours, 300µl of the hexane extract was added to 700µl of methanol, eluted through a Rainin Microsorb-MV column and analyzed on a Millipore Waters 712 WISP HPLC system. Synthetic β-carotene (Sigma-Aldrich, St. Louis, MO) was used in each independent run as a reference standard for calibration. Xanthophyll, α- and β-carotene were quantified at 450 nm and phytoene at 287 nm. Total carotenoids were calculated as the sum of all quantified pigments. All concentrations were described in µg g^{-1} dry weight (DW).

6.2.3 Genotypic evaluation and association analysis

Total genomic DNA of individual plants was isolated from four week old lyophilized leaves following the protocol described by Murray and Thompson[1] with modifications by Boiteux et al.[161]. DNA was quantified using Quantus PicoGreen ds DNA Kit (Life Technologies, Grand Island, NY) and normalized to 10ng/µl.

Genotyping was carried out using the Genotyping-by-Sequencing (GBS) technique in population 97837 and with the KASPar chemistry in population 70796. GBS, as described by Elshire et al.[23], was conducted at the Biotechnology Center, UW-Madison (WI, USA) with minimal modification and half-sized reactions. Briefly, DNA samples were digested with ApeKI, barcoded and pooled for sequencing and run on a single Illumina HiSeq 2000 lane, using single end, 100 nt reads and v3 SBS reagents (Illumina, San Diego, CA). The TASSEL-GBS pipeline version 4.3.7 was used to call SNPs as described in Bradbury et al.[24] and Glaubitz et al.[25].

To genotype 70796 population plants, a collection of 4,000 published SNPs[2] was evaluated by KBioscience using KASPar chemistry, which is a competitive allele-specific PCR SNP genotyping system using FRET quencher cassette oligos (http://www. KBioscience.co.uk/reagents/KASP/KASP.html). Approximately 980 out of 4000 SNPs were

Nature Genetics: doi:10.1038/ng.3565

polymorphic. After filtering for 10% missing data based upon marker and genotype, 920 SNPs markers were used for molecular mapping.

Marker-trait associations for both populations were carried out with molecular markers considered as fixed effects in a linear model implemented in the GLM function of TASSEL[24]. The carrot genome assembly v2.0 was used as a reference to identify marker locations. The genome-wide significance threshold was determined by the Bonferroni method[162].

To confirm associations, QTL analysis was carried for population 70796 using the R package qtl[163]. In order to analyze the major QTL, the single QTL analysis and LOD score calculations were done by standard interval mapping and marker regression (10,000 permutations, 0.001 assumed genotyping error rate). A logarithm of odds (LOD) threshold of 3.0 was used to identify QTLs while avoiding false positives. The confidence intervals for each of the QTL were defined as the 1.5 LOD drop off on either side of the peak of the QTL.

6.2.4 Fine-mapping

After the initial identification of the genomic region associated with carotenoid accumulation in both populations, an additional 165 and 130 samples from populations 97837 and 70796, respectively, were used to narrow down the *Y* gene region, to identify potential candidate genes controlling this trait. DNA was extracted from freeze-dried leaves as previously described. A set of 18 primer pairs were designed using Primer3 (ref. 164) targeting specific loci spanning the genomic sequences flanking the most significant markers associated with carotenoid accumulation (**Supplementary Table 80).** Additional samples were evaluated by PCR and Sanger sequencing as described in Iorizzo et al.[15].

We further used the resequencing data to relate polymorphisms and phenotypes in the region associated with the high carotenoid accumulation. To identify the haplotype block associated with pigmented vs. non-pigmented roots, SNPs covering the region associated with high carotenoid accumulation were loaded into TASSEL[24] and manually inspected to identify the start and end of the haplotype block. Sequence from the haplotype block and its flanking sequences were then used to carry the haplotype network analysis using PopArt v1.7 (ref. 165 with the following parameters: minimum spanning network analysis with Epsilon = 0.

Nature Genetics: doi:10.1038/ng.3565

Haploview v4.2 (ref. 166) was used to calculate and visualize linkage disequilibrium in the region associated with high carotenoid accumulation. F_{ST} analysis of the 1,393,431 original filtered SNPs was conducted in a pairwise fashion between each of the 35 resequenced genotypes using VCFTools[70] with default parameters. The top 1% of F_{ST} values were determined and visualized by a custom Perl script. Nucleotide diversity (π) was estimated in TASSEL[24] using the method described by Nei and Lin[167].

6.2.5 RNA-Sequencing

In population 97837, root tissue was collected from plants with yellow (yyY_2Y_2) and white (YYY_2Y_2) genotypes, with two biological replications per genotype, at 80 days after planting (DAP). In population 70796, root tissue was collected from plants with dark orange (yyy_2y_2) and pale orange (YYy_2y_2) genotypes, with three biological replications per genotype, at 100 DAP. Total RNA was extracted from whole root tissue using the TRIzol® Plus RNA Purification Kit (Life Technologies, Carlsbad, CA) in accordance with the manufacturer's protocol. RNA was treated for DNA contamination with the TurboDNA-free kit (Life Technologies, Carlsbad, CA). RNA quantity and integrity was confirmed with an Experion RNA StdSens Analysis kit (Bio-Rad, Hercules, CA). All samples had RQI values above 8.0.

For each biological replicate, a 133 nt insert size paired-end library was prepared at the Biotechnology Center, UW-Madison (WI, USA). Libraries were sequenced on Illumina HiSeq2000 lanes using 2 × 100 nt reads. Reads were filtered using Trimmomatic version 0.32 with adapter trimming and using a sliding window of length ≥50 and quality ≥28, ***i.e.*** "ILLUMINACLIP:adapterfna:2:40:15 LEADING:28 TRAILING:28 SLIDINGWINDOW:10:28 MINLEN:50".

Filtered reads were aligned to the *Daucus carota* v2.0 genome assembly using the program TopHat v2.0.12 (ref. 30). Non-default parameters used were "--mate-inner-dist -67 --mate-std-dev 50 --min-intron-length 20 --max-intron-length 10000 --library-type fr-unstranded --num-threads 14". The aligned read files were processed by Cufflinks v2.2.1 (ref. 61). Reads were assembled into transcripts with "cufflinks" using the carrot annotation v1.0 gene predictions as the reference gtf guide. Samples were combined with "cuffmerge", and then differential

Nature Genetics: doi:10.1038/ng.3565

expression analyzed with "cuffdiff", using non-default parameters of "--multi-read-correct --min-alignment-count=5". Using the abundance estimations, this performs tests for differential expression and regulation between the samples. Normalized counts of the mapped RNA sequences were used to calculate the relative abundances of transcripts expressed as Fragments Per Kilobase of exon per Million fragments mapped (FPKM). When testing for differential expression, biological replicates were included as a term in the mixed model analysis to account for experimental error. Testing for differential expression was done at the level of genes, isoforms, and promoters. Transcriptome sequence polymorphisms of plants with the high pigmented (dark orange and yellow) and low pigmented (pale orange and white) phenotypes were evaluated by manually identifying SNPs and indels in the candidate region to find polymorphisms associated with contrasting genotypes at the *Y* locus.

6.2.6 Weighted gene co-expression network analysis (WGCNA)

Fragments Per Kilobase of exon per Million fragments mapped (FPKM) were calculated using the protocol as described in **Supplementary Note 6.2.5**. A coefficient of variation (CV) filter of 0.7 was applied to the expression values in order to eliminate those genes that are either not variable in expression, not expressed in any genotype, or constitutively expressed across the four genotypes. This CV threshold was determined based on the number of genes to be analyzed (8,345 genes). Expression values were log2-transformed and the WGCNA package[168] in R with signed correlations was used to determine gene co-expression modules with a soft threshold value β of 10 and a treecut value of 0.6. The β and treecut parameters were chosen after assessing the quality of modules detected, and all other parameters used default settings. Custom Perl scripts were used to determine the expression level of genes within each module, and the candidate gene for the *Y* locus, DCAR_032551, was identified in the blue module. Functional annotation of genes within this module was determined by blastp[11] of protein sequences within this module to Arabidopsis TAIR10 (ref. 169) predictions and gene ontology enrichment analysis based on blastp best hits to TAIR10 was determined using AgriGO[170] and PANTHER[171].

Nature Genetics: doi:10.1038/ng.3565

6.2.7 Differentially expressed gene annotation

To understand the genome-wide transcriptome changes associated with the *Y* locus we manually annotated all genes that were simultaneously upregulated or downregulated in both yellow and dark orange samples, relative to the white and pale orange samples. Protein sequences from this subset of differentially expressed genes were extracted and used to find Arabidopsis homologs and predict subcellular localization. Protein sequences were aligned against the Arabidopsis database (https://www.arabidopsis.org/Blast/index.jsp) and the annotations from the most similar Arabidopsis homologs (>50% identity, minimum length 50aa) were recorded and manually curated. Genes that had their function tested *in vivo* or *in vitro* and reported in the literature were annotated, and their functional descriptors were noted. Gene ontology (GO) annotations were also reported. Subcellular localization of each carrot gene was predicted with TargetP[172], using the web-based predictor available at http://www.cbs.dtu.dk/services/TargetP/. The cellular location assignment was based on the predicted presence of any of the N-terminal presequences: chloroplast transit peptide (cTP), mitochondrial targeting peptide (mTP), or secretory pathway signal peptide (SP).

6.3 Results and discussion

6.3.1 Inheritance of carotenoid accumulation

Nine carrots from the F_2 mapping population 97837, including *YY, Yy,* and *yy* plants, were self-pollinated to form F_3 families. Segregation ratios in the F_3 families were not significantly different from expected ratios under the hypothesis of a single dominant gene (*Y*) conditioning lutein accumulation given the homozygous Y_2 Y_2 locus of that population (**Supplementary Table 38**). Similarly six carrots from the F_4 mapping population 70796 heterozygous of the *Y* locus were self-pollinated to form F_5 families. Again no significant deviation from the single-dominant gene model (3:1) was observed (**Supplementary Table 38**). Additionally, three *Yy* pOr carrots were testcrossed to the dOr recessive inbred, B493, and the expected 1:1 ratio was observed (**Supplementary Table 38**)**.** These results agree with previous studies that suggest lack or reduction of taproot pigmentation is controlled by the dominant *Y* locus[149,155,157].

Nature Genetics: doi:10.1038/ng.3565

6.3.2 Root pigment analysis

We evaluated 253 F_1 plants from the 97837 mapping population for lutein concentration using both a visual score (0 = yellow pigment absent and 1 = yellow pigment present) and HPLC (μg/g in dry weight) analysis. Lutein concentration varied from 0-66 μg g^{-1} dry weight (DW). From the HPLC data two distinct clusters were identified, those with below 10μg g^{-1} DW of lutein (white roots) and those with above 10μg g^{-1} DW of lutein (yellow roots) (**Supplementary Figure 40**). Based on these results, plants with greater than 10μg g^{-1} DW of lutein content in their storage roots were categorized as having a yyY_2Y_2 genotype, while those with less than 10μg g^{-1} DW were categorized as $Y\_Y_2Y_2$.

We evaluated 155 F_4 plants from the 70796 mapping population for total carotenoid content (primarily alpha- and beta-carotene) using both a visual score (0 = pale orange, pOr, and 1 = dark orange, dOr) and HPLC (μg/g in dry weight) analysis. Total carotenes varied from 22-1200 μg g^{-1} DW. From the HPLC data two distinct clusters were identified, those with less than 190 μg g^{-1} DW of total carotene (pOr roots) and those with above 190 μg g^{-1} DW of total carotene (dOr roots) (**Supplementary Figure 40**) (**Supplementary Figure 41**). Based on these results, plants with greater than 190μg g^{-1} DW total carotenoid content in their storage roots were categorized as yyy_2y_2, while those with less than 190μg g^{-1} DW were categorized as $Y\_y_2y_2$ genotype.

6.3.3 SNP identification

After evaluating samples with the v4.0 TASSEL GBS pipeline, population 97837 had 85,178 SNPs. After filtering for 10% missing data for marker and genotype and 10% minor allele frequency, 24,507 high quality SNPs were called in 70 plants (39 yellow and 31 white). In population 70796, approximately 980 out of the 4000 KASPar SNPs were polymorphic. After filtering for 10% missing data, 920 SNPs markers were used for molecular mapping in 155 samples. The distribution of markers across the nine chromosomes can be found in **Supplementary Tables 81, 82**. In population 97837, Chr 1 had the largest number of markers, 4,258, while the smallest numbers of markers were found on Chr 9, 930. Marker density across the chromosomes ranged from one SNP every 12,064 nt to 36,002 nt. In population 70796, Chr 5 had the densest marker distribution, 139, while Chr 2 was the least dense, 72. Additionally Chr 3,

Nature Genetics: doi:10.1038/ng.3565

Chr 4 and Chr 7 had no polymorphic markers, and all alleles were inherited from the cultivated B493 parent. Of the remaining six chromosomes marker density ranged from one SNP every 211,170 nt to 606,823 nt.

6.3.4 Molecular mapping of carotenoid accumulation

To identify the possible locus underlying the *Y* locus in the 97837 population we used the HPLC data (lutein content) along with the genotypes from our 24,507 GBS SNPs to identify marker-trait associations. Genome-wide tests of significant association were carried out using a standard GLM analysis. Inspection of the Q-Q plot confirmed no inflation in p-values. A region of high significance was found on Chr 5 (**Supplementary Figure 23).** Within this region two recombinants were found to flank a region of 114,778 nt. The flanking markers of this region were S5_24556774 and S5_24671552. A similar analysis was conducted to analyze total carotenoid accumulation in 70796. The HPLC data, total carotenoids, was used in conjunction with 920 KASPar markers. Again a significant region on Chr 5 was identified, with the markers K0536 and K0165 flanking the recombinants within this region of approximately 6 M nt (**Supplementary Figure 23**). Presuming the phenotypic variation in the two populations to be controlled by the same locus (*Y*) we used recombinants from both populations for fine-mapping.

In addition to GLM analysis, for population 70796 we carried out a standard QTL mapping based on recombination frequencies. The results confirmed the detection of a single QTL in chromosome 5 (position 39.8 cM) for total carotenoid, with high LOD value (21.2)(**Supplementary Table 83**) and its nearest marker K0536. These QTLs overlap with the region identified by GLM analysis (**Supplementary Figure 24**).

6.3.5 Fine-mapping, expression, and candidate gene identification

Fine mapping in population 97837 resulted in the identification of eight linkage blocks, spanning 115 kb on Chr 5 between markers S5_24556774 and 173.9, associated with high lutein accumulation **(Figure 4, Supplementary Figure 25**). Samples with linkage blocks between markers S5_24556774 and 173.9 harboring the “B” and “H” alleles had low lutein levels and were classified as White, W, whereas samples associated with the “A” allele had high lutein

Nature Genetics: doi:10.1038/ng.3565

content and were classified as Yellow, Y (**Supplementary Figure 25)**. These results are consistent with the hypothesis that high lutein controlled by the *Y* locus, which is a recessive trait[157]. Similarly, in population 70796, samples associated with linkage block "A" between markers 173.12.1 and 173.12.8 and spanning a region of 91 kb had high total carotenoid content (dOr samples) (**Supplementary Figure 25)**. The results of fine mapping confirmed the co-localization of the region of interest in populations 97837 and 70796, and identified an overlapping linkage block spanning 75 kb (**Figure 4**, **Supplementary Figure 25)** that explained all of the phenotypic variation (W vs. Y and pOr vs. dOr phenotypes) across the two populations, indicating that this region the carrot genome harbors the *Y* gene controlling carotenoid accumulation in carrot root. This region is included within the previously mapped QTL region associated with the *Y* trait[157] (**Figure 4**).

In total, eight genes were predicted in the 75 kb region of overlap (**Supplementary Table 39**). Five of these genes shared similarity with previously characterized genes and all genes contained a characterized conserved protein domain. Interestingly, none of the predicted genes in the 75 kb region were annotated as, or shared similarity with, known biosynthetic or regulatory genes involved in the isoprenoid pathway.

6.3.6 Transcriptome comparison

Comparative transcriptome data were used to evaluate sequence polymorphism comparing high (dark orange and yellow) and low (pale orange and white) pigmented phenotypes by manually identifying SNPs and indels in the candidate region to find polymorphisms associated with contrasting genotypes at the *Y* locus. Analysis of the eight predicted candidate gene sequences from Y, W, dOr and pOr plants identified synonymous SNPs in three genes in this region associated with low and high pigment accumulation, and an insertion of 212 nt in DCAR_032551 that was only present in plants with the recessive *yy* genotype (dOr and Y) (**Supplementary Table 39**). Presuming that the wild type allele represents the functional allele, the 212 nt insert occurs in the second exon of this gene. PCR amplification confirmed the insertion in the mapping populations (**Supplementary Figure 42)**.

Nature Genetics: doi:10.1038/ng.3565

Three genes were differentially expressed in the pOr vs. dOr transcriptome comparison in the 70796 population, and only one of these three genes, DCAR_032551 was differentially expressed between W and Y roots in the 97837 population. The candidate gene, DCAR_032551 was upregulated in the more high-pigmented roots of both populations, making this gene the most likely candidate controlling the *Y* trait. The differential expression for DCAR_032551 was also observed at the isoform level where both larger and smaller isoforms are produced as compared to wild-type, suggesting a possible change in the structure of the gene as a potential mechanism for differential expression. DCAR_032551 is a member of a plant-specific family of proteins with unknown function. The functionality of transcripts remains to be determined in future research.

7. Flavonoid and isoprenoid pathways

Secondary metabolites such as flavonoids and isoprenoids have played an important role throughout the history of carrot domestication. To establish a solid genomic framework and further study the regulatory mechanisms leading to the accumulation and differentiation of these metabolites we carried out a detailed annotation of candidate genes involved in the flavonoid and isoprenoid pathway.

7.1 Methods

A multiple step approach was used to identify and annotate genes involved in the flavonoid and isoprenoid pathways:

1) Peptide sequences of all carrot predicted genes were aligned against genes annotated in flavonoid and isoprenoid pathway in the KEGG database. Blastp[11] was carried out using default parameters and sequences with less than 50% identity, minimum length 50aa were excluded. Sequences with best blastp similarity to genes annotated in the flavonoid and isoprenoid pathways were retained and used for the next analysis;
2) Peptide sequences from other genomes sharing orthologous relationships with retained carrot genes from step1 were extracted from the prior genome evolution analysis (see **Supplementary Note 4**). Genes annotated in these two pathways from Arabidopsis and

Nature Genetics: doi:10.1038/ng.3565

tomato genomes were manually verified. All peptide sequences retained from step1 and step2 were then used to establish phylogenetic relationships and ensure supported clustering of carrot genes with known genes involved in the flavonoid and isoprenoid pathways. Multiple sequence alignments were generated with the ClustalW program[107]. Phylogenetic analyses were carried out using MEGA version 6 (ref. 45).

7.2 Flavonoid pathways

Flavonoids are extensively distributed in the plant kingdom and exhibit a variety of biological activities not only in plants, which produce these compounds, but also in animals, which take visual cues from flavonoids in fruits and leaves, and as a result of the intake of flavonoids in their diets[173].

In carrot, the flavonoid pathway leads to the biosynthesis of flavones and anthocyanin. Several studies have indicated that purple carrots mainly accumulate cyanidin derivatives[174], unlike many other plants that accumulate derivatives of several anthocyanidins. The accumulation of these pigments has played an important role in carrot domestication, since purple carrots were among the first documented colors of domesticated carrot recorded in Central Asia, Asia Minor, then Western Europe and finally in England between the 11th and 15th centuries[175]. To date seven genes involved in the flavonoid and anthocyanin biosynthetic pathway have been identified in carrot[28,176]. Our analysis identified 97 genes involved in the biosynthesis of flavonoids and anthocyanins (**Supplementary Table 84**). In contrast to grape and Arabidopsis, the DH1 carrot genome lacks the *anthocyanidin reductase* (ANR) gene. The ANR enzyme catalyzes the first committed step of the proanthocyanindin (PA) pathway, and perhaps this partially explains the low diversity of flavonoid derivatives detected in carrot.

7.3 Isoprenoid pathways: MEP and Carotenoid biosynthesis

The isoprenoid or terpenoid pathway is one of the most important and well-studied biosynthetic pathways in plants. It involves cross-talk between the cytosolic mevalonate (MVA) and plastid 2-C-methyl-D-erythritol 4-phosphate (MEP) pathways, to give rise to isopentenyl-diphosphate (IPP), the C5 building block required for the synthesis of a diverse group of natural

Nature Genetics: doi:10.1038/ng.3565

products that perform numerous biochemical functions in plants. A main branch of the isoprenoid pathway leads to the synthesis and accumulation of carotenoids, C40 terpenoid compounds formed by the condensation of eight isoprene units, within plastids. This pathway in carrot plays a major role not only for its importance in photosynthesis, plant growth, development, and response to the environment but also because of the role that volatile terpenoids and norisoprenoids play as odor and flavor cues that attract or repel animals, as well as carotenoids that are the source of numerous phytonutrients and dietary vitamin A.

Terpenes constitute a large class of compounds that serve multiple roles in plants including hormone biosynthesis, stress response and reproduction by attraction or repulsion of herbivores, pollinators and seed disseminators[177]. In addition to these important roles in plant physiology and ecology, volatile terpenoids are responsible for aroma and flavors that have a beneficial impact on humans as health promoting compounds[178]. In carrot, over 90 volatile compounds have been identified, with mono- and sesquiterpenoids by far the most abundant[179,180]. Genetic variation, tissue specificity and environmental factors play an important role on the diversity of terpenoid constitution in carrot[181,182,183]. Studies have demonstrated that the level and type of terpinolene, typically the most abundant monoterpene, was associated with undesirably harsh fresh carrot flavor[182] and an oxygenated form of terpinolene, linden ether had the highest relative flavor impact in cooked carrot[180]. Despite their importance in carrot quality, to date, only two *TPS* genes have been identified and characterized in carrot[184]. The *TPS* gene family in other plant genomes examined included anywhere from 19-113 members[185,93].

To date, 24 genes involved in the carotenoid biosynthetic pathway have been identified in carrot[186] and none of these genes are involved in the biosynthesis of carotenoid precursors in the MEP or MVA pathways. Our analysis of the carrot genome identified 24 genes involved in the biosynthesis of carotenoid precursors and 44 genes involved in the isoprenoid biosynthetic pathway (**Supplementary Tables 45, 85**), (**Supplementary Figure 43 Panel A**). The majority (63 of 68) of the carrot genes annotated in the MEP or carotenoid pathways shared ancestry with genes from at least one other plant genome evaluated in this study, reflecting conservation of this pathway across angiosperms[187]. Phylogenetic analysis of the four *DXS* genes, indicated that carrot retained at least one copy of each of the three *DXS* clades (**Supplementary Figure 44**), a gene family that specialized in synthetize isoprenoid/carotenoid precursors[188]. Expanded gene

Nature Genetics: doi:10.1038/ng.3565

families in carrot relative to other plant genomes included *GGPS*, carotenoid oxygenases (*CCDs* and *NCEDs*) and abscisic acid 8'-hydroxylase 4-like (*CYP707a-b-c*). *GGPS* is involved in the synthesis of carotenoid precursors, while *CCDs*, *NCEDs* and *CYP707s* are involved in the degradation of carotenoids into secondary products such as abscisic acid, norisoprenoids and strigolactone. The expansion of gene families that act upstream or downstream from the carotenoid pathway suggests that the carrot genome has evolved efficient metabolic machinery to initiate and further process the diversity of metabolites synthesized in the carotenoid pathway including ABA precursors. According to the reconstruction of the evolutionary history of duplicated genes in the carrot genome, over 50% of the annotated carotenoid oxygenase genes were retained after the two recent WGDs, Dc-α and Dc-β.

Expanded *CCD* genes in carrot include genes orthologous to Arabidopsis and tomato *CCD1*, genes involved in carotenoid cleavage and norisoprenoid flavor volatile production (**Supplementary Figure 43 Panel B)**. To date, norisoprenoid compounds that have been detected in carrot include farnesylacetone, geranylacetone, α-ionone and β-ionone[189]. Despite the diversity of norisoprenoids identified in carrot, only one *CCD1* gene that is associated with the biosynthesis of α- and β-ionone has been described[189]. Here we identified four *CCD1* genes including one that was previously described (DCAR_003216) and three new ones (**Supplementary Table 45**). A pair of *CCD1* genes (DCAR_022390 and DCAR_003216) was retained after the Dc-α WGD event and local tandem duplications contributed to further expansion of this set of genes on Chr 6. Among *CCD1* genes, DCAR_022386 was not expressed in DH1, indicating that this gene may represent a pseudogene in the DH genome.

7.4 Identification of terpene synthase genes

7.4.1 Methods

Carrot peptide sequences annotated as InterProScan ID IPR001906 and IPR005630 and containing the N terminal domains, PF011397 and PF03936, were extracted and manually analyzed for the presence of the conserved domains of terpene synthase DDXXD and DXDD[190]. Putative full-length TPS- predicted proteins identified in carrot (**Supplementary Table 46**)

Nature Genetics: doi:10.1038/ng.3565

along with representative known TPSs[191,185] from six other species, including two from *Physcomitrella patens*, 14 from *Selaginella moellendorffii*, 21 from *Sorghum bicolor*, 31 from *O. sativa*, 29 from *S. lycopersicum* and 33 from *A. thaliana,* were used to perform phylogenetic analysis with MEGA version 6 (ref. 45).

A multiple sequence alignment was generated with the ClustalW program[107]. The alignment was then truncated to ensure that sites were homologous. To create a phylogeny, we first tested which amino acid substitution model provided the maximum likelihood tree with the best AICc value[192] (using Akaike's information criterion, corrected for samples size) and further tested whether the gamma distribution estimation and/or proportion of invariable sites estimation improved the AICc value. The amino acid substitution models that were tested were: WAG, mtREV, Dayhoff, JTT, VT, Blosum62, and CpREV. The tree with the highest AICc value was obtained with the JTT+F model with estimation of the gamma (G) distribution. The phylogenetic tree was then rooted at the split between type I (*TPS-c*, *-e*, *-f* and *-h*) and type III (*TPS-a*, *-b* and *-g*) subfamilies. Clades were labeled according to their *TPS* sub family category and nodes were colored by species (**Supplementary Figure 45**).

7.4.2 Results and Discussion

Our analysis of the *TPS* gene family indicated that the *D. carota* genome has at least 30 *TPS* genes (**Supplementary Table 46**), a similar number to that found in tomato (29 *TPS*) and Arabidopsis (33 *TPS*). Among these 30 *TPS* genes, 22 contain the DDXXD domain, four contain the DXDD domain and none contained both motifs, similar to other studies in other angiosperm genomes. Neither domain was found in the remaining four TPS proteins, which may represent pseudogenes. Phylogenetic analysis of the carrot *TPS* genes indicated that two are in the *TPS-a* clade, 16 are in *TPS-b* clade, three in *TPS-c* clade, three in *TPS-e/f* clade and six in *TPS-g* clade (**Supplementary Figure 45)**. The subfamily *TPS-h*, as reported in previous studies[185] is unique to the lycopod *Selaginella*, a species basal to vascular plants, and was not found in carrot. Relative to tomato, a member of the Euasterid I sister clade, *TPSs* in the carrot genome are over-represented by subfamilies *-b* and *-g*, which form monoterpenes. Proteins responsible for the formation of primary metabolites (e.g. giberellins via ent-kaurene; subfamilies *-c* and *-e*) are represented in similar numbers to other species studies. Most carrot *TPS* genes (24) were found

Nature Genetics: doi:10.1038/ng.3565

to be expressed and several appear to be tissue-specific in expression (**Supplementary Table 46**).

TPS genes were presumably first introduced to the plant genomes in bryophyte mosses, such as *Phycomitrella*. Large lineage specific expansion of *TPS* genes has been observed in grape[190] and *Eucalyptus*[93], perhaps as a result of plant adaptation to certain environmental conditions. Over 50% (16 of 24) of the carrot *TPS* genes are organized in tandem arrays. Phylogenetic analysis clearly demonstrated that lineage-specific duplications contributed to the diversification of *TPS* genes in the carrot genome. Only three *TPS* genes resulted from the carrot Dc-α WGD event. Tandem duplications contributed to the expansion of the *TPS-b* and *-g* subfamilies, perhaps contributing to variation in plant development and adaption to biotic and abiotic stress, or selective pressure during domestication.

Nature Genetics: doi:10.1038/ng.3565

8. References

1. Murray, M.G. & Thompson, W.F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8**, 4321–4326 (1980).
2. Iorizzo, M. *et al.* Genetic structure and domestication of carrot (*Daucus carota* subsp. *sativus*) (Apiaceae). *Amer. J. Bot.* **100**, 930–938 (2013).
3. Arumuganathan, K. & Earle, E.D. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**, 208–218 (1991).
4. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* **1**, 18 (2012).
5. Cavagnaro, P.F. *et al.* A gene-derived SNP-based high resolution linkage map of carrot including the location of QTL conditioning root and leaf anthocyanin pigmentation. *BMC Genomics* **15**, 1118 (2014).
6. Simon, P.W., Matthews, W.C. & Roberts, P. Evidence for simply inherited dominant resistance to *Meloidogyne javanica* in carrot. *Theor. Appl. Genet.* **100**, 735–742 (2000).
7. Parsons, J., Matthews, W., Iorizzo, M., Roberts P. & Simon, P.W. QTL for *Meloidogyne incognita* nematode resistance in carrot. *Mol. Breeding* **35**, 114 (2014).
8. Voorrips, R.E. MapChart: Software for the graphical presentation of linkage maps and QTLs. *J. Hered.* **93**, 77–78 (1994).
9. Van Ooijen, J.W. JoinMap 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma BV, Wageningen, Netherlands (2006).
10. Iovene, M. *et al*. Comparative FISH mapping of *Daucus* species (Apiaceae family). *Chromosom. Res*. **19**, 493–506 (2011).
11. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
12. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
13. Stein L.D. et al. The Generic Genome Browser: A building block for a model organism system database. *Genome Research* **12,**1599–1610 (2002).
14. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754–1760 (2009).
15. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
16. Iorizzo, M. *et al. De novo* assembly of the carrot mitochondrial genome using next generation sequencing of whole genomic DNA provides first evidence of DNA transfer into an angiosperm plastid genome. *BMC Plant Biol.* **12**, 61 (2012).
17. Clement, N. *et al.* The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics* **26**, 38 (2009).
18. Chevreux *et al.* Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Research* **14**, 1147 (2004).

Nature Genetics: doi:10.1038/ng.3565

19. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

20. Bonfield, J.K. & Whitwham, A. Gap5—editing the billion fragment sequence assembly. *Bioinformatics* **26**, 1699-1703 (2010).

21. Robison, M.M. & Wolyn, D.J. A mitochondrial plasmid and plasmid-like RNA and DNA polymerases encoded within the mitochondrial genome of carrot (*Daucus carota* L.). *Curr. Genet.* **47**, 57–66 (2005).

22. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* **6**, e17288 (2011).

23. Elshire, R.J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**, e19379 (2011).

24. Bradbury, P.J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).

25. Glaubitz, J.C. *et al.* TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* **9**, e90346 (2014).

26. Dong, F. *et al.* Development and applications of a set of chromosome-specific cytogenetic DNA markers in potato. *Theor. Appl. Genet*. **101**, 1001–1007 (2000).

27. Iovene, M., Grzebelus, E., Carputo, D., Jiang, J. & Simon, P.W. Major cytogenetic landmarks and karyotype analysis in *Daucus carota* and other Apiaceae. *Amer. J. Bot* **95**, 793–804 (2008).

28. Iorizzo, M. *et al. De novo* assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. *BMC Genomics* **12**, 389 (2011).

29. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644–652 (2011).

30. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

31. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).

32. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573-580 (1999).

33. Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).

34. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265-W268 (2007).

35. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).

36. Garcia-Mas, J. *et al.* The genome of melon (*Cucumis melo* L.). *Proc. Natl. Acad. Sci. USA* **109,** 11872–11877 (2012).

37. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005)

38. Grzebelus, D., Yau, Y.-Y. & Simon, P.W. *Master*: A novel family of *PIF*/*Harbinger*-like transposable elements identified in carrot (*Daucus carota* L.). *Mol. Genet. Genomics* **275**, 450–459 (2006).

39. Macko-Podgorni, A., Nowicka, A., Grzebelus, E., Simon, P.W. & Grzebelus, D. *DcSto*: carrot *Stowaway*-like elements are abundant, diverse, and polymorphic. *Genetica* **141**, 255–267 (2013).

40. Gambin, T. *et al.* TIRfinder: a web tool for mining class II transposons carrying terminal inverted repeats. *Evol. Bioinform. Online* **9**, 17–27 (2013).

41. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).

42. Darzentas, N. Circoletto: visualizing sequence similarity with Circos. *Bioinformatics* **26**, 2620–2621 (2010).

43. Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).

44. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).

45. Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).

46. Tamura, K. *et al.* Estimating divergence times in large molecular phylogenies. *Proc. Natl. Acad. Sci. USA* **109**, 19333–19338 (2012).

47. Qin, C. *et al.* Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proc. Natl. Acad. Sci. USA* **111**, 5135–5140 (2014).

48. Novák, P., Neumann, P. & Macas, J. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* **11**, 378 (2010).

49. Arbizu, C., Ruess, H., Senalik, D., Simon, P.W. & Spooner, D.M. Phylogenomics of the carrot genus (*Daucus*, Apiaceae). *Amer. J. Bot.* **101**, 1666–1685 (2014).

50. Cheng, Z., Presting, G.G., Buell, C.R., Wing, R.A. & Jiang, J. High-resolution pachytene chromosome mapping of bacterial artificial chromosomes anchored by genetic markers reveals the centromere location and the distribution of genetic recombination along chromosome 10 of rice. *Genetics* **157**, 1749–1757 (2001).

51. Arbizu, C., Reitsma, K.R., Simon P.W. & Spooner, D.M. Morphometrics of *Daucus* (Apiaceae): a counterpart to a phylogenomic study. *Amer. J. Bot.* **101**, 2005–2016 (2014).

52. Spalik, K. *et al*. Amphitropic amphiantarctic disjunctions in Apiaceae subfamily Apioideae. *J. Biogeogr*. **37**, 1977–1994 (2010).

53. Lee, H.-R. *et al.* Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in *Oryza* species. *Proc. Natl. Acad. Sci. USA* **102**,11793–11798 (2005).

54. Bao, W. *et al.* Diversity of centromeric repeats in two closely related wild rice species, *Oryza officinalis* and *Oryza rhizomatis*. *Mol. Genet. Genomics* **275**, 421–430 (2006).

55. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).

56. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).

57. Majoros, W.H., Pertea, M. & Salzberg, S.L. TigrScan and GlimmerHMM: two open-source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).

58. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).

59. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).

60. Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M. & Buell, C.R. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genomics* **7**, 327 (2006).

61. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).

62. Elsik, C.G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).

63. Zdobnov, E.M. & Apweiler, R. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).

64. Nawrocki, E.P. & Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).

65. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).

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

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

68. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

69. Van der Auwera, G.A. *et al.* From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **11**, 11.10.1–11.10.33 (2013).

70. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

71. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

72. Lischer, H.E. & Excoffier, L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298–299 (2012).

73. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).

74. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).

75. Earl, D.A. & vonHoldt, B.M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genet. Resour*. **4**, 359–361 (2012).

76. Rosenberg, N.A. DISTRUCT: a program for the graphical display of population structure. *Mol. Ecol. Notes* **4**, 137–138 (2004).

77. Felsenstein, J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**,164–166 (1989).

78. Bremer, B. in *Asterids. The Timetree of Life*. (Eds. Hedges, S.B. & S. Kumar, S.) 177–187 (Oxford University Press, New York, NY, 2009).

79. Peng, Y. *et al.* De novo genome assembly of the economically important weed horseweed using integrated data from multiple sequencing platforms. *Plant Physiol*. **166**, 1241-1254 (2014).

80. Scaglione, D. *et al.* The genome sequence of the outbreeding globe artichoke constructed de novo incorporating a phase-aware low-pass sequencing strategy of F1 progeny. *Sci Rep.* **6**, 19427 (2016).

81. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012).

82. Xu, X. *et al.* Genome sequence and analysis of the tuber crop potato. *Nature* **475,** 189–195 (2011).

83. Denoeud, F. *et al.* The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345**, 1181–1184 (2014).

84. Wang, L. *et al.* Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome Biol.* **15**, R39 (2014).

85. Li, L., Stoeckert C.J.-Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).

86. Truco, M.J. *et al.* An ultra-high-density, transcript-based, genetic map of lettuce. *G3* **3**, 617–631 (2013).

87. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

88. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

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

90. Zhang, G. *et al.* Genome sequence of foxtail millet (*Setaria Italica*) provides insights into grass evolution and biofuel potential. *Nature Biotechnol.* **30**, 549–554 (2012).

91. Paterson, A.H, Bowers, J.E. & Chapman, B.A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. USA* **101**, 9903–9908 (2004).

92. Paterson, A.H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).

93. Myburg, A.A. *et al.*The genome of *Eucalyptus grandis*. *Nature* **510**, 356–362 (2014).

94. Shulaev, V. *et al.* The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* **43,** 109–116 (2011).

95. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).

96. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).

97. Salse, J. *et al*. Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc. Natl. Acad. Sci. USA* **106**, 14908–14913 (2009).

98. Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous − Paleogene boundary, *Genome Res*. **24**, 1334–1347 (2014).

99. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92–100 (2002).

100. Ermolaeva, M.D., Wu, M., Eisen, J.A. & Salzberg, S.L. The age of the *Arabidopsis thaliana* genome duplication. *Plant Mol. Biol.* **51**, 859–866 (2003).

101. Jiao, Y. *et al*. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**, R3 (2012).

102. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

103. Prüfer, K. *et al.* FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* **8**, 41 (2007).

104. Riechmann, J.L. *et al.* Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science* **290**, 2105–2110 (2000).

105. Dai, X., Sinharoy, S., Udvardi, M. & Zhao, P.-X. PlantTFcat: an online plant transcription factor and transcriptional regulator categorization and analysis tool. *BMC Bioinformatics* **14**, 321 (2013).

106. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).

107. Larkin, M.A. *et al.* Clustal W and Clustal X Version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).

108. Lang, D. *et al*. Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol. Evol.* **2**, 488–503 (2010).

109. Lu, F. *et al.* Comparative analysis of JmjC domain-containing proteins reveals the potential histone demethylases in *Arabidopsis* and rice. *J. Integr. Plant Biol.* **50,** 886–896 (2008).

110. Juranić, M. & Dresselhaus, T. Phylogenetic analysis of the expansion of the *MATH-BTB* gene family in the grasses. *Plant Signal. Behav.* **9**, e28242 (2014).

111. Smaczniak, C. *et al.* Characterization of MADS-domain transcription factor complexes in *Arabidopsis* flower development. *Proc. Natl. Acad. Sci. USA* **109**, 1560–1565. (2012).

112. Franzke, A., Lysak, M.A., Al-Shehbaz, I.A., Koch, M.A. & Mummenhoff, K. Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends Plant. Sci.* **16**, 108–116 (2011).

113. Kieffer, M., Neve, J. & Kepinski, S. Defining auxin response contexts in plant development. *Curr. Opin. Plant Biol.* **13**, 12–20 (2010).

114. Omidbakhshfard, M.A., Proost, S., Fujikura, U. & Mueller-Roeber, B. Growth-Regulating Factors (GRFs): a small transcription factor family with important functions in plant biology. *Mol. Plant* **8**, 998–1010 (2015).

115. Ciftci-Yilmaz, S. & Mittler, R. The zinc finger network of plants. *Cell. Mol. Life Sci.* **65**, 1150–1160 (2008).

116. Maia, L.G. *et al*. Caracterização do mutante no fator de transcrição ZF-GRF do nódulo de fixação de nitrogênio de *Medicago truncatula*. PhD dissertation (2013).

117. Tsukada, Y. *et al.* 2006. Histone demethylation by a family of JmjC domain-containing proteins. *Nature* **439**, 811–816 (2006).

118. Pagnussat, G.C. *et al.* Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis*. *Development* **132**, 603–614 (2005).

119. Yu, X. *et al.* Modulation of brassinosteroid-regulated gene expression by jumonji domain-containing proteins ELF6 and REF6 in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **105**, 7618–7623 (2008).

120. Searle, I.R., Pontes, O., Melnyk, C.W., Smith, L.M. & Baulcombe, D.C. JMJ14, a JmjC domain protein, is required for RNA silencing and cell-to-cell movement of an RNA silencing signal in *Arabidopsis*. *Genes Dev.* **24,** 986–991 (2010).

121. Lu. F., Cui, X., Zhang, S., Liu, C. & Cao, X. JMJ14 is an H3K4 demethylase regulating flowering time in *Arabidopsis*. *Cell Res.* **20**, 387–390 (2010).

122. Noh, B. *et al.* Divergent roles of a pair of homologous jumonji/zinc-finger-class transcription factor proteins in the regulation of Arabidopsis flowering time. *Plant Cell* **16**, 2601–2613 (2004).

123. Manassero, N.G., Viola, I.L., Welchen, E. & Gonzalez, D.H. TCP transcription factors: architectures of plant form. *Biomol. Concepts* **4**, 111–127 (2013).

124. Doebley, J., Stec, A. & Gustus, C. *teosinte branched1* and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics* **141**, 333–346 (1995).

125. Viola, I.L., Uberti Manassero, N.G., Ripoll, R. & Gonzalez, D.H. The *Arabidopsis* class I TCP transcription factor AtTCP11 is a developmental regulator with distinct DNA-binding properties due to the presence of a threonine residue at position 15 of the TCP domain. *Biochem J.* **435**, 143–155 (2011).

126. Chevalier, F. *et al.* GeBP and GeBP-like proteins are noncanonical leucine-zipper transcription factors that regulate cytokinin response in Arabidopsis. *Plant Physiol.* **146,** 1142–1154 (2008).

127. Argueso, C.T. *et al.* Two-component elements mediate interactions between cytokinin and salicylic acid in plant immunity. *PLoS Genet.* **8**, e1002448 (2012).

128. Perazza, D. *et al.* GeBP/GPL transcription factors regulate a subset of *CPR5*-dependent processes. *Plant Physiol.* **157**, 1232–1242 (2011).

129. Hwang, I., Chen, H.-C. & Sheen, J. Two-component signal transduction pathways in Arabidopsis. *Plant Physiol.* **129,** 500–515 (2002).

130. Strayer, C. *et al.* Cloning of the *Arabidopsis* clock gene *TOC1*, an autoregulatory response regulator homolog. *Science* **289**, 768–771 (2000).

131. Pin, P.A. *et al.* The role of a pseudo-response regulator gene in life cycle adaptation and domestication of beet. *Curr. Biol.* **22**, 1095–1101 (2012).

132. Swaminathan, K., Peterson, K. & Jack, T. The plant B3 superfamily. *Trends Plant Sci.* **13,** 647–655 (2008).

133. McCarty, D.R. *et al.* The *Viviparous-1* developmental gene of maize encodes a novel transcriptional activator. *Cell* **66**, 895–905 (1991).

134. Hu, Y.X., Wang, Y.X., Liu, X.F. & Li, J.Y. *Arabidopsis* RAV1 is down-regulated by brassinosteroid and may act as a negative regulator during plant development. *Cell Res.* **14**, 8–15 (2004).

135. Peng, F.Y. & Weselake, R.J. Genome-wide identification and analysis of the B3 superfamily of transcription factors in Brassicaceae and major crop plants. *Theor. Appl. Genet.* **126**, 1305–1319 (2013).

136. Sanseverino, W. *et al.* PRGdb 2.0: towards a community-based database model for the analysis of R-genes in plants. *Nucleic Acids Res.* **41**, D1167–D1171 (2012).

137. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput Biol.* **7**, e1002195 (2011).

138. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).

139. Jeong, S., Trotochaud, A.E. & Clark, S.E. The Arabidopsis *CLAVATA2* gene encodes a receptor-like protein required for the stability of the CLAVATA1 receptor-like kinase. *Plant Cell* **11**, 1925–1934 (1999).

140. Holub, E.B. The arms race is ancient history in *Arabidopsis*, the wildflower. *Nat. Rev. Genet.* **2**, 516–527 (2001).

141. Kim, S. *et al.* Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nat. Genet.* **46**, 270–278 (2014).

142. Walter, M.H. & Strack, D. Carotenoids and their cleavage products: biosynthesis and functions. *Nat. Prod. Rep.* **28**, 663–692 (2011).

143. Alder, A. *et al.* The path from β-carotene to carlactone, a strigolactone-like plant hormone. *Science* **335**, 1348–1351 (2012).

144. Ruiz-Sola, M.Á. & Rodriguez-Concepcion, M. Carotenoid biosynthesis in Arabidopsis: a colorful pathway. *Arabidop. Book* **10**, e0158 (2012).

145. Fraser, P.D. & Bramley, P.M. The biosynthesis and nutritional uses of carotenoids. *Prog. Lipid Res.* **43**, 228–265 (2004).

146. Biesalski, H.K., Chichili, G.R., Frank, J., von Lintig, J. & Nohr, D. Conversion of β-carotene to retinal pigment. *Vitam. Horm.* **75**, 117–130 (2007).

147. Laferriere, L. & Gabelman, W.H. Inheritance of color, total carotenoids, alpha-carotene, and beta-carotene in carrots, *Daucus carota*, L. *Proc. Am. Soc. Hort. Sci.* **93**, 408–418. (1968).

148. Umiel, N. & Gabelman, W.H. Inheritance of root color and carotenoid synthesis in carrot, *Daucus carota*, L.: orange vs. red. *J. Am. Soc. Hort. Sci.* **97**, 453–460 (1972).

149. Buishand, J.G. & Gabelman, W.H. Investigations on the inheritance of color and carotenoid content in phloem and xylem of carrot roots (*Daucus carota* L.). *Euphytica* **28**, 611–632 (1979).

150. Goldman, I.L. & Breitbach, D.N. Inheritance of a recessive character controlling reduced carotenoid pigmentation in carrot (*Daucus carota* L.). J*. Hered.* **87**, 380–382 (1996).

151. Santos, C.A. & Simon, P.W. QTL analyses reveal clustered loci for accumulation of major provitamin A carotenes and lycopene in carrot roots. *Mol. Genet. Genomics* **268**, 122–129 (2002).

152. Arango, J., Jourdan, M., Geoffriau, E., Beyer, P. & Welsch, R. Carotene hydroxylase activity determines the levels of both α-Carotene and total carotenoids in orange carrots. *Plant Cell* **26**, 2223–2233 (2014).

153. Bowman, M.J., Willis, D.K. & Simon, P.W. Transcript abundance of phytoene synthase 1 and phytoene synthase 2 is associated with natural variation of storage root carotenoid pigmentation in carrot. *J. Am. Soc. Hort. Sci.* **139**, 63–68 (2014).

154. Wang, H., Ou, C.G., Zhuang, F.Y. & Ma, Z.G. The dual role of phytoene synthase genes in carotenogenesis in carrot roots and leaves. *Mol. Breed.* **34**, 2065–2079 (2014).

155. Simon, P.W. Inheritance and expression of purple and yellow storage root color in carrot. *J. Hered*. **87**, 63–66 (1996).

156. Bradeen, J.M., Vivek, B.S. & Simon, P.W. Detailed genetic mapping of the Y_2 carotenoid locus in carrot. *J. Appl. Genet.* **38A**, 28–32 (1997).

157. Just, B.J., Santos, C.A., Yandell, B.S. & Simon, P.W. Major QTL for carrot color are positionally associated with carotenoid biosynthetic genes and interact epistatically in a domesticated x wild carrot cross. *Theor. Appl. Genet.* **119**, 1155–1169 (2009).

158. Just, B.J. Genetic mapping of carotenoid pathway structural genes and major gene QTLS for caroteniod accumulation in wild and domesticated carrot (*Daucus carota* L.) PhD dissertation (2004).

159. Simon, P.W & Wolff, X.Y. Carotenes in typical and dark orange carrots. *J. Agric. Food Chem.* **35**, 1017–1022 (1987).

160. Simon, P.W., Wolff, X.Y., Peterson, C.E. & Kammerlohr, D.S. High carotene mass carrot population. *HortSci.* **24**, 174–175 (1989).

161. Boiteux, L.S., Fonseca, M.E. & Simon, P.W. Effects of plant tissue and DNA purification method on randomly amplified polymorphic DNA-based genetic fingerprinting analysis in carrot. *J. Am. Soc. Hort. Sci.* **124**, 32–38 (1999).

162. Bland, J.M. & Altman, D.G. Multiple significance tests: the Bonferroni method *BMJ.* **310**, 170 (1995).

163. Broman, K.W. & Sen, S. A guide to QTL mapping with R/qtl. (Springer, New York, 2009).

164. Untergasser, A. et al. Primer3—New capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012)

165. Leigh, J.W. & Bryant, D. POPART: full-feature software for haplotype network construction. *Methods Ecol Evol* **6**, 1110-1116 (2015).

166. Barrett, J.C., Fry, B., Maller, J., & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-265 (2004).

167. Nei, M. & Li, W.H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA*. **76**, 5269–5273 (1979).

168. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559 (2008).

169. Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., Karthikeyan, A.S., Lee, Nelson, W.D., Ploetz, L., Singh, Wensel, A.S., & Huala, E. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.***40**, D1202–D1210 (2012).

170. Du, Z., Zhou, X., Ling, Y., Zhang, Z., & Su, Z. agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res.* **38**, W64-70 (2010).

171. Mi, H., Muruganujan, A., & Thomas, P.D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41**, D377-86 (2013).

172. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300** 1005–1016 (2000).

173. Winkel-Shirley, B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol.* **126**, 485–493 (2001).

174. Montilla, E.C., Arzaba, M.R., Hillebrand, S. & Winterhalter, P. Anthocyanin composition of black carrot (*Daucus carota* ssp. *sativus* var. *atrorubens* Alef.) cultivars Antonina, Beta Sweet, Deep Purple, and Purple Haze. *J. Agri. Food Chem.* **59**, 3385–3390 (2011).

175. Banga, O. Origin and domestication of the western cultivated carrot. *Genet. Agrar.* **17**, 357–370 (1963).

176. Yildiz, M. *et al.* Expression and mapping of anthocyanin biosynthesis genes in carrot. *Theor. Appl. Genet.* **126**, 1689–1702 (2013).

177. Nagegowda, D.A. Plant volatile terpenoid metabolism: biosynthetic genes, transcriptional regulation and subcellular compartmentation. *FEBS Lett.* **584**, 2965–2973 (2010).

178. Wagner, K.H. & Elmadfa, I. Biological relevance of terpenoids. Overview focusing on mono-, di- and tetraterpenes. *Ann. Nutr. Metab.* **47**, 95–106 (2003).

179. Buttery, R.G., Seifert, R.M., Guadagni, D.G., Black, D.R. & Ling, L. Characterization of some volatile constituents of carrots. *J. Agric. Food Chem.* **16**, 1009–1015 (1968).

180. Buttery, R.G. & Takeoka, G.R. Cooked carrot volatiles. AEDA and odor activity comparisons. Identification of linden ether as an important aroma component. *J. Agric. Food Chem.* **61**, 9063–9066 (2013).

181. Simon, P.W., Peterson, C.E. & Lindsay, R.C. Correlations between sensory and objective parameters of carrot flavor. *J. Agric. Food Chem.* **28,** 559–562 (1980).

182. Simon, P.W., Peterson, C.E. & Lindsay, R.C. Genotype, soil, and climate effects on sensory and objective components of carrot flavor. *J. Amer. Soc. Hort. Sci.* **107**, 644–648 (1982).

183. Simon, P.W. *et al.* in *Carrot. Handbook of Plant Breeding, Vegetables II.* (Eds. Prohens, J. & Nuez, F.) 327–357 (Springer, New York, NY, 2008).

184. Yahyaa, M. *et al.* Identification and characterization of terpene synthases potentially involved in the formation of volatile terpenes in carrot (*Daucus carota* L.) roots. *J. Agric. Food Chem.* **63**, 4870–4878 (2015).

185. Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* **66**, 212–229 (2011).

186. Just, B.J. *et al.* Carotenoid biosynthesis structural genes in carrot (*Daucus carota*): isolation, sequence-characterization, single nucleotide polymorphism (SNP) markers and genome mapping. *Theor. Appl. Genet*. **114**, 693–704 (2007).

187. Giuliano, G. Plant carotenoids: genomics meets multi-gene engineering. *Curr. Opin. Plant Biol.* **19**, 111–117 (2014).

188. Saladié, M., Wright, L.P., Garcia-Mas, J., Rodríguez-Concepción, M. & Phillips, M.A. The 2-C-methylerythritol 4-phosphate pathway in melon is regulated by specialized isoforms for the first and last steps. *J. Exp. Bot.* **65**, 5077–5092 (2014).

189. Yahyaa, M. *et al.* Formation of norisoprenoid flavor compounds in carrot (*Daucus carota* L.) roots: characterization of a cyclic-specific carotenoid cleavage dioxygenase 1 gene. *J. Agric. Food Chem.* **61**, 12244–12252 (2013).

190. Martin, D.M. *et al.* Functional annotation, genome organization and phylogeny of the grapevine (*Vitis vinifera*) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays. *BMC Plant Biol.* **10**, 226 (2010).

191. Falara, V. *et al.* The tomato terpene synthase gene family. *Plant Physiol.* **157**, 770–789 (2011).

192. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105 (2005).

Additional references cited in supplementary figures and supplementary tables

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

194. Guo, S. *et al.* The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat. Genet.* **45**, 51–58 (2013).

195. Xu, Q. *et al.* The draft genome of sweet orange (*Citrus sinensis*). *Nat. Genet.* **45**, 59–66 (2013).

196. Huang, S. *et al.* Draft genome of the kiwifruit *Actinidia chinensis*. *Nat. Commun.* **4,** 2640 (2013).

197. Varshney, R.K. *et al.* Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30,** 83–89 (2012).

198. Argout, X. *et al.* The genome of *Theobroma cacao*. *Nat. Genet.* **43,** 101–108 (2011).

199. Al-Mssallem, I.S. *et al.* Genome sequence of the date palm *Phoenix dactylifera* L. *Nat. Commun.* **4,** 2274 (2013).

200. Huang, S. *et al.* The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **41,** 1275–1281 (2009).

201. Wang, K. *et al.* The draft genome of a diploid cotton *Gossypium raimondii*. *Nat. Genet.* **44,** 1098–1103 (2012).

202. Velasco, R. *et al.* The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nat. Genet.* **42,** 833–839 (2010).

203. Varshney, R.K. *et al.* Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.* **31,** 240–246 (2013).

204. Yuan, Y. & Wessler, S.R. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl. Acad. Sci. USA* **108,** 7884–7889 (2011).

205. Martin-Trillo, M. & Cubas, P. TCP genes: a family snapshot ten years later. *Trends Plant Sci.* **15***,* 31–39 (2010).

206. Sakai, H. *et al.* ARR1, a transcription factor for genes immediately responsive to cytokinins. *Science* **294,** 1519–1521 (2015).

207. Tian, C., Wan, P., Sun, S., Li, J. & Chen, M. Genome-wide analysis of the GRAS gene family in rice and *Arabidopsis*. *Plant Mol. Biol.* **54**, 519–532 (2004).

208. Schauser, L., Roussis, A., Stiller, J. & Stougaard, J. A plant regulator controlling development of symbiotic root nodules. *Nature* **402,** 191–195 (1999).

209. Kim, J. H., Choi, D. & Kende, H. The AtGRF family of putative transcription factors is involved in leaf and cotyledon growth in Arabidopsis. *Plant J*. **36**, 94–104 (2003).

210. Kawaoka, A. *et al.* Functional analysis of tobacco LIM protein Ntlim1 involved in lignin biosynthesis. *Plant J*. **22,** 289–301 (2000).

211. Husbands, A., Bell, E.M., Shuai, B., Smith, H.M.S. & Springer, P.S. Lateral organ boundaries defines a new family of DNA-binding transcription factors and can interact with specific bHLH proteins. *Nucleic Acids Res.* **35,** 6663–6671 (2007).

212. Lijavetzky, D., Carbonero, P. & Vicente-Carbajosa, J. Genome-wide comparative phylogenetic analysis of the rice and Arabidopsis Dof gene families. *BMC Evol. Biol.* **3,** 17 (2003).

213. Parenicová L. *et al.* Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world. *Plant Cell* **15,** 1538–1551 (2003).

214. Nam, J., dePamphilis, C.W., Ma, H. & Nei, M. Antiquity and evolution of the MADS-box gene family controlling flower development in plants. *Mol. Biol. Evol.* **20**, 1435–1447 (2003).

215. Liu, H., Frankel, L.K. & Bricker, T.M. Functional complementation of the *Arabidopsis thaliana* psbo1 mutant phenotype with an N-terminally His6-tagged PsbO-1 protein in photosystem II. *Biochim. Biophys. Acta.* **1787**, 1029–1038 (2009)

216. Yakushevska, A.E. *et al.* Supermolecular organization of photosystem II and its associated light-harvesting antenna in *Arabidopsis thaliana*. *Eur. J. Biochem.* **268,** 6020–6028 (2001).

217. Liu, H., Frankel, L.K. & Bricker, T.M. Characterization and complementation of a psbR mutant in *Arabidopsis thaliana*. *Arch. Biochem. Biophys.* **489,** 34–40 (2009).

218. Thomas, L., Marondedze, C., Ederli, L., Pasqualini, S. & Gehring, C. Proteomic signatures implicate cAMP in light and temperature responses in *Arabidopsis thaliana*. *J. Proteomics* **83,** 47–59 (2013).

219. García-Cerdán, J.G. *et al.* The PsbW protein stabilizes the supramolecular organization of photosystem II in higher plants. *Plant J*. **65**, 368–381 (2011).

220. Seok, M. *et al.* AtFKBP16-1, a chloroplast lumenal immunophilin, mediates response to photosynthetic stress by regulating PsaL stability. *Physiol Plant.* **150**, 620–631 (2014).

221. Granlund, I., Hall, M., Kieselbach, T. & Schro, W.P. Light induced changes in protein expression and uniform regulation of transcription in the thylakoid lumen of *Arabidopsis thaliana*. *PLoS ONE* **4,** e5649 (2009).

222. Ihnatowicz, A. *et al.* Mutants for photosystem I subunit D of *Arabidopsis thaliana*: effects on photosynthesis, photosystem I stability and expression of nuclear genes for chloroplast functions. *Plant J*. **37**, 839–852 (2004).

223. Fristedt, R. & Vener, A.V. High light induced disassembly of Photosystem II supercomplexes in Arabidopsis requires STN7-dependent phosphorylation of CP29. *PLoS ONE* **6,** e24565 (2011).

224. Wientjes, E. & Croce, R. The light-harvesting complexes of higher-plant Photosystem I: Lhca1/4 and Lhca2/3 form two red-emitting heterodimers. *Biochem. J.* **485,** 477–485 (2011).

225. Fristedt, R. *et al.* PHOTOSYSTEM II PROTEIN33, a protein conserved in the plastid lineage, is associated with the chloroplast thylakoid membrane and provides stability to photosystem II supercomplexes in Arabidopsis. *Plant Physiol.* **167,** 481–492 (2015).

226. Tsunoyama, Y. *et al.* Blue light-induced transcription of plastid-encoded psbD gene is mediated by a nuclear-encoded transcription initiation factor, AtSig5. *Proc. Natl. Acad. Sci. USA* **101,** 3304–3309 (2004).

227. Yu, H. *et al.* Downregulation of chloroplast RPS1 negatively modulates nuclear heat-responsive expression of *HsfA2* and its target genes in *Arabidopsis*. *PLoS Genet.* **8,** e1002669 (2012).

228. Xing, Y. *et al.* MKK5 regulates high light-induced gene expression of Cu/Zn superoxide dismutase 1 and 2 in Arabidopsis. *Plant Cell Physiol.* **54,** 1217–1227 (2013).

229. Jurić, S. *et al.* Tethering of ferredoxin: NADP+ oxidoreductase to thylakoid membranes is mediated by novel chloroplast protein TROL. *Plant J*. **60**, 783–794 (2009).

230. Jaru-Ampornpan, P. *et al.* Mechanism of an ATP-independent protein disaggregase: II. distinct molecular interactions drive multiple steps during aggregate disassembly. *J. Biol. Chem.* **288,** 13431–13445 (2013).

231. Stengel, K.F., Holdermann, I., Cain, P., Robinson, C. & Wild, K. Structural basis for specific substrate recongition by the chloroplast signal recognition particle protein cpSRP43. *Science* **321,** 253–256 (2015).

232. Pontier, D., Albrieux, C., Joyard, J., Lagrange, T., & Block, M.A. Knock-out of the magnesium protoporphyrin IX methyltransferase gene in Arabidopsis. Effects on chloroplast development and on chloroplast-to-nucleus signaling. *J. Biol. Chem.* **282,** 2297–2304 (2007).

233. Takahashi, K., Takabayashi, A., Tanaka, A. & Tanaka, R. Functional analysis of light-harvesting-like protein 3 (LIL3) and its light-harvesting chlorophyll-binding motif in *Arabidopsis*. *J. Biol. Chem.* **289**, 987–999 (2014).

234. Moseley, J.L. *et al.* Reciprocal Expression of Two Candidate Di-Iron Enzymes Affecting Photosystem I and Light-Harvesting Complex Accumulation. *Plant Cell*, **14,** 673–688 (2002).

235. Sakuraba, Y., Tanaka, R., Yamasato, A. & Tanaka, A. Determination of a Chloroplast Degron in the Regulatory Domain of Chlorophyllide α Oxygenase. *J. Biol. Chem.* **284,** 36689–36699 (2009).

236. Wang, P. *et al.* One divinyl reductase reduces the 8-vinyl groups in various intermediates of chlorophyll biosynthesis in a given higher plant species, but the isozyme differs between species. *Plant Physiol.* **161,** 521–534 (2013).

237. Ganjewala, D., Kumar, S. & Luthra R. An account of cloned genes of Methyl-erythritol-4-phosphate pathway of isoprenoid biosynthesis in plants. *Curr. Issues Mol. Biol.* **11,** 35–46 (2008).

238. Zhang, Y. *et al.* Two Arabidopsis cytochrome P450 monooxygenases, CYP714A1 and CYP714A2, function redundantly in plant development through gibberellin deactivation. *Plant J.* **53,** 342–353 (2011).

239. Wuille, S. Role of WRINKLED1 in the transcriptional regulation of glycolytic and fatty acid biosynthetic genes in Arabidopsis. *Plant J*. **60**, 933–947 (2009).

240. Qin, F. *et al.* SPINDLY, a negative regulator of gibberellic acid signaling, is involved in the plant abiotic stress response. *Plant Physiol.* **157,** 1900–1913 (2011).

241. Morrone, D., Chen, X., Coates, R.M. & Peters, R.J. Characterization of the kaurene oxidase CYP701A3, a multifunctional cytochrome P450 from gibberellin biosynthesis. *Biochem. J.* **344,** 337–344 (2010).

242. Lewis, D.R. *et al.* A kinetic analysis of the auxin transcriptome reveals cell wall remodeling proteins that modulate lateral root development in Arabidopsis. *Plant Cell* **25,** 3329–3346 (2013).

243. Zhang Z.-B. *et al.* Arabidopsis Inositol Polyphosphate 6-/3-Kinase (*AtIpk2β*) Is Involved in Axillary Shoot Branching. *Plant Physiol.* **144,** 942–951 (2007).

244. Babiychuk, E. *et al.* Plastid gene expression and plant development require a plastidic protein of the mitochondrial transcription termination factor family. *Proc. Natl. Acad. Sci. USA* **108,** 6674–6679 (2011).

245. Tillich, M. *et al.* Chloroplast ribonucleoprotein CP31A is required for editing and stability of specific chloroplast mRNAs. *Proc. Natl. Acad. Sci. USA* **106,** 6002–6007 (2009).

246. Takano, A., Suetsugu, N., Wada, M. & Kohda, D. Crystallographic and functional analyses of J-domain of JAC1 essential for chloroplast photorelocation movement in *Arabidopsis thaliana*. *Plant Cell Physiol.* **51,** 1372–1376 (2010).

247. Ferro, M. *et al*. Integral membrane proteins of the chloroplast envelope: Identification and subcellular localization of new transporters. *Proc. Natl. Acad. Sci. USA* **99,** 11487–11492 (2002).

248. Nakamura, Y. *et al.* Differential metabolomics unraveling light/dark regulation of metabolic activities in Arabidopsis cell culture. *Planta* **227**, 57–66 (2007).

249. Bang, W.Y., Kim, S.W., Jeong, I.S., Koiwa, H. & Bahk, J.D. The C-terminal region (640–967) of *Arabidopsis* CPL1 interacts with the abiotic stress- and ABA-responsive transcription factors. *Biochem. Biophys. Res. Commun.* **372,** 907–912 (2008).

250. Brosseau, C. & Moffett, P. Functional and genetic analysis identify a role for Arabidopsis ARGONAUTE5 in antiviral RNA silencing. *Plant Cell* **27,** 1742–1754 (2015).

251. Kaewthai, N. *et al.* Group III-A *XTH* genes of Arabidopsis encode predominant xyloglucan endohydrolases that are dispensable for normal growth. *Plant Physiol.* **161,** 440–454 (2013).

252. Ozalvo, R. *et al.* Two closely related members of *Arabidopsis* 13-lipoxygenases (13-LOXs), LOX3 and LOX4, reveal distinct functions in response to plant-parasitic nematode infection. *Mol. Plant. Pathol.* **15,** 319–332 (2014).

253. Yelagandula, R. *et al.* The histone variant H2A.W defines heterochromatin and promotes chromatin condensation in *Arabidopsis*. *Cell* **158**, 98–109 (2014).

254. Ichikawa, T. *et al.* The FOX hunting system: an alternative gain-of-function gene hunting technique. *Plant J*. **48**, 974–985 (2006).

255. Knoetzel, J., Mant, A., Haldrup, A., Jensen, P.E. & Scheller, H V. PSI-O, a new 10-kDa subunit of eukaryotic photosystem I. *FEBS Lett.* **510,** 145–148 (2002).

256. Liu, J. & Last, R.L. A land plant-specific thylakoid membrane protein contributes to photosystem II maintenance in *Arabidopsis thaliana*. *Plant J*. **82**, 731–743 (2015).

257. Reinbothe, C. *et al.* A pentapeptide motif related to a pigment binding site in the major light-harvesting protein of photosystem II, LHCII, governs substrate-dependent plastid import of NADPH:protochlorophyllide oxidoreductase A. *Plant Physiol.* **148,** 694–703 (2008).

258. Dinh, T.T. *et al.* DNA topoisomerase 1α promotes transcriptional silencing of transposable elements through DNA methylation and histone lysine 9 dimethylation in *Arabidopsis*. *PLoS Genet.* **10,** e1004446 (2014).

259. Taochy, C. *et al.* The Arabidopsis root stele transporter NPF2.3 contributes to nitrate translocation to shoots under salt stress. *Plant J*. **83**, 466–479 (2015).

260. Kaneda, M. *et al.* ABC transporters coordinately expressed during lignification of Arabidopsis stems include a set of ABCBs associated with auxin transport. *J. Exp. Bot*. **62,** 2063–2077 (2011).

261. Kim, S. *et al.* An atypical soybean leucine-rich repeat receptor-like kinase, GmLRK1, may be involved in the regulation of cell elongation. *Planta* **229**, 811–821 (2009).

262. Fäldt, J., Arimura, G., Gershenzon, J., Takabayashi, J. & Bohlmann, J. Functional identification of AtTPS03 as (E)-β-ocimene synthase: a monoterpene synthase catalyzing jasmonate- and wound-induced volatile formation in *Arabidopsis thaliana*. *Planta* **216**, 745–751 (2003).

263. Denancé, N., Ranocha, P., Martinez, Y., Sundberg, B. & Goffner, D. Light-regulated compensation of wat1 (walls are thin1) growth and secondary cell wall phenotypes is auxin-independent. *Plant Signal Behav*. **5,** 1302–1304 (2010).

264. Brandão, M.M., Dantas, L.L., & Silva-Filho, M.C. AtPIN: *Arabidopsis thaliana* protein interaction network. *BMC Bioinform.* **10**, 454 (2009).

9. List of Supplementary Figures

Supplementary Figure 1: Scheme of the carrot genome assembly pipeline. In Phase I, quality filtered Illumina data from eight insert libraries (Supplementary Table 49) were assembled using SOAPdenovo (http://soap.genomics.org.cn) producing the carrot assembly v1.0, which included contigs and scaffolds. In Phase II, unambiguously aligned sequences from mapped molecular markers, BAC end sequences and 20 and 40 kb Illumina MPE were visualized in Gbrowse to manually inspect and correct chimeric regions and construct superscaffolds. This process produced the carrot assembly v1.1 which included contigs, scaffolds and superscaffolds. In Phase III, the integrated linkage map was used to anchor superscaffolds and construct the nine carrot pseudomolecules (chromosomes). The final assembly named carrot assembly v2.0 includes pseudomolecules and the remaining unanchored contigs, scaffolds and superscaffolds.

Supplementary Figure 2: Example of a chimeric scaffold in the carrot assembly v1.0. The yellow vertical highlight identifies the chimeric region on scaffold13 of carrot assembly v1.0 (Panel A and B) and its correction in the carrot assembly v2.1 (Panel C). Panel A) GBrowse window of scaffold13 (carrot assembly v1.0). a1: indicates the mapping location of markers unambiguously aligned to scaffold13; markers in purple mapped to LG3(Chr 3) and markers in brown mapped to LG5(Chr 5); a2-a3: 40 kb paired end reads (PE) and DH BAC end sequences that unambiguously aligned to scaffold13; Panel B) GBrowse enlarged window of scaffold13 covering the chimeric region. b1-b2; enlarged window of the region where neither the 40 kb nor the BAC end sequences spanned a contiguous connection on scaffold13; b3: misassembly point were the scaffold13 has been split; b4: 40 kb PE reads that unambiguously aligned to one side of the misassembled region on scaffold13. The PE reads in the left side unambiguously link to scaffold192, the PE reads on the right unambiguously link to scaffold 177. Panel C) GBrowse window of carrot assembly v2.0, CH3, superscaffold6 (CH3.6). c1: order of scaffolds in superscaffold6 spanning the corrected chimeric assembly. c2: mapping location of markers unambiguously aligned to superscaffold CH3.6; markers on both side of the corrected chimeric assembly mapped to LG3(Chr 3) at 30.4 cM; c3: BAC end sequences that unambiguously aligned to superscaffold CH3.6. The BAC end sequences span the region connecting the misassembly point on scaffold13 and its connection to scaffold192 at an expected average distance of 150 kb.

Supplementary Figure 3: Relationships between sequence depth and GC content of the carrot genome. A: Read depth distribution on the carrot genome assembly. B: Relationship between percent of GC content and sequencing depth. The x-axis represents the percent GC content; the y-axis represents the average sequence depth. A 10 kb non-overlapping sliding window was used to calculate the GC content and the average depth. C: Comparisons of percent GC content across five species, including carrot (*Daucus carota*), tomato (*Solanum lycopersicum*), potato (*Solanum tuberosum*), *Arabidopsis thaliana* and cucumber (*Cucumis sativus*).

Supplementary Figure 4: Comparison of the genetic map of population 85036 to the physical map of the DH1 anchored genome. The upper bars indicate pseudomolecules, and the lower bars represent linkage groups, corresponding to the nine chromosomes. The orange blocks and gray lines indicate scaffolds matched to the genetic map. Gray blocks indicate scaffolds that did not match with any markers. Triangles indicate break points between scaffolds.

Supplementary Figure 5: Evaluation of alignment consistency between the 85036 linkage map and the order of sequences in the nine carrot pseudomolecules. The analysis was carried out by comparing adjacent pairs of markers, and pairs with one, two or three intervening mapped markers. Blue bars represent the fraction of markers that matched the order of sequences in the nine pseudomolecules. Red bars represent the median genetic distance between discordant markers. Green bars represent the physical median distance between discordant markers. In pairwise comparisons of adjacent markers, those with one intervening marker, those with two, and those with three intervening markers, the percent of concordant markers ranged from 82 to 98%, and the median genetic distance between discordant markers ranged from 0.08 to 0.18 cM suggesting that inconsistent alignments likely reflect technical accuracy and low map density.

Supplementary Figure 6: FISH-based confirmation of the consistency and coverage of the carrot genome assembly at the telomeric regions of the DH1 chromosomes. Chr 2 (A-D), 6 (E-H), 8 (I-L) and 9 (M-P). DAPI stained pachytene Chr.2 (A), Chr 6 (E), Chr 8 (I) and 9 (M) are converted to black and white images. (B-F-J-N): FISH signals derived from BAC clones that unambiguously aligned to sequences close to the start ('S', green signals) and the ends ('L', red signals) of the pseudomolecules of Chr 2 (B), 6 (F), 8 (J) and 9 (N). BACs A9P11 and 2B20 (red signals) are specific for Chr 6 and 8, respectively. (C-G-K-O) Same chromosomes probed with a telomeric oligo-based probe ('T', orange signals). (D-H-L-P) Merged images demonstrating the co-localization of the 'S' (green signals/arrows) and 'L' (red signals/arrows) BACs with the telomeric probe (orange signals/arrows). Bar scale=5 um.

Supplementary Figure 7: MITE family analysis.
Panel a1-b1-c1: Histograms of unimodal (a1) bimodal (b1) and multimodal (c1) distributions of pairwise distances calculated using the Kimura 2-parameter mode among members of some MITE families. Sharp modal distributions suggest a rapid amplification burst rather than a gradual amplification.
Panel a2-b2-c2: Timetree analysis by Maximum Likelihood method for some MITE families. The length of the branch indicates relative divergence time among elements represented in each phylogenetic tree.
Colors of *DcSto* families' histograms represented in the a1, b1, c1 panels correspond to family colors represented in the divergence time tree (Except for *DcSto*2 and *Dcsto*12 which were too divergent to be aligned with *DcSto*7a, *DcSto*7b, *DcSto*1 and *DcSto*5).

Supplementary Figure 8: Distribution of distance from MITE or *Krak* insert site to nearest predicted gene in carrot DH1. MITE *DcSto* (Panel A, blue line) and *Krak* (Panel B, blue line) elements, and simulated datasets (red lines). The simulated datasets represent the distribution of 1M random insertion sites of 280 or 340 nt, which in turn represent the average sizes of *DcSto* and *Krak* elements, respectively. The maximum interval plotted represents the point where the simulation reaches 95% of the elements that do not overlap genes.

Nature Genetics: doi:10.1038/ng.3565

Supplementary Figure 9: Evolution of tandem repeats in the carrot genome and *Daucus* genus. Panel I: Estimated genome proportion of Cent-Dc-like and CL80 repetitive sequences across six *Daucus* species included in this study. Panel II: Figures on the left illustrate the similarity and structure of the A-B-C-D Cent-Dc monomers in K11, DH1 and *D. syrticus* (Dsyr) organized in a higher order repeat structure (HOR). *D. aureus* (Daur) harbors a single 40 nt monomer most similar to monomer A. The panel on the right illustrates the sequences of the alignments of the 40 nt monomers. Arrows above each base indicates the private polymorphic sites that each individual A-B-C-D monomer has accumulated. Panel III: Phylogenetic analysis of the single ABCD monomers extracted from DH1 BAC end sequences. Panel IV: Schematic illustration of the hypothetical evolution of carrot Cent-Dc sequence.

Supplementary Figure 10: FISH analysis of the CL80 repeat in *Daucus* species with 2n=20. (A) A somatic metaphase cell of *D. littoralis*. (B) FISH signals derived from the CL80 (green) and a telomeric probe (red). (C) Merged image of (A) and (B). CL80 repeat generated interstitial FISH signals on all chromosomes. (D) DAPI-stained meiotic pachytene chromosomes of *D. littoralis* converted to black and white image. (E) FISH signals from the CL80 repeat (red). (F) Merged image of (D) and (E). The interstitial CL80-repeat sites likely span the centromeres of *D. littoralis* chromosomes; yellow arrowheads in (D) and (F) point to several interstitial sites. (G-I) FISH of the telomeric probe (green, G) and the CL80 repeat (red, H) on pachytene chromosomes of *D. littoralis*. (I) Merged image of (G) and (H). (J) A somatic metaphase cell of *D. guttatus*. (K) FISH signals derived from CL80 repeat. (L) Image merged from (J) and (K); white arrows point to four unambiguous interstitial signals. (M) A somatic metaphase cell of *D. guttatus*. (N) FISH signals derived from the CL80 repeat (green) and the telomeric DNA probe (red). (O) A merged image of (M) and (N). Most CL80 signals co-localized with the telomeres. White arrows point to four interstitial signals. Bars = 5 μm.

Supplementary Figure 11: Representative roots of resequenced carrot accessions. Samples include eastern (C1, C2, C5) and western (C9, C12, C13) cultivated (*D. carota* subsp. *sativus*) carrot phenotypes, and examples of wild (*D. carota* subsp. *carota*) carrots. Details for each sample are reported in Supplementary Table 16.

Supplementary Figure 12: Comparative gene analysis. A: Maximum likelihood tree constructed with 312 single copy orthologous genes. Bootstrap values are shown at nodes. The scale is amino acid substitutions per site. B: Time divergence estimation of 13 dicot and monocot plants.

Supplementary Figure 13: Age distribution of 4DTv and Ks analyses. 4DTv analysis (Panel A) and Ks analysis (Panel B) are presented for genes from the Horseweed (*Conyza canadensis*), carrot, *A. thaliana*, kiwi and lettuce genomes. X-axis indicates 4DTv values and Ks distance, respectively; Y-axis indicates percentage of ortholog/paralog gene pairs.

Supplementary Figure 14: Age distribution of 4DTv for carrot gene paralogs descending from the seven ancestral core eudicot chromosomes. A1 to A19 are the ancestral protochromosomes.

Nature Genetics: doi:10.1038/ng.3565

Supplementary Figure 15: Gene family cluster analysis. The Poaceae group includes orthologous genes from *O. sativa* and *S. bicolor* genomes, representatives of the Monocot clade. The Rosids group includes orthologous genes from *A. thaliana*, *A. lyrata*, *B. rapa*, *C. papaya*, *P. persica* and *V. vinifera* genomes. The Asterids group includes orthologous genes from *D. carota*, *S. lycopersicum*, *S. tuberosum* and *A. chinensis* genomes.

Supplementary Figure 16: Phylogenetic analysis of the JMJD2 subfamily of JmjC transcription factors. The phylogenetic tree was constructed using the Neighbor-Joining method and a bootstrap test was performed with 1,000 iterations. Bootstrap values over 50% are shown. The scale bar (0.2) shows the number of amino acid substitutions per site. Previously characterized JmjC genes from *A. thaliana* are labeled with AT prefix[109]. JMJD2 sub-group 3 was expanded in carrot and it includes the functionally characterized Arabidopsis *REF6*. Carrot DCAR_016424 and DCAR_026201 were retained from the eudicot gamma whole genome triplication.

Supplementary Figure 17: Phylogenetic analysis of the TCP transcription factors. The phylogenetic tree was constructed using the Neighbor-Joining method and a bootstrap test was performed with 1,000 iterations. Bootstrap values over 50% are shown. The scale bar (0.2) shows the number of amino acid substitutions per site. Previously characterized Arabidopsis TCPs are labeled with AT prefix[125]. TCP sub-group 11 (highlighted with thick black branches) has expanded in carrot and it includes a functionally characterized Arabidopsis *At-TCP11*.

Supplementary Figure 18: Phylogenetic analysis of the GeBP transcription factors. The phylogenetic tree was constructed using the Neighbor-Joining method and a bootstrap test was performed with 1,000 iterations. Bootstrap values over 50% are shown. The scale bar (0.5) shows the number of amino acid substitutions per site. Annotated Arabidopsis GeBP genes from the plant transcription factor database are labeled with AT prefix. GeBP sub-group 1 has expanded in carrot and it includes four genes that are homologous to the functionally characterized Arabidopsis *GPL1-2-3* (ref. 126).

Supplementary Figure 19: Phylogenetic analysis of the response regulator genes (RR). The phylogenetic tree was constructed using the Neighbor-Joining method and a bootstrap test was performed with 1,000 iterations. Bootstrap values over 50% are shown. The scale bars (0.2 and 0.5) show the number of amino acid substitutions per site. A: phylogenetic tree of type A-B-C regulators. B: phylogenetic tree of PRR regulators. Previously characterized RR genes from *A. thaliana* are labeled with AT prefix[129].

Supplementary Figure 20: Phylogenetic analysis of the REM sub-groups 1 and 4 of the B3-domain transcription factors. The phylogenetic tree was constructed using the Neighbor-Joining method and a bootstrap test was performed with 1,000 iterations. Bootstrap values over 50% are shown. The scale bar (0.5) shows the number of amino acid substitutions per site. Previously characterized REM genes from *A. thaliana* are labeled with AT prefix[132]. Sub-group 1 was expanded in carrot and it includes the functionally characterized Arabidopsis *VRN1*. Sub-group 4 is carrot specific and it is the sub-group most closely related to sub-group1.

Supplementary Figure 21: Phylogenetic analysis of the CNL R gene class. The phylogenetic tree was constructed using the Neighbor-Joining method and a bootstrap test was performed with 1,000 iterations. Bootstrap values over 50% are shown. The scale bar (0.2) shows the number of amino acid substitutions per site.

Nature Genetics: doi:10.1038/ng.3565

Supplementary Figure 22: Carrot R genes. A: Distribution of candidate R genes in the nine carrot chromosomes. Brackets on the right side of each chromosome indicate the set of R genes organized in clusters (CL) or arrays (Ar), and the length in kb of the genomic region spanning each cluster (in parentheses). B: Summary distribution of each R gene class along the nine carrot chromosomes. Bars represent the percentage of genes for each family located on each chromosome. Numbers above the bars indicate the number of genes.

Supplementary Figure 23: Manhattan plots for marker-trait associations using GLM analysis. a) Lutein content in population 97837. b) Total carotenoids in population 70796. The gray dotted line indicates significance cut-off after using a Bonferroni correction.

Supplementary Figure 24: Population 70796 carotenoid QTL mapping results. Map on the left corresponds to LG5. Markers highlighted in red indicate mapping positions (cM) flanking the QTL detected for total carotenoid content. Map on the right indicates the corresponding physical position in carrot chromosome 5 with markers flanking the QTL interval. DCAR_032551 corresponds to the candidate gene controlling the *Y* locus. To optimize the visualization of the QTL region, the start (from position 0 to 15.3 cM) and the end (from position 51 to 63 cM) portions of LG5 are not shown.

Supplementary Figure 25: Fine mapping of the carrot *Y* locus. A: Linkage blocks associated with dark orange (dOr) and pale orange (pOr) root phenotypes. B: Linkage blocks associated with yellow (Y) and White (W) root phenotypes. Overlapping linkage blocks associated with dOr and Y phenotypes and spanning 75 kb were identified across the two populations as the most significant location harboring the gene controlling the *Y* locus.

Supplementary Figure 26: Schematic representation of the transcriptome polymorphisms detected in DCAR_032551. "Wild" indicates the wild type allele without the insertions. dOrF1 to dOrF4 indicate all the isoforms identified in the *Y* mutant which includes a 212 nt insertion in the second exon. The relative percent for each isoform is reported. Alt-y represents the *Y* allele identified in two resequenced samples, C1 and I2, that contain a 1nt insertion in the second exon.

Supplementary Figure 27: Haplotype network analysis. Panel B includes SNP data from all *D. carota* wild and cultivated accessions in the haplotype block associated with the *Y* locus. Panels A and C include SNPs from the regions covering 8 kb upstream and 8 kb downstream from the start and the end of the haplotype block, respectively.

Supplementary Figure 28: 17-mer estimated carrot genome size. The x-axis is depth (X). The y-axis is the proportion of sequences which represents the frequency at that depth, divided by the total frequency of all the depths.

Nature Genetics: doi:10.1038/ng.3565

Supplementary Figure 29: Multi-dimensional topography of carrot chromosomes. Chromosome 1 is illustrated in Fig. 1. a) Carrot integrated linkage map. Vertical bar to the left indicates the genetic distance in cM. Lines to the right connect a subset of markers to the assembled pseudomolecule. b) From the left to the right: linkage map distance (cM/Mb), predicted genes (% nucleotide per 200 kb), RNA-Seq from 20 different sequencing libraries (% nucleotide per 200 kb), class I and class II repetitive sequences (% nucleotide per 200 kb), non-coding RNA (% nucleotide per 200 kb), SNPs detected comparing resequencing data from 35 different genotypes (number of SNPs per 100 kb). Horizontal gray lines represent gaps in the pseudomolecule. c) DNA pseudomolecules. Gaps between superscaffolds are indicated by gray horizontal lines. Location of BAC probes hybridized to pachytene chromosomes are identified by horizontal green and red lines and labeled on the right. Horizontal orange lines indicate location of the telomeric repetitive sequence (T). Dark gray lines at the right indicate the location of BAC clone end sequences previously used to anchor the genetic map to carrot chromosomes[10].

Supplementary Figure 30: Anchoring the carrot genome assembly to the carrot reference genetic map. The carrot chromosome pseudomolecules (right) are shown in orange if superscaffold is oriented, blue if ambiguous orientation. Connections between superscaffolds are marked by triangles. Superscaffolds were anchored to the linkage groups (left) of the *D. carota* genetic bin map with 918 SNP markers.

Supplementary Figure 31: Distribution of haplotypes along the nine linkage groups from the mapping population 85036 using Checkmatrix. The 84 genotypes are arranged along the x-axis and the loci displayed in linear order along the y-axis. Red indicates parent A alleles; blue indicates parent B 85alleles; yellow indicates heterozygous loci; gray indicates missing data. The first column at the right lists the allele identifier. The second column indicates the number of inconsistent scores. The third column indicates the genetic distance. The proportion of each haplotype and the number of crossovers are below each genotype.

Supplementary Figure 32: Distribution of divergence rates of mobile elements (ME) in carrot DH1. The divergence rate was calculated between the ME elements identified in the genome by homology compared to the consensus sequence in the Repbase (panel A) and in the predicted TE library (Panel B). SINE elements are not included due their small representation among MEs.

Supplementary Figure 33: Phylogenetic tree of *DcSto* families in carrot DH1 based on the genetic distances calculated with the neighbor-joining method (NJ). To evaluate their phylogenetic relationships and the robustness of our classification, *DcSto* families with shared similarity from the analysis with Circoletto were analyzed together.

Supplementary Figure 34: Inter- and intra-specific similarity among families of carrot (*DcSto*), potato (*StSto*), pepper (*CaSto*) and tomato (*SlSto*) *Stowaway*-like elements. Each family is represented by a consensus sequence. Colored ribbons indicate regions of similarity based on blastn results. Colors represent similarity levels (blue<green<orange<red). The diagram was drawn with Circoletto[42].

Supplementary Figure 35: Synteny analysis comparing carrot to grape. **A**: Reconstruction of the grape ancestral gamma whole genome triplication associated with early diversification of the core eudicots. **B**: Distribution of 14,536 syntenic orthologous gene pairs between *Vitis vinifera* (x axis) and *Daucus carota* (y axis) chromosomes.

Supplementary Figure 36: Carrot genome duplications relative to grape. The nine chromosomes of carrot are represented on both x and y axis. Dots represent the positions of paralogous pairs of genes. Genome-wide distribution of the syntenic blocks corresponding to the seven ancestral eudicot protochromosomes. Each dot represents clusters of paralogs syntenic to the grape triplicated blocks. Clusters are painted in seven colors. For example A1 in blue represents triplicated blocks of grape Chr 2, Chr 15, Chr 16.

Supplementary Figure 37: Synteny analysis comparing carrot to kiwi and tomato. A: Distribution of 23,518 syntenic orthologous gene pairs between *Actinidia chinensis* (kiwi) (x axis) and *Daucus carota* (y axis) chromosomes. B: Distribution of 17,446 syntenic orthologous gene pairs between *Solanum lycopersicum* (tomato) (x axis) and *Daucus carota* (y axis) chromosomes.

Supplementary Figure 38: Phylogenetic analysis of the Zinc finger, GRF-type transcription factors. The phylogenetic trees were constructed using the Neighbor-Joining method and a bootstrap test was performed with 1,000 iterations. Bootstrap values over 50% are shown. The scale bars (0.5) show the number of amino acid substitutions per site. A: Phylogenetic tree of all potential GRF genes identified in the *D. carota* genome. B: structure of GRF protein sequences. Lines indicate the amino acid (aa) length of each gene. Blue boxes indicate the conserved GRF (IPR010666) domains.

Supplementary Figure 39: Schematic representation of the 97837 and 70796 pedigrees and derived progenies.

Supplementary Figure 40: Root phenotypes of the two mapping populations used in this study. a) Typical white (left) and yellow (right) phenotype in the 97837 population and b) dark orange (left) and pale orange (right) phenotype in the 70796 population. Box plots demonstrating the distribution of c) lutein in population 97837 and d) total carotenes in population 70796.

Supplementary Figure 41: Photograph of carrot roots from population 70796. Dark orange samples (top left) are associated with high carotenoid accumulation and are recessive at both the *Y* and *Y2* loci (yyy_2y_2).

Supplementary Figure 42: 212 nt indel in the *Y* candidate gene in population 70796. Four samples on left are dark orange (dOr) and four samples on right are pale orange (pOr). Ladder shown is GeneRuler 1 kb Plus DNA ladder from Fermentas. Numbers of the left side indicates the size of each band in nt. Primers (yY-W) flanking the indel can be found in Supplementary Table 70.

Nature Genetics: doi:10.1038/ng.3565

Supplementary Figure 43: Annotation and evolution of the isoprenoid pathway in the carrot genome. (A) Schematic reconstruction of the isoprenoid pathway in carrot. Numbers in parenthesis indicate for each gene the number of annotated homologs/paralogs. Numbers highlighted in red indicates genes that have been retained from WGD. (B) Interspecific phylogenetic analysis of carrot *CCD* and *NCED* genes across multiple genomes. The phylogenetic tree was constructed using the Neighbor-Joining method and a bootstrap test was performed with 1,000 iterations. Bootstrap values over 50% are shown.

Supplementary Figure 44: Phylogenetic analysis of the *1-deoxyxylulose 5-phosphate synthase* (*DXS*). The phylogenetic tree was constructed using the Neighbor-Joining method and a bootstrap test was performed with 1,000 iterations. Bootstrap values over 50% are shown. The scale bars (0.1) show the number of amino acid substitutions per site. Previously characterized *DXS* genes from *A. thaliana* are labeled with AT prefix. The gene DCAR_030576 which was upregulated in both Y and dOr samples clusters with *DXS* group 1.

Supplementary Figure 45: Interspecific phylogenetic analysis and classification of terpene synthase (TPS) genes from *Daucus carota*. The phylogenetic tree shows all potential TPS genes identified in *D. carota* genome and known TPS genes from six other species. TPS subfamilies are indicated on the circumference of the circle. The scale bar (0.5) shows the number of amino acid substitutions per site.

10. List of Supplementary Tables

Supplementary Table 1: Estimation of carrot genome size based on 17 K-mer statistics.

Supplementary Table 2: Carrot genome sequence assembly organized by pseudomolecules (chromosomes).

Supplementary Table 3: Statistics of quality filtered Illumina PE reads mapped to the carrot assembly v2.0.

Supplementary Table 4: Assessment of the gene space coverage using publicly available carrot EST data.

Supplementary Table 5: Assessment of the gene space coverage based upon alignment of the RNA-Seq data obtained from 20 different tissues and treatments, against the carrot assembly v2.0.

Supplementary Table 6: Evaluation of gene space coverage using core eukaryotic gene mapping approach (CEGMA).

Supplementary Table 7: Summary of assembly quality assessment carried out by alignment of paired-end sequences to the carrot genome assembly v2.0.

Supplementary Table 8: Comparison of the carrot genome assembly with other plant genomes assembled with whole genome sequence technologies.

Supplementary Table 9: Organization of repetitive sequences in the carrot DH1 genome.

Nature Genetics: doi:10.1038/ng.3565

Supplementary Table 10: Comparative analysis of the main tandem repeat clusters observed in the carrot genome and other *Daucus* species.

Supplementary Table 11: General statistics of predicted protein-coding genes based on homology, *de novo*, and consolidated final set analysis.

Supplementary Table 12: Summary of carrot gene annotation based on homology or functional classification.

Supplementary Table 13: Summary of evidence for GLEAN gene models for carrot.

Supplementary Table 14: Top 50 over- and under-represented InterPro domains identified comparing the annotations of the carrot genome with lettuce, kiwi, potato, tomato and grape genomes. For each IPR domain percentages and absolute numbers detected in each species are reported. Fields highlighted in green represent IPR domains that are overrepresented in carrot protein families while fields highlighted in red represent underrepresented domains.

Supplementary Table 15: ncRNA (tRNA, rRNA, snRNA and miRNA) genes in the carrot genome.

Supplementary Table 16: Plant materials used for resequencing.

Supplementary Table 17: Summary of SNP analysis of plant materials used for resequencing.

Supplementary Table 18: Diversity levels of resequenced carrots.

Supplementary Table 19: Protein datasets used for gene family analyses.

Supplementary Table 20: Summary of duplicated blocks and retained carrot genes associated with the Dc-α, Dc-β and Dc-γ peaks 4DTv plot.

Supplementary Table 21: Summary of gene depth in the carrot genome at each duplicated block.

Supplementary Table 22: GO categories significantly enriched for the Dc-α WGD retained genes.

Supplementary Table 23: Most abundant (top 50) InterPro (IPR) protein domains of carrot specific gene families identified by ortholog comparative analysis with 13 genomes. Rows highlighted in green represent IPR domains that are over-represented in predicted carrot genes.

Supplementary Table 24: Most abundant (top 50) InterPro (IPR) protein domains of unclustered carrot genes identified by ortholog comparative analysis with 13 genomes. Rows highlighted in green represent IPR domains that are over-represented in predicted carrot genes.

Supplementary Table 25: Regulatory genes (RG including: transcription factors, TF; transcription regulators, TR; chromatin remodeling genes, CR) identified in the carrot genome and 10 other plant genomes. Cells highlighted in green indicated RG families over-represented in the carrot genome, cells highlighted in red indicate RG families under-represented in the carrot genome. Cells highlighted with a thick box border represent those Arabidopsis annotated RG families used to calculate the sensitivity score.

Nature Genetics: doi:10.1038/ng.3565

Supplementary Table 26: Numbers of regulatory genes in 11 plant genomes for major regulating gene families as classified by PlantTFcat database (http://plantgrn.noble.org/PlantTFcat/).

Supplementary Table 27: Summary of duplication modes predicted for each gene and each regulatory gene family based on their physical location, orthologous/paralogous relationships and whole genome duplication history as described in Supplementary note section 4.1. Rows highlighted in green represent RG families found significantly correlated with post-WGD retention by Lang *et al.*[108]. The list of RG families above LisH (blank row 41) represent those families that contain >20 predicted genes.

Supplementary Table 28: Overview of the gene duplication analysis of carrot regulatory gene (RG) classes. Numbers in each column indicate number of genes.

Supplementary Table 29: Disease resistance genes identified in the carrot genome and eight other plant genomes. Cells highlighted in green indicated RG families over-represented in the carrot genome, cells highlighted in red indicate RG families under-represented in the carrot genome.

Supplementary Table 30: Summary of OrthoMCL cluster analysis carried out for four NBS R gene classes. The analysis included predicted R genes from carrot and eight other genomes.

Supplementary Table 31: Summary of OrthoMCL analysis for CNL R genes. OrthoMCL clusters containing carrot CNL R genes were classified as lineage specific (carrot, Euasterids or Asterids) and either expanded or contracted. For each cluster, cells highlighted in green indicate an expansion, cells highlighted in red indicate a contraction for the carrot genome relative to all plants, to Euasterids and to Asterids evaluated in this study.

Supplementary Table 32: Summary of OrthoMCL analysis for NL R genes. OrthoMCL clusters containing carrot NL R genes were classified as lineage specific (carrot, Euasterids or Asterids) and either expanded or contracted. For each cluster, cells highlighted in green indicate an expansion, cells highlighted in red indicate a contraction for the carrot genome relative to all plants, to Euasterids and to Asterids evaluated in this study.

Supplementary Table 33: Summary of OrthoMCL analysis for TNL R genes. OrthoMCL clusters containing carrot TNL R genes were classified as lineage specific (carrot, Euasterids or Asterids) and either expanded or contracted. For each cluster, cells highlighted in red indicate a contraction for the carrot genome relative to all plants, to Euasterids and to Asterids evaluated in this study.

Supplementary Table 34: Summary of OrthoMCL analysis for N R genes. OrthoMCL clusters containing carrot N R genes were classified as lineage specific (carrot, Euasterids or Asterids) and either expanded or contracted. For each cluster, cells highlighted in red indicate a contraction for the carrot genome relative to all plants, to Euasterids and to Asterids evaluated in this study.

Supplementary Table 35: Overview of the gene duplication analysis of carrot R gene classes CNL and NL. Numbers in each column indicate number of genes.

Nature Genetics: doi:10.1038/ng.3565

Supplementary Table 36: Number of predicted R genes organized in clusters or arrays in the carrot genome.

Supplementary Table 37: Organization of predicted R gene clusters identified in the carrot genome.

Supplementary Table 38: Segregation ratios observed in carrot populations derived from family 97837 and 70796.

Supplementary Table 39: Summary of the annotation, polymorphisms (SNPs and indels), and comparative transcriptome analysis of predicted genes located in the genomic region associated with carotenoid accumulation in the 97837 and 70796 populations. In population 97837 the region of interest spans from gene predictions DCAR_017884 to DCAR_017894. In population 70796 the region of interest spans from gene prediction DCAR_017887 to DCAR_017902. Cells highlighted in green represent predicted genes located in the 75kb overlapping region controlling the carotenoid accumulation *Y* locus in both populations. Annotations for each predicted gene were retro-viewed from the carrot genome annotation. Polymorphism detection analysis between high pigmented types, dark orange (dOr) and yellow (Y), versus low pigmented types, pale orange (pOr) and white (W), are reported in columns J and K. Results of the differential expression analysis comparing dark orange with pale orange samples are reported in columns M-Z. Results of the differential expression analysis comparing white (W) and yellow (Y) samples are reported in columns AB-AO.

Supplementary Table 40: Polymorphisms (SNPs and indels) detected in the genomic region covering the haplotype block associated with high carotenoid phenotypes in the reseqeunced samples.

Supplementary Table 41: List of carrot genes co-expressed with DCAR_032551.

Supplementary Table 42: GO overrepresentation test carried out using AGRIGO.

Supplementary Table 43: GO overrepresentation test carried out using PANTHER.

Supplementary Table 44: Annotation of carrot genes upregulated in both yellow and dark orange (*yy*) storage roots of plants from mapping populations. Highlighted DCAR_032551 is the only gene both upregulated and located in the mapped region of the *Y* gene (Supplementary Table 38.

Supplementary Table 45: Annotation of carrot genes downregulated in yellow or dark orange (*yy*) storage roots of plants from mapping populations.

Supplementary Table 46: Summary of carrot candidate genes involved in the plastid 2-C-methyl-D-erythritol 4-phosphate (MEP) and carotenoid pathways.

Supplementary Table 47: Summary of TPS candidate genes identified in the carrot V2.0 genome annotation.

Supplementary Table 48: Possible gene interactions of AT3G55240, the *Y* candidate gene homolog, in Arabidopsis using two-hybrid protein-protein screening array.

Nature Genetics: doi:10.1038/ng.3565

Supplementary Table 49: Statistics of raw data generated by sequencing the carrot DH1 genomic DNA and the BAC clone ends.

Supplementary Table 50: Carrot sequencing statistics after filtering.

Supplementary Table 51: Statistics of the carrot genome assembly, phase I.

Supplementary Table 52: Statistics of data used to construct the integrated carrot linkage maps.

Supplementary Table 53: Summary statistics of the full map and bin integrated carrot linkage maps.

Supplementary Table 54: Spearman's rank correlation (r) of the marker order for the comparison among the integrated *D. carota* linkage map and three population-specific maps, 70349, 70796 and Br1091×HM1. The test was carried out with data from all three maps, including redundant markers.

Supplementary Table 55: Statistics of the carrot genome assembly, phase II.

Supplementary Table 56: Subtelomeric BAC clones DNA used for FISH experiments.

Supplementary Table 57: Characteristics of carrot DH1 *Tourist*-like and *Stowaway*-like miniature inverted repeat transposable elements (MITEs).

Supplementary Table 58: Copy numbers of *Stowaway*-like MITEs in Asterid genome species.

Supplementary Table 59: Copy numbers and within-family similarity of potato, pepper and tomato *Stowaway*-like MITEs.

Supplementary Table 60: Cluster ID, length and consensus sequences of clusters containing tandem repetitive sequences identified in DH1 genome and in other D*aucus* species resequencing data using RepeatExplorer (RE) analysis (see section 2.1.3 for details).

Supplementary Table 61: Pairwise percent similarity among Cent-Dc monomers. Numbers in parentheses indicate the number of sequences used to carry out this analysis. DcarK11 monomer sequences were extracted from a Sanger sequenced plasmid K11 containing the putative centromeric repeat of carrot (Cent-Dc)(See section 1.4.2.3). BES monomers were extracted from 17 DH1 BACs containing the Cent-Dc motif. DH1, Dsyr and Daur Cent-Dc monomers were extracted from whole genome sequences produced in this study.

Supplementary Table 62: Summary of gene ortholog analysis conducted on 14 sequenced genomes.

Supplementary Table 63: Summary of syntenic blocks detected comparing the carrot genome with itself and with the grape, tomato and kiwi genomes**.**

Supplementary Table 64: Time estimates of carrot (Dc) and *S. lycopersicum* (Sl) WGD and species divergence.

Nature Genetics: doi:10.1038/ng.3565

Supplementary Table 65: Estimates of carrot genome paralogous block depth under the Dc-α, Dc-β and Dc-γ peaks. Paralogous pairs with 4DTv values ranging from 0.2-0.3 were used to estimate the depth of the Dc-α blocks; paralogous pairs with 4DTv values ranging from 0.3-0.6 were used to estimate the depth of the Dc-β blocks; paralogous pairs with 4DTv values ranging from 0.6-1 were used to estimate the depth of the Dc-γ blocks; Numbers in parentheses indicate the number of genes in duplicated blocks.

Supplementary Table 66: Summary of OrthoMCL analysis carried out for eight RG families with carrot and 10 other genomes.

Supplementary Table 67: Characterization of the evolutionary history of the GRF RG family in the carrot genome.

Supplementary Table 68: Summary of OrthoMCL analysis for GRF Regulatory genes (RG). OrthoMCL clusters containing carrot GRF genes were classified as lineage specific (carrot, Euasterids or Asterids) and either expanded or contracted. For each cluster, cells highlighted in green indicate an expansion for carrot, Euasterids and Asterids relative to other plant genomes used in this study.

Supplementary Table 69: Characterization of the evolutionary history of the JmjC RG family in the carrot genome.

Supplementary Table 70: Summary of OrthoMCL analysis for JmjC Regulatory genes (RG). OrthoMCL clusters containing carrot JmjC genes were classified as lineage specific (carrot, Euasterids or Asterids) and either expanded or contracted. For each cluster, cells highlighted in green indicate an expansion, cells highlighted in red indicate a contraction for carrot, Euasterids and Asterids relative to other plant genomes used in this study.

Supplementary Table 71: Characterization of the evolutionary history of the TCP RG family in the carrot genome.

Supplementary Table 72: Summary of OrthoMCL analysis for TCP Regulatory genes (RG). OrthoMCL clusters containing carrot TCP genes were classified as lineage specific (carrot, Euasterids or Asterids) and either expanded or contracted. For each cluster, cells highlighted in green indicate an expansion, cells highlighted in red indicate a contraction for the carrot genome relative to all plants, to Euasterids and to Asterids evaluated in this study.

Supplementary Table 73: Characterization of the evolutionary history of the GeBP RG family in the carrot genome.

Supplementary Table 74: Summary of OrthoMCL analysis for GeBP Regulatory genes (RG). OrthoMCL clusters containing carrot GeBP genes were classified as lineage specific (carrot, Euasterids or Asterids) and either expanded or contracted. For each cluster, cells highlighted in green indicate an expansion, cells highlighted in red indicate a contraction for the carrot genome relative to all plants, to Euasterids and to Asterids evaluated in this study.

Supplementary Table 75: Characterization of the evolutionary history of the response regulator (RR) RG family in the carrot genome.

Nature Genetics: doi:10.1038/ng.3565

Supplementary Table 76: Summary of OrthoMCL analysis for RR Regulatory genes (RG). OrthoMCL clusters containing carrot RR genes were classified as lineage specific (carrot, Euasterids or Asterids) and either expanded or contracted. For each cluster, cells highlighted in green indicate an expansion, cells highlighted in red indicate a contraction for the carrot genome relative to all plants, to Euasterids and to Asterids evaluated in this study.

Supplementary Table 77: Characterization of the evolutionary history of the ARF RG family in the carrot genome.

Supplementary Table 78: Characterization of the evolutionary history of the B3-domain/RAV (B3/RAV) RG family in the carrot genome.

Supplementary Table 79: Summary of OrthoMCL analysis for B3/RAV Regulatory genes (RG). OrthoMCL clusters containing carrot B3/RAV genes were classified as lineage specific (carrot, Euasterids or Asterids) and either expanded or contracted. For each cluster, cells highlighted in green indicate an expansion, cells highlighted in red indicate a contraction for the carrot genome relative to all plants, to Euasterids and to Asterids evaluated in this study.

Supplementary Table 80: Primers used for carrot fine-mapping and indel detection.

Supplementary Table 81: Marker distribution and density in the 97837 population.

Supplementary Table 82: Marker distribution and density in the 70796 population.

Supplementary Table 83: Summary of QTL for total carotenoids in the population 70796.

Supplementary Table 84: Carrot predicted peptides annotated as candidate genes involved in the flavonoid and anthocyanin biosynthetic pathways.

Supplementary Table 85: Summary of isoprenoid pathway genes across 11 sequenced genomes.

Nature Genetics: doi:10.1038/ng.3565