SUPPLEMENTARY INFORMATION

doi:10.1038/nature21370

1. Reference genome sequencing and assembly

1.1. Plant material

We sequenced *Chenopodium quinoa* Willd. (quinoa) accession PI 614886 (also known as NSL 106399 and QQ74), which was originally collected in Chile and belongs to the Coastal ecotype. Unless noted otherwise, all analyses reported herein were performed with accession PI 614886. This accession is publicly available from the Germplasm Resources Information Network (GRIN; http://www.ars-grin.gov/index.html) of the United States Department of Agriculture (USDA) Agricultural Research Service (ARS) and has been included in previous published assessments of genetic diversity of coastal and highland quinoas[80], in which it clustered with other coastal varieties.

1.2. PacBio sequencing and assembly

DNA was extracted from leaf and flower tissue of a single soil-grown plant that had been placed in the dark 48 h before tissue harvest. DNA was prepared as described in the "Preparing *Arabidopsis* Genomic DNA for Size-Selected ~20 kb SMRTbell™ Libraries" protocol (http://www.pacb.com/wp-content/uploads/2015/09/Shared-Protocol-Preparing-Arabidopsis-DNA-for-20-kb-SMRTbell-Libraries.pdf). DNA was purified twice with Beckman Coulter Genomics AMPure XP magnetic beads and assessed by standard agarose gel electrophoresis and Thermo Fisher Scientific Qubit Fluorometry. 100 Single-Molecule Real-Time (SMRT) cells were run on the PacBio RS II system with the P6-C4 chemistry by DNALink (Seoul, Republic of Korea). A total of 6,037,280 PacBio post-filtered reads was generated from the 100 SMRT cells. This produced a total of 75,132,015,080 bp of single-molecule sequencing data, with an average read length of 12,444 bp. *De novo* assembly was conducted using the smrtmake assembly pipeline (https://github.com/PacificBiosciences/smrtmake) with the setting GENOME_SIZE = 750,000,000. Smrtmake starts with the filtering step for the SMRT reads (Options --filter='MinReadScore=0.80,MinSRL=500,MinRL=100') and then performs an error correction (CUTOFF option setting with GENOME_SIZE 750Mb*30). In the next step, the Celera Assembler generates the draft assembly using the error-corrected reads. The draft assembly was then polished for the final assembly using the quiver algorithm.

1.3. Integrating BioNano optical maps with the PacBio assembly

Quinoa plants were grown in soil for three weeks in a greenhouse at Brigham Young University (Provo, UT, USA) and then placed in the dark for two days. High molecular weight DNA was isolated and labelled from young leaf tissue according to standard BioNano protocols. Specifically, DNA was digested by the single-stranded nicking endonuclease Nt.BspQI and labelled with a fluorescent-dUTP nucleotide analogue using *Taq* polymerase. Nicks were ligated with *Taq* DNA ligase and the backbone of the labelled DNA was stained using the intercalating dye YOYO-1. Labelled DNA was imaged automatically using the BioNano Irys system and *de novo* assembled into consensus physical maps using the BioNano IrysView analysis software. The final *de novo* assembly used only single molecules with a minimum length of 150 kb and eight labels per molecule. The p-values for the initial assembly, extension of the assembly, and chimera detection were set to 10^{-8}, 10^{-9}, and 10^{-15}, respectively. Hybrid scaffolds were identified using IrysView's hybrid scaffold alignment subprogram using a p-value of 10^{-8} for initial and final alignment and 10^{-13} for chimera detection and merging.

1.4. Chicago library preparation and sequencing

Using the same DNA prepared for PacBio sequencing, a Chicago library was prepared as described previously[10]. Briefly, 500 ng of high molecular weight genomic DNA (mean fragment size ~100 kb) was reconstituted into chromatin *in vitro* and fixed with formaldehyde. Fixed chromatin was then digested with *Dpn*II, the 5' overhangs were filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed, and the DNA was purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was sheared to ~350 bp mean fragment size, and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were then isolated using streptavidin beads before PCR enrichment of the library. The library was sequenced on an Illumina HiSeq 2500 in rapid run mode to produce 122 million 2X100 bp read pairs, providing 51.6X physical coverage (1-50 kb pairs).

1.5. Scaffolding the PacBio and BioNano assemblies with HiRise

Chicago sequence data (in FASTQ format) was used to scaffold various quinoa input assemblies using HiRise, a software pipeline designed specifically for using Chicago data to assemble genomes[10]. Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analysed by HiRise to produce a likelihood model, and the resulting likelihood model was used to identify putative mis-joins and score prospective joins.

1.6. Short-read Illumina sequencing

DNA was extracted from leaf tissue of a single soil-grown plant using the Qiagen DNeasy Plant Mini Kit. 500-bp paired-end (PE) libraries were prepared using the NEBNext Ultra DNA Library Prep Kit for Illumina. Sequencing was performed using an Illumina HiSeq 2000 machine at King Abdullah University of Science and Technology (KAUST; Thuwal, Saudi Arabia). Reads were processed with Trimmomatic (v0.33)[42] to remove adapter sequences and leading and trailing bases with a quality score below 20 and reads with an average per base quality of 20 over a 4-bp sliding window. Reads < 75 nucleotides in length after trimming were removed from further analysis, and the remaining high-quality reads were assembled with Velvet (v1.2.10)[43] using a k-mer of 75. The assembly contained 838,071,669 bp in 1,040,940 contigs with an N50 of 2,175 bp (in 95,338 contigs), with the largest contig being 45,571 bp.

1.7. Validating the reference genome assembly

The assembly was validated using the Illumina short-reads described above (66X coverage) and bacterial artificial chromosome end sequences (BESs)[81] data. Of the 851,664,032 Illumina short reads generated, 99.4% of the reads were successfully mapped back to the final assembly, with 97.7% properly paired. The short-read assembly was compared to the reference assembly using BLASTN. When considering only the top two BLASTN matches for each of the short-read contigs, a total of 659,305,355 bp (47%) of the quinoa final assembly was covered by the short-read assembly. Given the high repetitive content of the quinoa genome and the limitations of short-read sequencing, it is very likely that repetitive regions in the Illumina assembly would have collapsed in the assembly. Thus, it is not unexpected that such a BLAST search of the short-read assembly back onto the reference assembly would yield a much lower overall coverage.

To extend the validation further, we allowed for multiple BLASTN hits from the short-read assembly onto the reference assembly and filtered for BLASTN hits with an E-value < $5x10^{-4}$. This increased the total bases covered to 1,203,491,061 bp, which represents 86.6% of the total quinoa genome. To further validate the genome, a total of 2,106 BESs were aligned to the reference quinoa genome assembly using BLASTN. After filtering for hits with E-values < $1x10^{-100}$, and insert sizes that were too large (> 370 kb) or too small (< 10 kb), 109 scaffolds with a total of 286 Mb (representing 22% of the genome) could be validated by the BES data.

2. Reference genome annotation

2.1. Plant growth and RNA extraction

RNA was extracted from the following greenhouse-grown samples: whole young plants (with 6-8 true leaves) grown in soil; roots, leaves and petioles, apical meristems, lateral meristems, stems, and flowers and immature seeds from mature plants grown in soil. RNA was also isolated from roots and shoots of soil-grown plants in control conditions or exposed to heat or drought, and hydroponically-grown plants in control conditions or exposed to low phosphate. For the soil treatments, plants were grown in well-watered conditions in a growth chamber for three weeks at 20°C and 12 h daily light. Plants were then either left in these conditions with (control) or without (drought) water, or were transferred to a second growth chamber with conditions of 12 h light at 37°C and 12 h dark at 32°C (heat). After one additional week, roots and shoots were harvested separately for all plants and snap-frozen in liquid nitrogen. The hydroponic growth system was based on Conn *et al*.[82] Briefly, seeds were sown on germination medium containing 0.7% agar and grown for two weeks in tanks containing basal nutrient solution (BNS), after which plants were transferred to larger, aerated tanks containing BNS. After one additional week of growth, plants were either transferred to tanks containing fresh BNS (control) or tanks containing fresh BNS lacking KH_2PO_4 and supplemented with a compensatory amount of KCl (low phosphate). One week after all treatments began, roots and shoots were harvested separately for all plants and snap-frozen in liquid nitrogen. Frozen tissue from all samples was ground using either a mortar and pestle or a Spex Geno/Grinder, and RNA was isolated using the Zymo Direct-zol RNA MiniPrep Kit. RNA quality was assessed using an Agilent 2100 BioAnalyzer.

2.2. Illumina RNA-Seq

Sequencing libraries were prepared using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina. 100-bp PE sequencing was performed using an Illumina HiSeq 2000 machine at KAUST.

2.3. PacBio Iso-Seq

Sub-samples of root and shoot RNA from plants grown in hydroponics under control conditions (described in section 2.1. above) were purified with Beckman Coulter Genomics AMPure XP magnetic beads and assessed with standard agarose gel electrophoresis and Thermo Fisher Scientific Qubit Fluorometry. RNA was fractionated into three libraries consisting of differently sized RNA (1 - 2 kb, 2 - 3 kb and 3 - 6 kb). A total of 9 SMRT cells for each RNA sample was run on the PacBio RS II system with the P6-C4 chemistry by DNALink. A total of 836,322 reads covering 2,292,247,217 bp and 699,876 reads covering 1,696,155,060 bp was produced for the shoot and root RNA libraries, respectively. Sequencing reads were processed with the RS_IsoSeq protocol of SMRT Analysis (v2.2), and polished consensus sequences were produced with the ToFU pipeline[83].

2.4. Characterisation of repetitive sequences

Repeat families found in the genome assemblies of quinoa, *C. pallidicaule*, and *C. suecicum* (see Supplementary Information 3.) were first independently identified *de novo* and classified using the software package RepeatModeler[49]. RepeatModeler depends on the programs RECON and RepeatScout for the *de novo* identification of repeats within the genome. After the classification process, the output data file from RepeatModeler for each of the genome assemblies was used as a custom repeat library by the program RepeatMasker[50] to discover and identify repeats within the respective genomes. The results of repeat classification are summarized in Supplementary Table 1.

2.5. *Ab initio* gene model prediction

AUGUSTUS[51] was used as the main *ab initio* prediction software for the genomes. First, coding sequences from *Amaranthus hypochondriacus, Beta vulgaris, Spinacia oleracea* and *Arabidopsis thaliana* (102,149 in total) were concatenated to create a master list of genes. Fifty percent of the genes from this master list were used to train the AUGUSTUS model, and the remaining genes were used for validation purposes. Two rounds of prediction optimisation were done with the software package provided by AUGUSTUS. Next, RNA-Seq reads from different tissues and abiotic stress, and full-length transcripts generated from Iso-Seq were mapped onto the reference genome using Bowtie 2[52] and GMAP[53], respectively. Hints with locations of potential intron-exon boundaries were generated from the alignment files with the software package BAM2hints in the MAKER package[54]. MAKER with AUGUSTUS (intron-exon boundary hints provided from RNA-Seq and Iso-Seq) was then used to predict genes in the repeat-masked reference genome. To help guide the prediction process, peptide sequences from *B. vulgaris* and the original quinoa full-length transcript (provided as EST evidence) were used by MAKER during the prediction. To help assess the quality of the gene prediction, AED scores were generated for each of the predicted 44,776 genes as part of the MAKER pipeline. Genes were further characterised for their putative function by performing a BLAST search of the peptide sequences against the UniProt database. PFAM domains and InterProScan ID were added to the gene models using the scripts provided in the MAKER package. Results from the annotation process are summarized in Supplementary Table 2. Gene density and GC content were plotted using DensityMap[84].

Supplementary Table 1. Repeat classification in quinoa, C. pallidicaule, and C. suecicum.

Repeat Class Description			*C. quinoa* coverage		*C. pallidicaule* coverage		*C. suecicum* coverage	
Repeat Type	Order	Superfamily	bp	%	bp	%	bp	%
Class I TEs	All	All	619,249,241	45.26	91,893,102	27.27	161,750,700	30.09
	LTR	All	502,546,481	36.84	86,430,495	25.65	146,937,459	27.34
		Cassandra	278,710	0.02	41,007	0.01	90,116	0.02
		Caulimovirus	317,559	0.02	247,300	0.07	611,789	0.11
		Copia	113,928,967	8.22	23,612,635	7.01	22,611,546	4.21
		ERV1	323,293	0.02	236,306	0.07	694,800	0.13
		ERVK	0	0.00	0	0.00	90,940	0.02
		Gypsy	394,369,811	28.46	62,911,943	18.67	125,016,497	23.26
		Ngaro	109,720	0.01	0	0.00	0	0.00
	LINE	All	116,487,771	8.41	5,394,489	1.60	14,744,198	2.74
		CR1	0	0.00	42,355	0.01	79,503	0.01
		CRE	227,331	0.02	0	0.00	0	0.00
		CRE-II	2,940,296	0.21	1,084,086	0.32	948,796	0.18
		Jockey	2,660,398	0.19	142,053	0.04	524,260	0.10
		L1	97,249,157	7.02	2,826,654	0.84	9,853,874	1.83
		L1-Tx1	0	0.00	120,050	0.04	681,275	0.13
		L2	0	0.00	160,450	0.05	1,344,056	0.25
		R1	1,366,328	0.10	324,814	0.10	1,178,447	0.22
		RTE-BovB	1,341,413	0.10	750,095	0.22	568,510	0.11
		Penelope	8,810,321	0.64	0	0.00	0	0.00
		Tad1	4,644,879	0.34	0	0.00	0	0.00
	SINE	tRNA	214,989	0.01	68,118	0.02	69,043	0.01
Class II TEs	All	All	86,409,760	6.24	25,803,170	7.66	20,180,396	3.76
	TIR	CMC-EnSpm	26,309,462	1.90	5,365,500	1.59	7,255,775	1.35
		Dada	0	0.00	0	0.00	131,709	0.02
		hAT-Ac	8,512,388	0.61	4,565,652	1.35	3,060,223	0.57
		hAT-Tag1	3,264,709	0.24	270,251	0.08	425,005	0.08
		hAT-Tip100	1,126,378	0.08	397,681	0.12	484,464	0.09
		MuLE-MuDR	28,752,390	2.08	10,748,226	3.19	6,234,758	1.16
		TcMar-Mogwai	670,729	0.05	597,692	0.18	0	0.00
		PIF-Harbinger	2,084,193	0.15	1,094,376	0.32	598,704	0.11
		TcMar-Stowaway	6,870,731	0.50	2,806,151	0.83	2,488,801	0.46
		hAT-Charlie	184,042	0.01	0	0.00	0	0.00
		Sola	5,801,619	0.42	0	0.00	0	0.00
	Crypton	Crypton	422,446	0.03	0	0.00	0	0.00
	Maverick	Maverick	2,966,516	0.21	186,214	0.06	0	0.00
Low complexity			2,783,502	0.20	793,813	0.24	1,214,711	0.23
Simple repeat			25,128,515	1.81	6,234,759	1.85	12,046,245	2.24
snRNA			179,784	0.01	0	0.00	11,504	0.00
Unclassified			145,488,187	10.50	46,122,498	13.69	97,402,803	18.12

Supplementary Table 2. Annotation statistics for quinoa, C. pallidicaule, and C. suecicum.

	C. quinoa	*C. pallidicaule*	*C. suecicum*
Total number of genes	44,776	17,961	21,861
Total coding region (bp)	57,064,233	23,414,073	28,057,137
Average length of genes (bp)	1,274	1,303	1,283
Largest gene (bp)	15,933	15,411	16,194
Number of single-exon genes	6,864	2,665	3,384

2.6. Annotation validation with BUSCO

Genome assembly and annotation completeness was assessed using the plantae database of 956 single copy orthologs using BUSCO[18] with the BLAST E-value cutoff set to 10^{-5} (Supplementary Table 3).

Supplementary Table 3. BUSCO analysis of genome annotations from quinoa, C. pallidicaule, and C. suecicum.

	C. quinoa	*C. pallidicaule*	*C. suecicum*
Complete single-copy BUSCOs	906 (94.8 %)	886 (92.7 %)	871 (91.1 %)
Complete duplicated BUSCOs	834 (87.2 %)	315 (32.9 %)	313 (32.7 %)
Fragmented BUSCOs	24 (2.5 %)	34 (3.6 %)	38 (4.0 %)
Missing BUSCOs	26 (2.7 %)	36 (3.8 %)	47 (4.9 %)
Total BUSCO groups searched	956	956	956

2.7. Small RNA isolation and Illumina sequencing

Small RNAs were isolated from the total RNA extracted from the hydroponics control and low phosphate samples described above. Small RNA libraries were prepared using the Illumina TruSeq Small RNA Library Prep Kit, and sequencing was performed with an Illumina HiSeq 2000 machine. The sequenced reads were processed using trim galore v0.4.0 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), trimming small RNA adaptor sequences used for the small RNA preparation and low-quality reads. We selected all sequences between 19 and 26 nucleotides in length, resulting in 102,858,577 reads for small RNA analysis. Briefly, we used the pipeline as provided by ShortStack[85], using default parameters except for the 'foldsize=1000'. We detected 523,752 loci with clustered mappings of small RNAs across the genomes, of which 483,702 are likely to be from RNAi mediated processing of small RNAs. By the stringent miRNA detection algorithm implemented in ShortStack, we detected 67 candidate miRNA with canonical secondary hairpin structures. ShortStack tends to be very stringent in calling *de novo* microRNAs, keeping false positive rates low at the expense of high false negative rates. Indeed, during benchmarking and internal optimisation of the microRNA detection protocols using *A. thaliana* small RNA sequencing data, we noticed that only 17 out of 165 known microRNAs are detected by the default ShortStack algorithm (data not shown). Specificity was high, as the 17 were out of 20 detected by ShortStack (thus these are called high confidence miRNA candidates).

To increase sensitivity, we relaxed the stringency to create intermediate confidence candidates. Briefly, we selected all candidate loci which satisfy all criteria of microRNA loci except for detecting sequenced

miRNA star ('N15' category from ShortStack detection algorithm). Exploratory analyses showed enrichment of known microRNAs in highly expressed loci of this category (data not shown). When we selected those with high expression levels (number of mapped reads being higher than 25% of the high confidence miRNA candidates), in *A. thaliana*, we detected 68 putative loci (20 high confidence and 48 intermediate confidence), which overlaps with 45 out of 165 known microRNAs. Using these same criteria, we detected 67 high and 204 intermediate confidence candidate microRNAs in the quinoa genome. We matched the identified miRNA stem-loop precursor to known miRNA families using Rfam scan (http://rfam.xfam.org, Rfam 12.1, April 2016). For example, we found multiple putative miRNA loci that likely belong to a family of highly conserved, highly expressed mir166 (Rfam ID RF00075). The identified miRNAs also share sequence homology with known miRNAs in the miRBase database (http://www.mirbase.org) (Supplementary Table 4).

3. Sequencing and assembly of *C. pallidicaule* and *C. suecicum*

3.1. Plant material

The diploid species *C. pallidicaule* (PI 478407) and *C. suecicum* (BYU 1480) were chosen as representatives of the A and B sub-genomes of quinoa, respectively, according to published phylogenetic inferences[22].

3.2. Short-read Illumina sequencing and assembly

DNA was extracted from each diploid species and sent to the Beijing Genomic Institute (BGI, Hong Kong, China) where one 180-bp PE library and two mate-pair libraries with insert sizes of 3 and 6 kb were prepared and sequenced on the Illumina HiSeq platform to obtain 2 X 100-bp reads for each library. The generated reads were trimmed using the quality-based trimming tool Sickle (https://github.com/najoshi/sickle) with a quality PHRED score cutoff of 20. The trimmed reads were then assembled using the ALLPATHS-LG assembler[47] using the recommended default parameters, and genome size was estimated using a k-mer analysis as part of the ALLPATHS assembly process. GapCloser v1.12, a subtool for SOAPdenovo2[48] (Short Oligonucleotide Analysis Package), was used to resolve N spacers and gap lengths produced by the ALLPATHS-LG assembler. The GapCloser-corrected assembly is hereafter referred to as the *C. pallidicaule* or *C. suecicum* genome assembly.

3.3. Illumina RNA-Seq

Transcriptomes for each diploid species were developed using tissue samples collected from 21-28 day old hydroponically grown plants. For *C. pallidicaule,* tissue samples included leaf, root, immature flower buds and apical meristem tips grown in standard hydroponic media as well as leaf and root samples grown in hydroponic media supplement with 300 mM NaCl. For *C. suecicum*, tissue samples included leaf, root, stem and whole inflorescence grown in standard hydroponic media. Tissue samples were flash frozen in liquid nitrogen and sent to BGI where

Supplementary Table 4. miRNA family classification based on Rfam scan.

Target	Name	ClusterID	Position (start)	Position (stop)	Strand	GC freq.	Score	E-value
mir-160	RF00247	Cluster_181048	91	174	+	0.54	58.1	4.70E-13
mir-160	RF00247	Cluster_181048	174	91	-	0.54	51.7	2.90E-11
mir-160	RF00247	Cluster_511922	1	85	+	0.54	64.6	3.50E-15
mir-166	RF00075	Cluster_157797	1	95	+	0.58	65.2	1.20E-16
mir-166	RF00075	Cluster_177575	90	176	+	0.45	82.7	2.10E-21
mir-166	RF00075	Cluster_384456	217	295	+	0.52	71.3	5.30E-18
mir-166	RF00075	Cluster_493553	38	116	+	0.58	74.6	2.10E-19
mir-166	RF00075	Cluster_519369	1	80	+	0.51	72.1	1.10E-18
mir-172	RF00452	Cluster_277599	12	132	+	0.31	90.7	3.70E-22
mir-172	RF00452	Cluster_277599	132	12	-	0.31	79.2	5.50E-19
mir-172	RF00452	Cluster_388226	343	467	+	0.46	70.3	5.80E-16
mir-172	RF00452	Cluster_508069	11	135	+	0.48	70.6	1.80E-16
mir-393	RF02516	Cluster_169733	42	141	+	0.35	89.3	6.30E-22
mir-393	RF02516	Cluster_169733	142	43	-	0.35	62.3	2.20E-14
mir-399	RF00445	Cluster_19699	98	6	-	0.47	60.8	6.90E-16
mir-399	RF00445	Cluster_19699	6	98	+	0.47	58.5	3.50E-15
mir-399	RF00445	Cluster_32841	1	120	+	0.52	60.4	7.20E-16
mir-399	RF00445	Cluster_32847	14	110	+	0.42	60.8	4.90E-16
mir-399	RF00445	Cluster_32847	110	14	-	0.42	57.2	6.60E-15
mir-399	RF00445	Cluster_506691	28	123	+	0.45	64.6	3.40E-17
mir-399	RF00445	Cluster_506709	16	107	+	0.51	62.5	6.60E-16
MIR162_2	RF00742	Cluster_113602	1	87	+	0.48	68.2	2.50E-15
MIR162_2	RF00742	Cluster_37750	1	87	+	0.48	68.2	2.20E-15
MIR167_1	RF00640	Cluster_177944	1	123	+	0.36	57.7	4.40E-13
MIR171_1	RF00643	Cluster_108503	2	88	+	0.45	54.0	1.60E-10
MIR171_1	RF00643	Cluster_221895	1	101	+	0.41	58.0	5.40E-11
MIR171_1	RF00643	Cluster_270920	1	94	+	0.41	61.7	2.10E-12
MIR171_1	RF00643	Cluster_400983	1	148	+	0.30	54.2	4.10E-10
MIR171_1	RF00643	Cluster_436398	1	87	+	0.45	54.0	1.40E-10
MIR171_1	RF00643	Cluster_79082	18	175	+	0.29	62.8	3.80E-12
MIR171_1	RF00643	Cluster_93139	1	100	+	0.40	59.8	6.00E-12
MIR390	RF00689	Cluster_246008	7	192	+	0.31	77.9	2.00E-18
MIR396	RF00648	Cluster_108449	38	250	+	0.42	83.9	1.10E-19
MIR396	RF00648	Cluster_436440	3	115	+	0.32	83.7	8.40E-20
MIR397	RF00704	Cluster_269218	1	114	+	0.33	57.7	1.60E-11
MIR398	RF00695	Cluster_261303	1	103	+	0.36	67.1	1.80E-15
MIR398	RF00695	Cluster_8943	1	102	+	0.35	65.9	6.90E-15
MIR408	RF00690	Cluster_383633	179	292	+	0.49	68.6	9.60E-18
MIR535	RF00714	Cluster_124189	1	114	+	0.41	58.6	2.20E-14

200-bp short-insert libraries were prepared and sequenced using the HiSeq platform to obtain 2 X 100-bp reads. Using default parameters with the minimum k-mer coverage set to five and graph bubble

popping set to > 0.9, ABySS v1.3.7[86] was used to assemble the *de novo* transcriptome for each diploid species using 11 unique k-mers between k = 40 and k = 90 in increments of 5. Unitigs from all k-mer assemblies were combined and redundancies were removed using CD-HIT-EST[87] with a clustering threshold of 0.98 identity. CAP3[88] and ABySS were then used to identify overlaps (> 100 bp) and scaffold unitigs. GapCloser v1.12 was used to fill gaps created during the scaffolding process. Redundant sequences were again removed using CD-HIT-EST and, in an attempt to remove incomplete sequences, the consensus scaffolds were filtered at a minimum length of 200 bp to produce the final set of scaffolds.

3.4. Repeat characterisation and *ab initio* prediction

The gene models for *C. pallidicaule* and *C. suecicum* were predicted as described above for quinoa. As full-length transcript sequences were not available for either species, gene model predictions were only supported by intron-exon information from the RNA-Seq sequences. The same trained genome model from quinoa was used for prediction, and genes were similarly functionally characterised by UniProt, PFAM domains and InterPro. Annotation results are summarized in Supplementary Table 2.

4. Re-sequencing of additional species and accessions

4.1. Plant material

The following quinoa accessions were chosen for DNA re-sequencing (Supplementary File 5): 0654, Ollague, Real, Pasankalla (BYU 1202), Kurmi, CICA-17, Regalona (BYU 947), Salcedo INIA, G-205-95DK, Cherry Vanilla (BYU 1439), Chucapaca, Ku-2, PI 634921 (Ames 22157), Atlas, and Carina Red. The following accessions of *C. berlandieri* were sequenced: var. *boscianum* (BYU 937), var. *macrocalycium* (BYU 803), var. *zschackei* (BYU 1314), var. *sinuatum* (BYU 14108), and subsp. *nuttaliae* (“Huauzontle”). Two accessions of *C. hircinum* (BYU 566 and BYU 1101) were also sequenced.

4.2. Illumina sequencing

Sequencing of Atlas and Carina Red is described below (Supplementary Information 7.2.2.). For all other accessions, DNA was extracted from leaf tissue of soil-grown plants using the Qiagen DNeasy Plant Mini Kit. 100-bp PE libraries were prepared using the NEBNext Ultra DNA Library Prep Kit for Illumina. Sequencing was performed using an Illumina HiSeq 2000 machine at KAUST. Reads were trimmed using Trimmomatic, as described above. Trimmed reads were mapped to the reference assembly using BWA (v0.7.10)[55]. Read alignments were manipulated with SAMtools (v1.1)[45], and the mpileup function of SAMtools was used to call SNPs. Custom Perl scripts were used to filter SNPs for a depth of at least 10 and a SNP allele frequency of > 75%. Summary statistics for re-sequencing lines are shown in Supplementary Table 5.

Supplementary Table 5. Re-sequencing statistics.

Line	Reads generated	Total bp generated	Trimmed reads	Reads mapped (%)	Number of SNPs
C. quinoa O654	125,379,816	12,663,361,416	106,939,716	95.07	396,768
C. quinoa Ollague	108,722,734	10,980,996,134	90,375,536	98.55	521,026
C. quinoa Real	105,553,640	10,660,917,640	85,524,948	98.47	391,498
C. quinoa Pasankalla	104,838,972	10,588,736,172	86,746,706	98.47	553,923
C. hircinum BYU 566	81,941,702	8,276,111,902	73,065,872	98.75	223,771
C. quinoa Kurmi	85,774,884	8,663,263,284	77,619,466	98.60	351,531
C. berlandieri var. *boscianum*	72,500,738	7,322,574,538	63,472,948	92.48	323,526
C. berlandieri var. *macrocalycium*	77,883,792	7,866,262,992	68,769,534	92.78	429,776
C. quinoa CICA-17	100,375,172	10,137,892,372	85,007,202	98.61	415,672
C. quinoa Regalona	95,168,346	9,612,002,946	82,902,596	99.32	116,415
C. quinoa Salcedo INIA	88,237,122	8,911,949,322	74,782,454	98.47	269,310
C. quinoa G-205-95DK	111,375,902	11,248,966,102	96,342,744	98.96	343,258
C. hircinum BYU 1101	91,129,968	9,204,126,768	79,980,668	94.92	673,515
C. berlandieri var. *zschackei* (BYU 1314)	85,734,472	8,659,181,672	75,378,462	92.73	529,864
C. berlandieri subsp. *nuttaliae* Huauzontle	92,568,362	9,349,404,562	80,730,310	91.67	840,650
C. quinoa Cherry Vanilla	89,273,886	9,016,662,486	78,166,384	98.41	450,244
C. quinoa Chucapaca	93,325,818	9,425,907,618	74,220,692	99.47	36,023
C. quinoa Ku-2	130,610,478	13,191,658,278	109,438,110	99.34	131,691
C. quinoa PI 634921	89,436,886	9,033,125,486	71,954,956	99.33	47,892
C. berlandieri var. *sinuatum*	148,081,766	14,956,258,366	127,030,434	92.33	2,491,879
C. quinoa Atlas	316,492,592	39,401,220,212	261,505,978	98.75	1,335,514
C. quinoa Carina Red	346,159,396	43,099,867,862	278,916,504	86.29	797,151
C. quinoa PI 614886	994,572,016	100,451,773,616	851,664,032	99.37	-

5. Phylogenetic analyses

5.1. Phylogeny of quinoa accessions and related species

The genomic variants of all 25 sequenced taxa (Supplementary File 5) relative to the reference sequence were called based on the mapped Illumina reads in 25 bam files using SAMtools. To call variants in the reference accession (PI 614886), Illumina sequencing reads were mapped to the reference assembly. Variants were then filtered using VCFtools[61] and SAMtools, and the qualified SNPs were combined into a single VCF file. This VCF file was used as an input into SNPhylo[62] to construct the phylogenetic relationship using maximum likelihood and 1,000 bootstrap iterations, with the consideration of linkage disequilibrium blocks.

5.2. Phylogeny of quinoa sub-genomes, *C. pallidicaule*, and *C. suecicum*

Using OrthoMCL (see 6.1 below), orthologous gene sets containing two copies in quinoa and one copy each in *C. pallidicaule*, *C. suecicum*, and *B. vulgaris* were. In total, 7,433 gene sets were chosen, and their amino acid sequences were aligned individually for each set using MAFFT[58]. The 7,433 alignments were converted into PHYLIP format files by the seqret command in the EMBOSS package[59]. Individual gene trees were then constructed using the maximum likelihood method using proml in PHYLIP[60] with default parameters. A total of 5,807 trees supported the consensus topology in which one quinoa gene was more similar to *C. pallidicaule* and the other was more similar to *C. suecicum*.

5.3. *FLOWERING LOCUS T* (*FT*) gene tree and synteny analysis

For the identification and phylogenetic analysis of *FT* homologs in quinoa, a database with protein sequences from quinoa was established using the CLC Main Workbench (CLCbio, v6.9). The protein sequence from the *A. thaliana* flowering time gene *FT* was used as a BLAST query. Filtering for hits with an E-value < $1e^{-3}$ and with RNA-Seq evidence resulted in the identification of four quinoa proteins: AUR62013052, AUR62010060, AUR62006619, and AUR62000271. One quinoa protein (AUR62013052) appeared to be comprised of two tandem repeats which were separated for the purposes of phylogenetic analysis. Specifically, CqFT1B-1 was created from AUR62013052 by combining the first four exons of AUR62013052 and extending the fourth exon seven nucleotides to include the next in-frame stop codon. CqFT1B-2 was created from AUR62013052 by creating a new first exon from position 1,724,564 – 1,724,697 (in reverse orientation on the minus strand) and combining this with the last four exons of AUR62013052. For the construction of the phylogenetic tree, protein sequences from these five quinoa FT homologs were aligned using Clustal Omega[63] along with two *B. vulgaris* (gene models: BvFT1-miuf.t1, BvFT2-eewx.t1) and one *A. thaliana* (AT1G65480.1) homolog. Phylogenetic analysis was performed with MEGA[64] (v6.06). The JTT model was selected as the best fitting model. The initial phylogenetic tree was estimated using the neighbor joining method (bootstrap value = 50, Gaps/ Missing Data Treatment = Partial Deletion, Cutoff 95%), and the final tree was estimated using the maximum likelihood method with a bootstrap value of 1,000 replicates. The syntenic relationships between the coding sequences of the chromosomal regions surrounding these *FT* genes were visualised using the CoGE[65] GEvo tool and the Multi-Genome Synteny Viewer[66].

5.4. Phylogenetic analysis of bHLH peptides

All non-quinoa sequences were taken from Mertens *et al*[39]. The alignment of proteins was performed with Clustal Omega[63]. All positions with less than 95% site coverage were eliminated, resulting in a total of 70 positions used in the final dataset. The evolutionary history was inferred by using the maximum likelihood method based on the JTT matrix-based model[67]. The consensus tree with indicated support from 500 bootstrap replicates is presented in Extended Data Fig. 9. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model. All phylogenetic analyses were conducted in MEGA7[89]. Subclades are defined based on Heim *et al*[90]. Quinoa sequences with no expression (FPKM < 1) in the Kurmi × 0654 population were removed from the analysis.

6. Comparative genomics

6.1. Dating the ancestral tetraploidisation event

The rate of synonymous substitutions per synonymous site (Ks) was calculated for gene pairs identified individually in quinoa, *C. suecicum*, and *C. pallidicuale* using the recommended settings of the CoGe

SynMap tool[70]. Ks values were binned in increments of 0.01, and the peak in quinoa was estimated at a Ks value of 0.1. The date of the tetraploidisation event was calculated as Ks/2μ, where μ is the mutation rate. Two mutation rates were use: the estimated rate for *A. thaliana* (1.5×10^{-8})[23], and the estimated rate for core eukaryotes (8.1×10^{-9})[24].

6.2. Distinguishing the quinoa sub-genomes

Trimmed PE Illumina sequencing reads that were used for the *de novo* assembly of *C. suecicum* and *C. pallidicaule* were mapped onto the reference quinoa genome using the default settings of BWA. From both alignments, only properly paired reads were retained for downstream analyses. The read depth coverage for every base in the quinoa genome from the *C. suecicum* and *C. pallidicaule* mapping was calculated using the program GenomeCoverage in the BEDtools package[68]. A custom Perl script was used to calculate the percentage of each scaffold with more than 5X coverage from both diploids. As the genome is very repetitive and short reads are unlikely to bridge the repeat-rich region of the reference assembly, the repetitive fraction of each scaffold was also calculated and summarised in Supplementary File 6. By examining the percentage of base pairs in each scaffold covered by the mapping of each diploid and the percentage of repetitive base pairs in each scaffold, scaffolds were assigned to the A or B sub-genome if > 65% of the bases were covered by reads from one diploid and < 25% of the bases were covered by reads from the other diploid. 156 scaffolds covering 202,614,493 bp (14.6% of the quinoa genome) were assigned to the A sub-genome, and 410 scaffolds covering 646,250,932 bp (46.6%) were assigned to the B sub-genome. Of the remaining unassigned scaffolds (totalling 536,591,419 bp, 38.7% of the genome), 2462 scaffolds (236,529,454 bp, 17.1% of the genome) were found to be very repetitive (covering 80% or more of the scaffold), and were thus unlikely to be mapped properly by the Illumina short reads. The result of the sub-genome assignment is summarised in Supplementary File 6.

The relationship between the quinoa sub-genomes and the diploid species *C. pallidicaule* and *C. suecicum* was presented in a circle proportional to their sizes using Circos[69]. Orthologous regions in the three species were identified using BLASTN searches of the quinoa genome against each diploid genome individually. Single top BLASTN hits longer than 8kb were selected and presented as links between the quinoa genome assembly (arranged in chromosomes, see Supplementary Information 7.3.) and the two diploid genome assemblies on the Circos plot (Fig. 2a).

The 18-24J minisatellite repeat was previously shown to be more abundant in the B sub-genome of quinoa and related species[26]. A BLAST search of the quinoa genome using the 18-24J sequence was performed, and the number of 18-24J repeats observed in each of the scaffolds is summarised in Supplementary File 6.

6.3. Analysis of sub-genome synteny

The positions of homoeologous pairs of A- and B-sub-genome pairs (see 5.2 above) were plotted within the context of the 18 chromosomes using Circos.

6.4. Comparisons to *B. vulgaris*

The RefBeet-1.2.fna and BeetSet-2.unfiltered_genes.1408.gff3 sequence and annotation files for *B. vulgaris* were obtained from The *Beta vulgaris* Resource website ((http://bvseq.molgen.mpg.de/index.shtml) in September, 2016. Scaffolds which were known to be ordered and oriented within each chromosome were concatenated with no gap sequence to form 9

pseudomolecules. Syntenic regions between these *B. vulgaris* chromosomes and those of quinoa were identified using the recommended settings of the CoGe SynMap tool and visualized using MCScanX[71] and VGSC[72]. For the purposes of visualization, quinoa chromosomes CqB05, CqA08, CqB11, CqA15, and CqB16 were inverted.

6.5. Identification of orthologous genes

Orthologous and paralogous gene clusters were identified using OrthoMCL[28] (Supplementary Table 6). Recommended settings were used for all-against-all BLASTP comparisons (Blast+ v2.3.0[56]) and OrthoMCL analyses. Protein datasets for *Amaranthus hypochondriacus* were obtained from Phytozome (https://phytozome.jgi.doe.gov) and for *B. vulgaris* and *S. oleracea* from The *Beta vulgaris* Resource website (http://bvseq.molgen.mpg.de/index.shtml). All sequences were downloaded in March, 2016. Custom Perl scripts were utilised to process OrthoMCL outputs for visualisation with InteractiVenn[57].

Supplementary Table 6. Statistics of OrthoMCL analysis.

	C. quinoa	*A. hypochondriacus*	*B. vulgaris*	*S. oleracea*	*C. quinoa*	*C. pallidicaule*	*C. suecicum*	*B. vulgaris*
Clusters	15,211	13,585	14,972	14,110	16,542	14,054	15,217	14,292
Proteins	38,588	18,432	19,294	17,468	37,336	16,706	19,537	18,688
Singletons	6,188	4,585	7,604	4,179	7,440	1,255	2,393	8,210
Total	44,776	23,017	26,898	21,647	44,776	17,961	21,930	26,898

7. Linkage mapping and genetic marker analyses

7.1. Kurmi × 0654 population

7.1.1. Plant material

A population segregating for saponin content was created by crossing the low-saponin (sweet) variety Kurmi and the high-saponin (bitter) variety 0654. Homozygous high- and low-saponin F2 lines were identified by planting 12 F3 seeds derived from each F2 line, harvesting F4 seed from these F3 plants, and then performing foam tests on the F4 seed. Phenotyping was validated using gas chromatography/mass spectrometry (GC/MS), as described below. One plant each of 20 homozygous high-saponin, 20 homozygous low-saponin, and 20 heterozygous F3 lines were grown in soil in a greenhouse at KAUST. At maturity, inflorescences containing a mixture of flowers and seeds at various stages of development were harvested from individual plants and snap-frozen in liquid nitrogen.

7.1.2. RNA extraction and Illumina sequencing

RNA extraction and Illumina sequencing were performed as described above. Sequencing was performed for the parents (Kurmi and 0654) and 45 individual F3 progeny.

7.1.3. SNP calling

Sequencing reads from all lines were trimmed using Trimmomatic, as described above, and mapped to the reference assembly using TopHat[44], and SNPs were called using SAMtools mpileup, as described above. Two datasets were generated from these SNPs, with the datasets being combined at the end. For the first dataset, SNPs were filtered to remove positions for which read depth was < 8 and the SNP

allele frequency was ≤ 12.5%. Additionally, all positions for which data for either parent was missing were removed. The second dataset was based on all positions for which one of the parents was missing. For these positions, all positions for which at least one parent or any individuals in the population was heterozygous were removed. The total number of nucleotides called in the population was calculated for each position, and all positions for which more than two nucleotides were called were excluded. Positions for which a called nucleotide was only observed once in the population were also excluded. In any position that met these criteria, the nucleotide of the missing parent was inferred to be the second nucleotide segregating in the population. This second dataset containing inferred nucleotides for one parent was added to the first dataset for which the nucleotide of both parents was known.

7.1.4. Linkage mapping

The same set of SNPs described above (Supplementary Information 7.1.3) was used for constructing a linkage map of the Kurmi × 0654 population. Markers were assigned to linkage groups on the basis of the grouping by JoinMap v4.1 and homology of markers (same scaffold and marker position in the same 1 Mb bin on the assembly). Using the maximum likelihood algorithm of JoinMap, the order of the markers was determined; using this as start order and fixed order, regression mapping in JoinMap was used to determine the cM distances.

7.1.5. Analysis of differentially expressed genes

Genes differentially expressed between bitter and sweet lines and between green and red lines were identified using default parameters of the Cuffdiff function of the Cufflinks program[46].

7.1.6. Mapping the betalain stem colour locus

The phenotype segregated as a single gene in the F2 progeny (70 red, 22 green), and scoring stem colour in 51 F3 individuals enabled mapping of the trait to chromosome 2 (CqB03), where it mapped to the same position as a SNP marker from Scaffold 1995 (3,473,993 bp) (Supplementary File 1). This SNP causes an amino acid change (Ala to Gly) in an annotated peroxidase gene (AUR62012343) in the pigmented parent, and expression of the gene is significantly lower in the pigmented compared to non-pigmented progeny (Supplementary File 3). Peroxidase is known to regulate the stability of betalain pigments[91]. An additional 65 candidate genes lie within a 1-Mb window surrounding the mapped SNP marker (Supplementary File 4), including four other peroxidase genes, and a gene (AUR62012346) annotated as being homologous to *CYP76AD1*, which encodes for a cytochrome P450 which has also been shown to be required for production of the red betalain pigment in *Beta vulgaris* (sugar beet)[92], a member of the same family (Amaranthaceae) as quinoa. RNA-Seq analysis showed that this gene is expressed at significantly higher levels in pigmented plants than in non-pigmented plants (Supplementary File 3).

7.2. Atlas x Carina Red population

7.2.1. Plant material

Quinoa varieties Atlas and Red Carina (also called Carina Red) are two registered varieties based on single F8-lines. Carina Red seeds contain saponins in the outer fruit layer and these seeds are bitter while Atlas seeds are virtually free of saponins (< 0.1 g kg^{-1}) and are non-bitter.

Crosses between several Atlas genotypes as the female and Carina Red as the male parent were made by applying pollen from Carina Red on the Atlas flower heads. Using this method, selfing cannot be precluded, so the offspring from the Atlas flower heads were tested to find the F1 plants (red stem and leaf colour). About 20% of the offspring proved to be from a real cross. F2 plants were generated from a few F1 plants sown in isolation, and one F2 population was selected as the mapping population. In this F2 population, 742 plants were produced and leaf material was collected. F3 families were obtained by harvesting seed from F2 plants. The bulk seed of each F3-family was tested for the occurrence of saponins using a foam test and taste test. Out of the 742 F3 families, 175 (23.6 %) did not contain saponins (no foam and no bitter taste). This confirms the monogenetic recessive nature of the non-bitter trait. This identified which of the F2 plants were homozygous for the non-bitter trait as the mother genotype determines the phenotype of its offspring.

7.2.2. DNA extraction and Illumina sequencing

DNA from 94 non-bitter F2 genotypes was obtained for sequencing. DNA was also obtained from Atlas and Carina Red plants, although not from the same plants used in the crosses. A full lane of Illumina HiSeq2500 PE sequencing was allocated to the two genotypes from the parental lines, yielding a total sequence length of about 30 times the assembly size for each of the two parental lines. A separate library preparation of DNA of each of the 94 F2 genotypes was used to obtain data per individual genotype. Four lanes of Illumina 2500PE were used for the sequencing of the 94 F2 genotypes, yielding a total coverage for the whole set of about 160 times the assembly length, or approximately 1.8X coverage per F2 genotype.

7.2.3. Bulk segregant analysis

The read mapper BWA with the method '*mem*' was used to map all reads to the reference assembly. The bam files were sorted and indexed using PicardTools (http://broadinstitute.github.io/picard/), and variants were called in the entire set of bam files using SAMtools mpileup. A merged bam file with all combined F2 genotypes was created and used as a separate sample in SAMtools mpileup. The output vcf file was used to call variants with SAMtools BCFtools without filtering. The variants found in Atlas, Carina Red, and the merged F2 set helped to identify the real variants in the low coverage data for the individual F2 lines and to discard the read errors. In this way genotype calls were generated for all genotypes.

Many variant positions showed polymorphisms between Atlas and Carina Red, where either Atlas had genotype 0/0 (homozygous reference) and Carina Red 1/1 (homozygous alternative allele), or vice versa. In a low number of cases both parents were different from the reference genome assembly. A considerable number of variant positions gave 'heterozygous' scores for Atlas or Carina Red. These variants are indicative of positions for which sequencing reads map with equal similarity to two positions on the reference assembly. Such markers can be used for mapping, but give less certainty on the genotype scores and were therefore discarded for the mapping.

Using the variant positions for which the parents were homozygous and polymorphic, genotype calls were generated for the 94 F2 genotypes by summing up read counts over a sliding window of 500 variants. Over each 500-variant stretch, all reads with Atlas alleles were summed, and all reads with the Carina Red allele were summed. A genotype score of 1/1 (homozygous Atlas) was called when the frequency of reads was higher than a threshold level (usually >95 %, but depending on the read count

percentage of the alleles in the parents). A genotype score of 0/0 (homozygous Carina Red) was called when the read count of the alternative allele was above this threshold. Scores in between the thresholds (with reference allele frequencies usually between 25 and 75%) resulted in heterozygous genotype scores (0/1). We call the markers constructed in this way 'context500' markers (c500 markers) as we use 'contexts' of 500 consecutive variants to add up the read counts for adequate coverage for genotype calls in this low-coverage situation.

7.2.4. Linkage mapping

Over 4,000 c500 markers were obtained, and these were subjected to strict quality control. First, consecutive c500-markers needed to have a neighbouring c500-marker with genotype scores that only showed one recombination, given that no more recombination is expected in the distance between two such markers. Second, c500 markers were not allowed to significantly deviate from the expected single-locus segregation ratio of 1:2:1. Third, markers with skewed allele frequency were discarded (Atlas allele frequency < 25% in the 94 F2 lines). Markers with high Atlas allele frequency were not discarded, as those markers were of interest as being potentially associated with the non-bitter locus. Finally, c500-markers needed a minimum average depth per genotype of 200 reads. A set of 1,125 markers remained after this strict selection. Markers were assigned to linkage groups using JoinMap, with regression mapping used to obtain the genetic maps per linkage group. Linkage group names were determined by identifying markers on the Atlas × Carina Red map that share homology with markers on the previously published quinoa linkage map[13] (hereafter referred to as the SNP511 map). This was done by performing a BLAST search of the PCR fragments produced for the KASP marker analysis against the reference assembly. Based on the BLAST hits, these markers were renamed to include the four-digit scaffold number and the eight-digit position number. Markers on both maps coming from the same scaffold were used as anchors for naming the linkage groups in the Atlas × Carina Red map.

Mapping placed the non-bitter locus on quinoa Scaffold 3489. Read alignments from Atlas, Carina Red, and the merged F2 individuals from the region surrounding the non-bitter locus were visually inspected using the Integrative Genome Viewer (IGV)[93], revealing a distinct variant allele in the F2 merged bam file. To investigate this, the underlying reads from position 350,000 – 360,000 in Scaffold 3489 were extracted from the original fastq files using a custom Perl script and subsequently assembled *de novo* using SOAPdenovo2 with a k-mer size of 87. Resulting local assemblies were manually curated and appropriately inserted into the reference sequence assembly, producing an assembly we call V3.1alt1. Reads from the F2 genotypes were mapped against reference V3.1alt1 using the sampe module of BWA, sorted and merged using SAMtools, and again visualized using IGV. This visualisation supported the hypothesis that insertions are present around AUR62017204 in the sweet F2 progeny of the Atlas × Carina Red population, indicating that the Atlas parent used for this cross differs in this region from the plant used to generate re-sequencing data. Further investigation using PCR followed by restriction digests and/or sequencing indicated that in insertion is indeed present (data not shown).

7.3. Linkage map integration

Three maps were integrated: the Kurmi × 0654 map, the Atlas × Carina Red map, and the SNP511 map, with the Kurmi × 0654 map being used as the reference for the positions of anchor markers and scaling. We selected markers from the same scaffold that were in the same 10,000-bp bin in the assembly. The anchor markers on the alternative map received the position of the Kurmi × 0654 map

anchor marker in the integrated map. This process was repeated with anchor markers at the 100,000-bp bin level. The assumption is that at the 100,000-bp bin level recombination should essentially be zero. On this level, a regression of cM position on both maps yielded $R^2 > 0.85$ and often > 0.9, so the regression line can easily be used for interpolating the positions of the alternative map towards the corresponding position on the Kurmi × 0654 map. All Kurmi × 0654 markers went into the integrated map on their original position. A summary of the integrated map is shown in Supplementary Table 7.

7.4. Chromosome pseudomolecules

Pseudomolecules were assembled by concatenating scaffolds based on their order and orientation as determined from the integrated linkage map. An AGP ('A Golden Path') file was made that describes the positions of the scaffold-based assembly in coordinates of the pseudomolecule assembly, with 100 N's inserted between consecutive scaffolds. Based on these coordinates, custom scripts were used to generate the pseudomolecule assembly and to recoordinate the annotation file.

Markers from the three different linkage maps were assigned to scaffolds on the basis of unique homology of the DNA-sequence of the marker to specific scaffolds. In most cases markers from a scaffold were mapped uniformly to a single linkage group. However, for 46 scaffolds, markers were found on two linkage groups; for three scaffolds, markers were found

Supplementary Table 7. Linkage map statistics.

LG	Sub-genome	Markers	Scaffolds	Length (cM)	Estimated length[a] (Mb)
01	B	870	82	178	137.0
02	A	304	28	149	62.4
03	B	472	78	127	73.0
04	A	422	33	128	58.6
05	B	307	31	95	78.7
06	B	257	24	95	76.5
07	A	532	84	143	121.2
08	A	265	18	60	14.3
09	A	61	6	53	18.8
10	B	424	32	107	63.7
11	B	335	53	96	70.1
12	A	330	26	90	55.6
13	A	173	3	48	38.5
14	A	297	20	155	60.7
15	A	339	31	121	57.4
16	B	537	56	161	81.8
17	B	353	72	142	82.9
18	B	226	16	86	32.0
Total	-	6,504	693	2,034	1,183.3

[a]For scaffolds mapping to multiple LGs, the scaffold length was divided among the LGs in proportion to the percentage of the scaffold mapping to each LG.

on three linkage groups. For scaffolds in which this was observed in at least two of the three linkage maps, it was decided to split that scaffold into two (or three parts) and to allocate the parts to the linkage group on which markers of these parts were found. For markers in the SNP511 and Kurmi × 0654 maps, the starting position on the assembly was used to determine the placement of parts of the scaffolds on the different linkage groups. For markers in the Atlas × Carina Red map, both the starting and ending positions were used as the c500 markers cover rather large lengths in some cases.

The split was put in between the two closest marker positions that are on different linkage groups. As a first choice, the average position was taken as the border between to split parts, so long as this position was not within an annotated gene model. If the first proposed split position was in a gene model, then the middle position between this gene model start position and the previous gene model end position was taken as split position. Further, it was tested whether a Dovetail junction was present within the range between the two closest marker positions on different linkage groups; if so, the middle position in the NNN-range in the Dovetail junction was taken as the border between two parts.

Only in very few cases, inconsistencies of marker positions occurred in the form of overlapping ranging on the assembly on different linkage groups. In most cases this occurred with markers from the SNP511-map and in most cases this concerned only single SNP511-markers. The concordance between the Kurmi × 0654 and Atlas × Carina Red maps was then used to decide not to split on the basis of single SNP511-markers.

8. Saponin analyses

8.1. Determining total saponin content

Quinoa seeds were provided to BioProfile Testing Laboratories, LLC (Minneapolis, MN) for measurement of total saponin content. For this analysis, 5 g of unwashed seeds were ground and added to 30 mL water-saturated butanol (one part water with five parts n-butanol). Samples were placed on a shaker for 15 min and then centrifuged for five min at 2500 rpm. Butanol was evaporated from aliquots of the supernatant in a 90°C water bath, and samples were analysed by HPLC using a Gemini C-18 column (0 min with 75% water 25% acetonitrile, 20 min with 60% water 40% acetonitrile, 21 min with 75% water 25% acetonitrile).

8.2. Quinoa seed scanning electron microscopy (SEM)

Quinoa seeds were cut in half with a sharp razor blade and mounted on double-sided carbon tape on an aluminium stub. To minimise surface charging, samples were coated with 5 nm thick Au/Pd using K575X sputter coater (Quorum Technologies). A Quanta 200 FEG SEM equipped with an Everhart-Thornley detector was used for imaging, which was performed at an accelerating voltage of 5 kV, spot size 2.5, working distance 9 mm, and tilted at 30°. Measurements of the inner and outer seed coat layers were taken as indicated in Extended Data Fig. 8. Measurements were taken from 5 sweet and 5 bitter F3 lines of the Kurmi × 0654 population. For each line, measurements were taken from 3 seeds, at 3 different sites in each seed, with 3-5 inner and 3-5 outer measurements being taken at each site. Normality of residuals in the ANOVA was tested using the Shapiro-Wilk test and homogeneity of variance was tested using the Bartlett's test, using Genstat 18th Edition[94].

8.3. Imaging MS

Quinoa seeds were embedded in a 2% carboxymethylcellulose solution and frozen above liquid nitrogen. Sections of 50 µm thickness were obtained using a Reichardt-Jung Frigocut 2800N, modified to use a Feather C35 blade holder and blades at -20°C using a modified Kawamoto method[73]. Briefly, the sample block was first trimmed, then cryofilm (type 2C(9)) was gently adhered to the surface of the block, and sections were taken. The film with attached section was transferred to a chilled glass slide with pre-mounted double-sided conductive carbon tape and gently adhered to the surface. The frozen slide with section was transferred into a chilled 50 ml tube then freeze dried for 16 h using a Martin Christ ALPHA 1-4 LDplus freeze dryer (John Morris Scientific, Chatswood, VIC, Australia), set to −55°C and an operating pressure of 1 mBar.

2,5-dihydroxybenzoic acid (Sigma-Aldrich) matrix (40 mg ml^{-1} in 70% methanol) was applied using a HTX TM-Sprayer (HTX Technologies LLC, Carrboro, NC, USA) with attached LC20-AD HPLC pump (Shimadzu Scientific Instruments, Ermington, NSW, Australia) with the following settings: temperature 65°C, nitrogen gas pressure 10 psi, solvent flow rate of 0.1 ml min^{-1}, 8 passes at a rate of 1200 mm min^{-1}, using 2 mm spacing with 90° offset for alternate passes and a 1 mm offset for repeat passes. Sections were vacuum dried in a desiccator prior to analysis. The optical image was generated using an Epson 4400 Flatbed Scanner at 4800 dpi. For mass spectrometric analyses, a Bruker SolariX XR with 7T magnet was used with the following settings: 50 × 50 µm laser spot array, the minimum laser spot size was set with a random raster within a 35 µm area, the laser power set to 38%, 750 shots per pixel, mass range set to 150-3000 *m/z*, with optimised ion transmission between 400-1500 *m/z*, acquisition time set to 2 Megawords generating a transient of 1.46 s providing a mass resolving power of approximately 260,000 @ 400 *m/z*, with the instrument calibrated to known masses of elemental red phosphorous (Sigma-Aldrich) clusters. Images were generated using Bruker Compass FlexImaging 4.1. Data were normalised to the TIC, and brightness optimisation was employed to enhance visualisation of the distribution of selected compounds. Individual spectra were recalibrated using Bruker Compass DataAnalysis 4.4 to internally lock masses of known DHB clusters: $C_{14}H_9O_6$ = 273.039364 and $C_{21}H_{13}O_9$ = 409.055408 *m/z*. Accurate mass measurements for individual saponins and identified compounds were run using Continuous Accumulation of Selected Ions (CASI) using mass windows of 50-100 *m/z* and a transient of 4 Megaword generating a transient of 2.93 s providing a mass resolving power of approximately 390,000 @ 400 *m/z*. Lipids were putatively assigned by searching the LipidMaps database[74] (www.lipidmaps.org) and lipid class confirmed by Collision Induced Dissociation using a 10 *m/z* window centred around the monoisotopic peak with collision energy of between 15-20 V.

8.4. Saponin accumulation during seed development

To measure saponin accumulation during seed development, quinoa flowers were marked at anthesis, and seeds were sampled at 12, 16, 20, and 24 days after anthesis. A pool of 5 seeds from each time point was analysed using GC/MS.

Quantification of saponins was performed indirectly by quantifying oleanolic acid (OA) derived from the hydrolysis of saponins extracted from quinoa seeds. Seeds were immersed in 1 ml of 80% ethanol containing 10.0 µg of 2-hydroxytetradecanoic acid (as internal standard), and vortexed at 3,000 rpm for 30 s. The extracted solvent was evaporated to dryness and the sample hydrolysed using 2 ml of 2.5 N hydrochloric acid at 90°C for 2 h. The solution was cooled, supplemented with 0.25 g of NaCl and extracted twice with 1 ml of ethyl acetate. The ethyl acetate extraction solution was treated with 0.5 g

of sodium carbonate, centrifuged and the solvent was evaporated to dryness. For derivatisation, 100 µl of bis(trimethylsilyl)trifluoroacetamide was added to the dried sample and incubated at 70°C for 30 min. One microliter of the derivatised solution was analysed using single quadrupole GC-MS system (Agilent 7890 GC/5975C MSD) equipped with EI source at ionisation energy of 70 eV. The temperature of the ion source and mass analyser was set to 230°C and 150°C, respectively, and solvent delay of 7.0 min. The mass analyser was auto tuned according to the manufacturer's manual and the scan was set from 35 to 700 with a scan speed of 2 scans/s. Chromatography separation was performed using DB-5MS fused silica capillary column (30m x 0.25 mm I.D., 0.25 µm film thickness; Agilent J&W Scientific), chemically bonded with 5% phenyl 95% methylpolysiloxane cross-linked stationary phase. Helium was used as the carrier gas with constant flow rate of 1.0 ml min^{-1}.

The quantification of OA in each sample was performed using a standard curve based on standards of OA. Standard solutions were processed as saponin samples. Extracted ion chromatogram (EIC) peak area of ion 202 Da for OA-TMS derivative and 272 Da for 2-hydroxytetradecanoic acid-2TMS derivative was used for quantification of OA in the samples.

8.5. Saponin measurements in bitter and sweet seeds

To confirm previous measurements of saponins using foam test measurements, we performed GC/MS measurements on lines of the mapping population. To verify the absence of saponins in sweet lines, a pool of 12 seeds was subjected to GC/MS as described above. Also, to confirm the presence of saponins in bitter lines, 12 single seeds of each line were subjected to GC/MS as described above. GC/MS revealed quantitative differences in OA-based saponins in the bitter lines; hence, detailed analysis using liquid chromatography/MS (LC/MS) was performed on two contrasting lines: 7 and 75.

Quinoa seeds were ground into powder and 50 mg of the flour was extracted in an ultrasonic bath (20 min) using 1 ml of 80% methanol-0.1% acetic acid with 40 µg of digoxin as internal standard. After centrifugation at 1000 g for 10 min, the supernatant was removed, and the extraction was repeated once more without internal standard. The supernatants were collected, and evaporated of methanol to obtain extracts in water containing acetic acid. The crude extracts were loaded onto a pre-equilibrated column (HyperSep C18 500 mg/3 ml SPE, precondition with 3 ml methanol, and then 3 ml water). Subsequently, the columns were washed with 3 ml of water and saponins were eluted with 3 ml of methanol, creating a C18 fraction in which the saponin eluted. Samples were evaporated to dryness, and the residues were dissolved in 300 µl of 60% methanol. The samples were filtered through 0.2 µm PTFE filters before LC-MS/MS analysis. Analysis of saponins in quinoa was performed on a Dionex Ultimate 3000 UHPLC system coupled with a Q-Exactive plus mass spectrometer (Orbitrap detector, Thermo Scientific). Chromatographic separation was carried out on a Phenomenex Kinetex C18 (100 × 2.1 mm, 5 µm) column, at 35°C. The mobile phase A and B was 0.1% formic acid-95% acetonitrile-5% water and 0.1% formic acid-95% water-5% acetonitrile, respectively. The gradient used was 0-25 min, 20% - 35% A; 25 - 40 min, 35% - 45% A; 40 - 50 min, 45% - 100% A; 50 - 55 min, 100% A; 55 - 56 min, 100% - 20% A; 56 - 65 min, 20% A. The flow rate was 250 µl min^{-1}. The Q Exactive plus mass spectrometer was equipped with a heated electrospray ionisation source and operated in negative-ion mode. The spray voltage, capillary temperature and vaporiser temperature were set at 2.50 kV, 250°C and 310°C, respectively. The sheath gas, auxiliary gas, sweep gas and S-lens RF level were set at 50.0, 13.0, 3.0 l min^{-1} and 50 V, respectively. Nitrogen was used for the spray stabilisation, higher-energy collision dissociation (HCD) cell, and damping gas in the C-trap. The instrument was calibrated in

negative ion mode every day. The analyses were performed in the Full MS and all-ion-fragmentation (AIF) negative-ion mode. The mass spectrometer acquired a Full MS scan and an AIF MS scan at a resolution of 35,000 and 70,000, respectively. The automatic gain control (AGC) target (number of ions to fill the C-Trap) was set to 10^6 with a maximum injection time of 50 ms. The Full MS and the AIF MS scan ranges were respectively set to *m/z* 400–1,500 and *m/z* 200–1,500 with microscan 1. All of the ions from the quadrupole were sent to the HCD collision cell where they were fragmented at a normalised collision energy of 25.0 eV (z = 1). Normality of residuals in the ANOVA was tested using the Shapiro-Wilk test and homogeneity of variance was tested using the Bartlett's test, using Genstat 18th Edition [94]. A Games Howell post hoc test was used to assign significant groups. A summary of the most abundant saponins detected in these lines is show in Supplementary Table 8.

Supplementary Table 8. The most abundant saponins observed in seeds of quinoa lines 7 and 75 in the Kurmi × 0654 population.

Peak	Line 7 (µg/g)[1]	Line 75 (µg/g)[a]	RT (min)	Formula	*m/z* experimental	m/z difference (ppm)	MS/MS *m/z*[b]
1	13.18 ±0.97	36.38 ±1.86	4.07	$C_{48}H_{77}O_{21}$	989.49821	1.94508	781[M-Hex-H]$^-$, 619[M-Hex-Hex-H]$^-$, 487[M-Hex-Hex-Pen-H]$^-$
2	1.34 ±0.16	1.55 ±0.02	5.51	$C_{56}H_{89}O_{28}$	1209.55704	2.02999	1001[M-Hex-H]$^-$, 839[M-Hex-Hex-H]$^-$, 677[M-Hex-Hex-Hex-H]$^-$, 515[M-Hex-Hex-Hex-Hex-H]$^-$
3	3.64 ±0.15	8.96 ±0.03	5.53	$C_{42}H_{67}O_{16}$	827.44510	1.97907	619[M-Hex-H]$^-$, 487[M-Hex-Pen-H]$^-$
4	4.97 ±0.39	8.82 ±0.50	7.07	$C_{50}H_{79}O_{23}$	1047.50439	2.51330	839[M-Hex-H]$^-$, 677[M-Hex-Hex-H]$^-$, 531[M-Hex-Hex-dHex-H]$^-$
5	21.33 ±0.99	72.92 ±3.03	7.90	$C_{54}H_{85}O_{26}$	1149.53486	1.22213	941[M-Hex-H]$^-$, 795[M-Hex-dHex-H]$^-$, 633[M-Hex-dHex-Hex-H]$^-$, 487[M-Hex-dHex-Hex-dHex-H]$^-$
6	38.55 ±1.43	116.34 ±4.29	8.03	$C_{48}H_{75}O_{21}$	987.48295	2.34954	779[M-Hex-H]$^-$, 617[M-Hex-Hex-H]$^-$, 485[M-Hex-Hex-Pen-H]$^-$
7	2.84 ±0.14	5.92 ±0.05	8.21	$C_{50}H_{79}O_{23}$	1047.50360	1.75344	839[M-Hex-H]$^-$, 677[M-Hex-Hex-H]$^-$, 515[M-Hex-Hex-Hex-H]$^-$
8	1.11 ±0.05	4.98 ±0.17	8.61	$C_{44}H_{69}O_{18}$	885.45045	1.70795	677[M-Hex-H]$^-$, 515[M-Hex-Hex-H]$^-$
9	2.08 ±0.17	3.79 ±0.10	9.12	$C_{44}H_{69}O_{18}$	885.45058	1.85258	677[M-Hex-H]$^-$, 515[M-Hex-Hex-H]$^-$
10	15.47 ±0.78	3.37 ±0.01	10.19	$C_{42}H_{67}O_{16}$	827.44555	-0.93181	619[M-Hex-H]$^-$, 487[M-Hex-Pen-H]$^-$
11	48.31 ±2.64	133.17 ±10.99	10.56	$C_{55}H_{87}O_{27}$	1179.54606	1.73203	971[M-Hex-H]$^-$, 809[M-Hex-Hex-H]$^-$, 647[M-Hex-Hex-Hex-H]$^-$, 515[M-Hex-Hex-Hex-Pen-H]$^-$
12	356.18 ±8.53	660.50 ±14.06	10.72	$C_{49}H_{77}O_{22}$	1017.49284	1.61059	809[M-Hex-H]$^-$, 647[M-Hex-Hex-H]$^-$, 515[M-Hex-Hex-Pen-H]$^-$
13	77.48 ±3.46	176.54 ±6.54	13.30	$C_{43}H_{67}O_{17}$	855.44035	2.30976	647[M-Hex-H]$^-$, 501[M-Hex-dHex-H]$^-$
14	18.26 ±0.80	0 ±0	15.24	$C_{48}H_{75}O_{21}$	987.48291	2.31153	779[M-Hex-H]$^-$, 617[M-Hex-Hex-H]$^-$, 485[M-Hex-Hex-Pen-H]$^-$
15	202.29 ±7.62	49.66 ±0.16	19.76	$C_{43}H_{67}O_{17}$	855.43900	0.72922	647[M-Hex-H]$^-$, 515[M-Hex-Pen-H]$^-$
16	3.22 ±0.21	12.87 ±0.23	20.07	$C_{38}H_{59}O_{13}$	723.39765	2.12250	515[M-Hex-H]$^-$
17	46.59 ±2.28	167.39 ±1.56	21.00	$C_{42}H_{67}O_{15}$	811.45000	1.78802	603[M-Hex-H]$^-$, 471[M-Hex-Pen-H]$^-$
18	73.25 ±2.28	0 ±0	30.50	$C_{42}H_{67}O_{15}$	811.45026	1.71194	603[M-Hex-H]$^-$, 471[M-Hex-Pen-H]$^-$
19	14.16 ±0.70	2.38 ±0.10	34.90	$C_{36}H_{57}O_{10}$	649.39699	1.94993	471[M-Pen-H]$^-$

[a]For relative quantification, data are means±SES of three technical replicates.
[b]The selected ion to give the formula is [M+HCOO]$^-$.
Hex, hexose, dHex – deoxyhexose; Pen, pentose.
Colour indicates relative abundance to each other (red, high; blue, low).
RT, retention time

8.6. Saponin identification in quinoa

The preparation of 20 mg of seeds was performed according to metaSysX standard procedure, a modified protocol from Giavalisco *et al*.[75] Samples were measured with a Waters ACQUITY

Supplementary Table 9. Saponins identified in quinoa using LC/MS.

Peak ID[a]	Compound Name[b]	m/z mean	m/z diff (ppm)	RT diff	Adduct	Chemical Sum Formula	Intensity
PN_1	AG 533 (Hex-Pent)	989.497	NA	NA	[M+H]+	$C_{48}H_{76}O_{21}$	NA
PN_2	AG487 (Hex-Hex-Pent) a	1149.533	NA	NA	[M+HCOOH-H]-	$C_{53}H_{84}O_{24}$	7486571
PN_3	AG487 (Hex-Hex-Pent) b	1149.533	-1.936	0.031	[M+HCOOH-H]-	$C_{53}H_{84}O_{24}$	103949903
PN_4	AG487 (Hex-Pent)	987.481	-2.186	0.035	[M+HCOOH-H]-	$C_{47}H_{74}O_{19}$	173071999
PN_5	AG487 (Hex-Pent)	941.475	-1.689	-0.048	[M-H]-	$C_{47}H_{74}O_{19}$	13019045
PN_6	AG489 (Hex-Hex-HexA)	1149.534	2.214	0.037	[M-H]-	$C_{54}H_{86}O_{26}$	1036574
PN_7	AG489 (Hex-Hex-Pent)	1151.549	-1.946	0.025	[M+HCOOH-H]-	$C_{53}H_{86}O_{24}$	2453289
PP_1	AG489 (Hex-Hex-Pent)	1107.555	-2.608	0.012	[M+H]+	$C_{53}H_{86}O_{24}$	4595591
PN_8	AG489 (Hex-Pent) a	989.496	-1.477	0.024	[M+HCOOH-H]-	$C_{47}H_{76}O_{19}$	53802637
PN_9	AG489 (Hex-Pent) b	989.496	NA	NA	[M+HCOOH-H]-	$C_{47}H_{76}O_{19}$	13053718
PP_2	AG489 (Hex-Pent) c	945.504	NA	NA	[M+H]+	$C_{47}H_{76}O_{19}$	8789018
PN_10	Hed (Hex-Hex-Pent)	1135.554	0.157	0.032	[M+HCOOH-H]-	$C_{53}H_{86}O_{23}$	24757400
PN_11	Hed (Hex-Pent)	973.501	-2.064	0.033	[M+HCOOH-H]-	$C_{47}H_{76}O_{18}$	191008645
PN_12	Hed (Pent) a	811.449	-2.979	0.018	[M+HCOOH-H]-	$C_{41}H_{66}O_{13}$	162845538
PN_13	Hed (Pent) b	811.450	-1.561	0.017	[M+HCOOH-H]-	$C_{41}H_{66}O_{13}$	4089842
PN_14	Hed (Pent-Hex)	973.502	-1.937	0.038	[M+HCOOH-H]-	$C_{47}H_{76}O_{18}$	16087860
PN_15	OA (Hex-HexA) a	955.491	NA	NA	[M-H]-	$C_{48}H_{76}O_{19}$	1645550
PN_16	OA (Hex-HexA) b	955.491	NA	NA	[M-H]-	$C_{48}H_{76}O_{19}$	34692800
PN_17	OA (Hex-HexA) c	953.439	NA	NA	[M-H]-	$C_{47}H_{70}O_{20}$	103918182
PN_18	OA (Hex-Hex-HexA)	1117.543	NA	NA	[M-H]-	$C_{54}H_{86}O_{24}$	16254522
PN_19	PA (Hex-Hex) a	1047.502	-2.146	0.024	[M+HCOOH-H]-	NA	20073166
PN_20	PA (Hex-Hex) b	1047.502	-0.729	0.029	[M+HCOOH-H]-	NA	23935392
PN_21	PA (Hex-HexA)	1015.475	NA	NA	[M-H]-	NA	22307362
PP_3	PA (Hex-HexA)	1017.488	NA	NA	[M+H]+	NA	5401413
PN_22	PA (Hex-Hex-Hex)	1209.555	-1.248	0.031	[M+HCOOH-H]-	NA	3052276
PN_23	PA (Hex-Hex-HexA)	1177.527	NA	NA	[M-H]-	NA	52403817
PN_24	PA (Hex-Hex-Pent) a	1179.544	-0.264	0.031	[M+HCOOH-H]-	$C_{54}H_{86}O_{25}$	17463995
PN_25	PA (Hex-Hex-Pent) b	1179.544	-0.873	0.037	[M+HCOOH-H]-	$C_{54}H_{86}O_{25}$	139879134
PN_26	PA (Hex-Pent)a	1017.491	-3.464	0.031	[M+HCOOH-H]-	$C_{48}H_{76}O_{20}$	217482883
PN_27	PA (Hex-Pent) b	1017.492	-1.161	0.025	[M+HCOOH-H]-	$C_{48}H_{76}O_{20}$	17043061
PN_28	PA (Pent) a	855.438	NA	NA	[M+HCOOH-H]-	$C_{42}H_{66}O_{15}$	300217546
PP_4	PA (Pent) a	828.472	NA	NA	[M+NH4]+	$C_{42}H_{66}O_{15}$	29492528
PN_29	PA (Pent) b	855.439	-2.387	0.023	[M+HCOOH-H]-	$C_{42}H_{66}O_{15}$	72315811
PN_30	PA (Pent-Hex)	1017.491	-2.995	0.031	[M+HCOOH-H]-	$C_{48}H_{76}O_{20}$	85088415
PN_31	PA (Pent-HexA)	985.966	NA	NA	[M-H]-	$C_{48}H_{74}O_{21}$	17491627
PP_5	PA (Pent-HexA)	987.478	NA	NA	[M+H]+	$C_{48}H_{74}O_{21}$	5368711
PN_32	PA (Pent-HexA)	985.464	NA	NA	[M-H]-	$C_{47}H_{72}O_{19}$	89920214
PN_33	SA (Hex-Hex) a	1031.507	NA	NA	[M+HCOOH-H]-	NA	9548630
PP_6	SA (Hex-Hex) a	1004.541	NA	NA	[M+NH4]+	NA	1703802
PN_34	SA (Hex-Hex) b	1031.507	NA	NA	[M+HCOOH-H]-	NA	NA
PP_7	SA (Hex-Hex) b	987.514	NA	NA	[M+H]+	NA	NA
PN_35	SA (Hex-HexA)	999.481	NA	NA	[M-H]-	NA	162719
PN_36	SA (Hex-Hex-HexA)	1161.532	NA	NA	[M-H]-	$C_{55}H_{86}O_{26}$	30580697

[a] PP denotes a peak identified in positive mode, PN in negative mode.
[b] PA: phytolaccagenic acid, Hed: hederagenin, SA: serjanic acid, OA: oleanolic acid, AG533, AG489, AG515, AG487 refer to new aglycones with a specific *m/z*.
Pen, pentose; Hex, hexose; HexA, corresponding sugar acid, a and b denote saponins with similar m/z means, but with different retention times
RT, retention time

Reversed Phase Ultra Performance Liquid Chromatography (RP-UPLC) coupled to a Thermo-Fisher Exactive mass spectrometer which consists of an electrospray ionisation source and an Orbitrap mass analyser. A C18 column was used for the hydrophilic measurements. Chromatograms were recorded in Full Scan MS mode (Mass Range [100−1,500]). Extraction of the LC-MS data was accomplished with the software REFINER MS 7.5 (GeneData). Saponins detected in the reference quinoa accession are shown in Supplementary Table 9.

8.7. Computational 3D modelling of bHLH protein structures

Template search models for AUR62017206, AUR62017204 and AUR62010677 were performed with SwissModel[76]. Homology models for the bHLH region were built using the transcription factors Myc, Max and the sterol regulatory element binding protein 1A as 3D support, which have 27%, 30% and 25% sequence identity, respectively, for the bHLH region. PDB templates 1an2, 1nlw, 4h10, 1hlo, 1nkp, 1am9 were used, and the resulting models were compared individually. Best models (as judged by the QMEAN value) were obtained for all three quinoa bHLH sequences when using 1nkp as a template. 1nkp is the crystal structure of Myc-Max recognising DNA. Resulting homology models for the bHLH region displayed good model quality indicators: model QMEAN/sequence identities were -0.15/29%, 0.16/31% and -0.28/27% for AUR62017206, AUR62017204 and AUR62010677, respectively. 3D models were visualised using Pymol. Sequence alignment and computational homology modelling showed that the residues determining the specificity for the E box motif (CACGTG) are strictly conserved and positioned as seen in Myc or Max (Extended Data Fig. 9b-d). This conservation strongly suggested that AUR62017206, AUR62017204 and AUR62010677 have the same DNA specificity as Myc or Max for CACGTG. In particular the presence of a key arginine indicates that all three bHLH bind to class A, and not to non-canonical class B (CAGCTC) E box motifs. Our modelling further supports that basic residues from the N-terminal helix and the C-terminal lysine from the loop region are capable of engaging non-specific interactions with the DNA backbone, akin to Myc or Max. The residue composition of the C-terminal helix of the bHLH motif is compatible with a coiled-coil leucine zipper dimerisation domain as seen in Myc, Max and other bHLH transcription factors, supporting that the same dimeric arrangement occurs in AUR62017206, AUR62017204 and AUR62010677.

Downstream of the bHLH is a C-terminal domain with a predicted ββαββα fold. This C-terminal domain is linked to the bHLH by a serine and asparagine rich region that is predicted to be flexible and varies in length (~15 residues in AUR62017206 to ~40 residues in AUR62017204; Extended Data Fig. 9b). The alternatively-spliced isoform of AUR62017204 (AUR62017204-AS) found in sweet lines lacks this C-terminal ββαββα domain.

Although β/α repeats are very common, 3D structures of significant sequence similarity to the full-length ββαββα sequence have not yet been determined, because neither Hidden Markov Models nor gene threading approaches gave significant hits. BLAST searches using the quinoa bHLH C-terminal domain retrieved only sequences from flowering plants, suggesting that it evolved in angiosperms. Moreover, in flowering plants, almost all bHLH sequences with homology to the quinoa bHLHs AUR62017204, AUR62017206 and AUR62010677 have this domain (at the time of the search, only one uncharacterised and unreviewed sequence from spinach, SOVF_157670, lacked this C-terminal domain), suggesting that it is an essential requirement for the biological function. 3D structural homology searches (DALI) using *ab initio* structures (QUARK) gave the most significant hits to the C-terminal domain of the *Escherichia coli* arginine-repressor (DALI Z-score of 5.8, 2.5 Å r.m.s.d to PDB

entry 1xxc). This domain mediates protein oligomerisation and arginine binding, which influences the oligomer stability[95]. It is therefore plausible that the C-terminal domains of AUR62017204, AUR62017206 and AUR62010677 serve a similar purpose, namely multimerisation that might be modulated by small-molecule ligands. A multimerisation function is further supported by the similarity of the predicted 3D fold to one of the ACT domains (although the secondary structure elements are permutated in this family; see PDB entry 1ZPV for closest match) that forms dimers and acts as a regulatory domain[96]. Moreover, compared to canonical bHLH transcription factors in animals, all plant proteins with similarity to AUR62017204, AUR62017206 and AUR62010677 have a dimerisation coiled-coil domain that is substantially shorter (by 30%-50%). Hence, the dimerisation strength of the plant bHLHs might be less than that of animal bHLHs, and the C-terminal domain's capacity to contribute to (possibly ligand-influenced) dimerisation might be required for stable DNA binding and transcription factor activity.

References

80. Christensen, S. A. *et al.* Assessment of genetic diversity in the USDA and CIP-FAO international nursery collections of quinoa (*Chenopodium quinoa* Willd.) using microsatellite markers. *Plant Genetic Resources: Characterization and Utilization* **5,** 82–95 (2007).
81. Jarvis, D.E. *et al.* Simple sequence repeat marker development and genetic mapping in quinoa (*Chenopodium quinoa* Willd.). *J. Genet.* **87,** 39–51 (2008).
82. Conn, S. J. *et al.* Protocol: optimising hydroponic growth systems for nutritional and physiological analysis of *Arabidopsis thaliana* and other plants. *Plant Methods* **9,** 4 (2013).
83. Gordon, S. P. *et al.* Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *Plos One* **10,** e0132628 (2015).
84. Guizard, S., Piégu, B. & Bigot, Y. DensityMap: a genome viewer for illustrating the densities of features. *BMC Bioinformatics* **17,** 204 (2016).
85. Axtell, M. J. ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* **19,** 740–751 (2013).
86. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19,** 1117–1123 (2009).
87. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
88. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **9,** 868–877 (1999).
89. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33,** 1870–1874 (2016).
90. Heim, M. A. *et al.* The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. *Mol. Biol. Evol.* **20,** 735–747 (2003).
91. Gandía-Herrero, F & García-Carmona, F. Biosynthesis of betalains: yellow and violet plant pigments. *Trends Plant Sci.* **18,** 334–343 (2013).
92. Hatlestad, G. J. *et al.* The beet *R* locus encodes a new cytochrome P450 required for red betalain production. *Nat. Genet.* **44,** 816–820 (2012).
93. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–192 (2013).
94. VSN International (2011). GenStat *for Windows* **14th** Edition. VSN International, Hemel Hempstead, UK. Web page: GenStat.co.uk
95. Duyne, G. D. V., Ghosh, G., Maas, W. K. & Sigler, P. B. Structure of the oligomerization and L-arginine binding domain of the arginine repressor of *Escherichia coli*. *J. Mol. Biol.* **256,** 377–391 (1996).
96. Chipman, D. M. & Shaanan, B. The ACT domain family. *Curr. Opin. Struc. Biol.* **11,** 694–700 (2001).