ARTICLE

OPEN

doi:10.1038/nature21370

The genome of *Chenopodium quinoa*

David E. Jarvis[1]*, Yung Shwen Ho[1]*, Damien J. Lightfoot[1]*, Sandra M. Schmöckel[1]*, Bo Li[1]*, Theo J. A. Borm[2], Hajime Ohyanagi[3], Katsuhiko Mineta[4], Craig T. Michell[5], Noha Saber[1], Najeh M. Kharbatia[6], Ryan R. Rupper[7], Aaron R. Sharp[7], Nadine Dally[8], Berin A. Boughton[9], Yong H. Woo[1], Ge Gao[1], Elio G. W. M. Schijlen[10], Xiujie Guo[1], Afaque A. Momin[3], Sónia Negrão[1], Salim Al-Babili[1], Christoph Gehring[1], Ute Roessner[9], Christian Jung[8], Kevin Murphy[11], Stefan T. Arold[3], Takashi Gojobori[3], C. Gerard van der Linden[2], Eibertus N. van Loo[2], Eric N. Jellen[7], Peter J. Maughan[7] & Mark Tester[1]

Chenopodium quinoa **(quinoa) is a highly nutritious grain identified as an important crop to improve world food security. Unfortunately, few resources are available to facilitate its genetic improvement. Here we report the assembly of a high-quality, chromosome-scale reference genome sequence for quinoa, which was produced using single-molecule real-time sequencing in combination with optical, chromosome-contact and genetic maps. We also report the sequencing of two diploids from the ancestral gene pools of quinoa, which enables the identification of sub-genomes in quinoa, and reduced-coverage genome sequences for 22 other samples of the allotetraploid goosefoot complex. The genome sequence facilitated the identification of the transcription factor likely to control the production of anti-nutritional triterpenoid saponins found in quinoa seeds, including a mutation that appears to cause alternative splicing and a premature stop codon in sweet quinoa strains. These genomic resources are an important first step towards the genetic improvement of quinoa.**

Quinoa (*Chenopodium quinoa* Willd., $2n = 4x = 36$) is a highly nutritious crop that is adapted to thrive in a wide range of agroecosystems. It was presumably first domesticated more than 7,000 years ago by pre-Columbian cultures and was known as the 'mother grain' of the Incan Empire[1]. Quinoa has adapted to the high plains of the Andean Altiplano (>3,500 m above sea level), where it has developed tolerance to several abiotic stresses[2–4]. Quinoa has gained international attention because of the nutritional value of its seeds, which are gluten-free, have a low glycaemic index[5], and contain an excellent balance of essential amino acids, fibre, lipids, carbohydrates, vitamins, and minerals[6]. Quinoa has the potential to provide a highly nutritious food source that can be grown on marginal lands not currently suitable for other major crops. This potential was recognized when the United Nations declared 2013 as the International Year of Quinoa, this being one of only three times a plant has received such a designation.

Despite its agronomic potential, quinoa is still an underutilized crop[7], with relatively few active breeding programs[8]. Breeding efforts to improve the crop for important agronomic traits are needed to expand quinoa production worldwide. To accelerate the improvement of quinoa, we present here the allotetraploid quinoa genome. We demonstrate the utility of the genome sequence by identifying a gene that probably regulates the presence of seed triterpenoid saponin content. Moreover, we sequenced the genomes of additional diploid and tetraploid *Chenopodium* species to characterize genetic diversity within the primary germplasm pool for quinoa and to understand sub-genome evolution in quinoa. Together, these resources provide the foundation for accelerating the genetic improvement of the crop, with the objective of enhancing global food security for a growing world population.

Sequencing, assembly and annotation

We sequenced and assembled the genome of the coastal Chilean quinoa accession PI 614886 (BioSample accession code SAMN04338310) using single-molecule real-time (SMRT) sequencing technology from Pacific Biosciences (PacBio) and optical and chromosome-contact maps from BioNano Genomics[9] and Dovetail Genomics[10]. The assembly contains 3,486 scaffolds, with a scaffold N50 of 3.84 Mb and 90% of the assembled genome contained in 439 scaffolds (Table 1). The total assembly size of 1.39 gigabases (Gb) is similar to the reported size estimates of the quinoa genome (1.45–1.50 Gb (refs 11,12)). To combine scaffolds into pseudomolecules, an existing linkage map from quinoa[13] was integrated with two new linkage maps. The resulting map (Extended Data Fig. 1) of 6,403 unique markers spans a total length of 2,034 centimorgans (cM) and consists of 18 linkage groups (Supplementary Table 7), corresponding to the haploid chromosome number of quinoa. Pseudomolecules (hereafter referred to as chromosomes, which are numbered according to a previously published single-nucleotide polymorphism (SNP) linkage map[13]) were created by anchoring 565 scaffolds to the linkage map, representing 1.18 Gb (85%) of the total assembly length (Table 1, Supplementary Data 1, Supplementary Data 2). This assembly represents a substantial improvement over the previously published quinoa draft genome sequence, which contained more than 24,000 scaffolds with 25% missing data[14].

Predicted protein-coding and microRNA genes (Supplementary Table 4) were annotated using a combination of *ab initio* prediction and transcript evidence gathered from RNA sequenced from multiple tissues using both RNA-seq and PacBio isoform sequencing (Iso-Seq) approaches (Extended Data Fig. 2a). The annotation contains 44,776 gene models (Supplementary Table 2, Extended Data Fig. 2b), which is in line with sequenced tetraploid species[15], and includes 33,365 genes with annotation edit distance (AED)[16,17] values ≤ 0.3 (Extended Data Fig. 2c). Of the genome, 64% was found to be repetitive, including a large proportion of long terminal repeat (LTR) transposable elements (Supplementary Table 1). A majority (97.3%) of the 956 genes in the Plantae BUSCO dataset[18] were identified in the annotation

[1]King Abdullah University of Science and Technology (KAUST), Biological and Environmental Sciences & Engineering Division (BESE), Thuwal, 23955-6900, Saudi Arabia. [2]Wageningen University and Research, Wageningen UR Plant Breeding, Wageningen, The Netherlands. [3]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Biological and Environmental Sciences & Engineering Division (BESE), Thuwal, 23955-6900, Saudi Arabia. [4]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences & Engineering Division (CEMSE), Thuwal, 23955-6900, Saudi Arabia. [5]King Abdullah University of Science and Technology (KAUST), Red Sea Research Center (RSRC), Biological and Environmental Sciences & Engineering Division (BESE), Thuwal, 23955-6900, Saudi Arabia. [6]King Abdullah University of Science and Technology (KAUST), Analytical Core Lab, Thuwal, 23955-6900, Saudi Arabia. [7]Brigham Young University, Department of Plant and Wildlife Sciences, College of Life Sciences, Provo, Utah 84602, USA. [8]Plant Breeding Institute, Christian-Albrechts-University of Kiel, Olshausenstr. 40, D-24118 Kiel, Germany. [9]Metabolomics Australia, The School of Biosciences, The University of Melbourne, Parkville, Victoria 3010, Australia. [10]PRI Bioscience, Plant Research International, Wageningen UR, Wageningen, The Netherlands. [11]Washington State University, Department of Crop and Soil Sciences, College of Agricultural, Human, and Natural Resource Sciences, Pullman, Washington 99164-6420, USA.
*These authors contributed equally to this work.

Table 1 | Assembly statistics for quinoa, *C. pallidicaule* and *C. suecicum*

	C. pallidicaule Diploid ($2n=2x=18$)	*C. suecicum* Diploid ($2n=2x=18$)	*C. quinoa* Allotetraploid ($2n=4x=36$)			
Illumina	✓	✓				
PacBio			✓	✓	✓	✓
BioNano				✓	✓	✓
Dovetail					✓	✓
Linkage map						✓
Total assembly size (bp)	337,010,935	536,949,265	1,325,007,020	1,395,179,653	1,385,456,844	1,183,321,377
Longest scaffold (bp)	2,949,784	1,614,553	11,561,360	11,561,360	23,816,425	137,416,624*
Number of contigs	-	-	4,232	-	-	-
N50 contig length (bp)	-	-	1,663,340	-	-	-
L50 contig count	-	-	216	-	-	-
Number of scaffolds	3,013	11,198	-	4,014	3,486	18*
N50 scaffold length (bp)	356,818	105,389	-	2,450,933	3,846,917	74,588,639*
L50 scaffold count	243	1,285	-	177	105	7*
N90 scaffold length (bp)	55,204	27,807	-	157,165	249,904	53,127,663*
L90 scaffold count	1,215	5,075	-	800	439	14*
Missing bases (%)	2.52	7.49	0.00	4.53	4.56	4.44

*Based on scaffolds assigned to pseudomolecules.

(Supplementary Table 3), which is suggestive of a complete assembly and annotation. The utility of the assembly, linkage maps, and annotation was demonstrated by mapping the betalain locus and identifying candidate genes underlying stem pigmentation (Supplementary Information 7.1.6), which is often used as a morphological marker in breeding programs.

Evolutionary history of quinoa

Quinoa resulted from the hybridization of ancestral A- and B-genome diploid species[19]. Single-gene sequencing studies previously identified pools of North American and Eurasian diploids as candidate sources of the A and B sub-genomes, respectively[20–22], with hybridization occurring somewhere in North America. To understand genome structure and evolution in quinoa further, we sequenced, assembled, and annotated the A-genome diploid *C. pallidicaule* (commonly called cañahua or kañiwa) and the B-genome diploid *C. suecicum*[21] (Fig. 1a, Table 1). A high proportion of orthologous gene pairs in quinoa showed similar rates of synonymous substitutions per synonymous site (K_s), indicative of a whole-genome duplication event (Fig. 1b). This probably represents the hybridization of ancestral diploid species, because a similar peak was not observed in *C. pallidicaule* or *C. suecicum* (Fig. 1b). Using mutation rates calculated for *Arabidopsis thaliana*[23] and for core eukaryotes[24], we estimate the tetraploidization to have occurred 3.3–6.3 million years ago.

Multiple interfertile tetraploid species have arisen from the ancestral tetraploid following hybridization, including *C. berlandieri* and *C. hircinum*, although the evolutionary relationships among quinoa and its diploid and tetraploid relatives remain unclear[25]. To begin to resolve these issues, we re-sequenced 15 additional quinoa samples representing the two major recognized groups of quinoa: highland and coastal (Supplementary Data 5). We also sequenced five accessions of *C. berlandieri* and one accession each of *C. hircinum* from the Pacific and Atlantic Andean watersheds (Supplementary Data 5). Phylogenetic analysis of these taxa indicates that North American *C. berlandieri* is the basal member of the species complex (Fig. 1c). Quinoa was thought to have been domesticated from *C. hircinum* in a single event, from which coastal quinoa was later derived (Fig. 1d, arrow 1); however, our sequencing data place a *C. hircinum* sample basal to coastal ecotypes (Fig. 1c), suggesting the possibility that quinoa was domesticated independently in highland and coastal environments (Fig. 1d, arrows 2a and 2b, respectively). Future analyses with deeper sampling of quinoa and *C. hircinum* will help clarify the relationship between *C. hircinum* and quinoa, as well as provide germplasm for breeding broadly adapted coastal quinoa cultivars for warm-season production. The SNPs identified between these accessions and the reference quinoa genome—a total of 7,809,381 (Extended Data Fig. 3, Supplementary Table 5), including 2,668,694 that are specific to quinoa—will be useful in assessing genetic diversity and identifying genomic regions associated with desirable traits.

Analysis of sub-genome structure

By mapping sequencing reads from *C. pallidicaule* and *C. suecicum* onto the quinoa scaffold assembly, and by performing BLASTN searches of each diploid against the quinoa assembly, 156 and 410 quinoa scaffolds (totalling 202.6 and 646.3 Mb) were assigned to the A and B sub-genomes, respectively (Fig. 2a, Supplementary Data 6). A mini-satellite repeat (18-24J) previously shown to be more abundant in the B sub-genome[26] is over-represented in scaffolds assigned to the B sub-genome (Supplementary Data 6). Nine chromosomes were assigned to each sub-genome (chromosomes hereafter designated as *Cq*A or *Cq*B, followed by the chromosome number), with the B sub-genome accounting for a larger percentage of both the genetic (1,087 cM) and physical (660 Mb) sizes than the A sub-genome (946 cM, 524 Mb). This result was not unexpected, given the differences in the estimated genome sizes of *C. suecicum* (815 Mb) and *C. pallidicaule* (452 Mb) based on *k*-mer analyses.

Visualization of the chromosomal locations of 5,807 homoeologous gene pairs revealed a high degree of synteny between the A and B sub-genomes (Fig. 2b). A small number of homoeologous pairs (3.1%) mapped within the same sub-genome, suggesting that recombination and chromosomal rearrangements have occurred between the A and B sub-genomes. For example, we identified homoeologous A and B sub-genome regions located in the B sub-genome chromosomes *Cq*B05 and *Cq*B03. The genes in the region of ~70–72 Mb of *Cq*B03 are phylogenetically more similar to the A-genome diploid *C. suecicum* and therefore probably originated from the A sub-genome chromosome *Cq*A12 (Fig. 2c). We also found evidence of large chromosomal rearrangements within sub-genomes, complicating the assignment of homoeologous chromosome pairs. For example, orthologue analysis clearly identifies *Cq*B05 and *Cq*A12 as homoeologous, although the same analysis is much more complicated with other chromosomes (Fig. 2b). To clarify these relationships, we identified syntenic regions between chromosomes of the diploid *Beta vulgaris*[27] ($n=9$) and the A and B sub-genome chromosomes of quinoa (Fig. 2d). These results

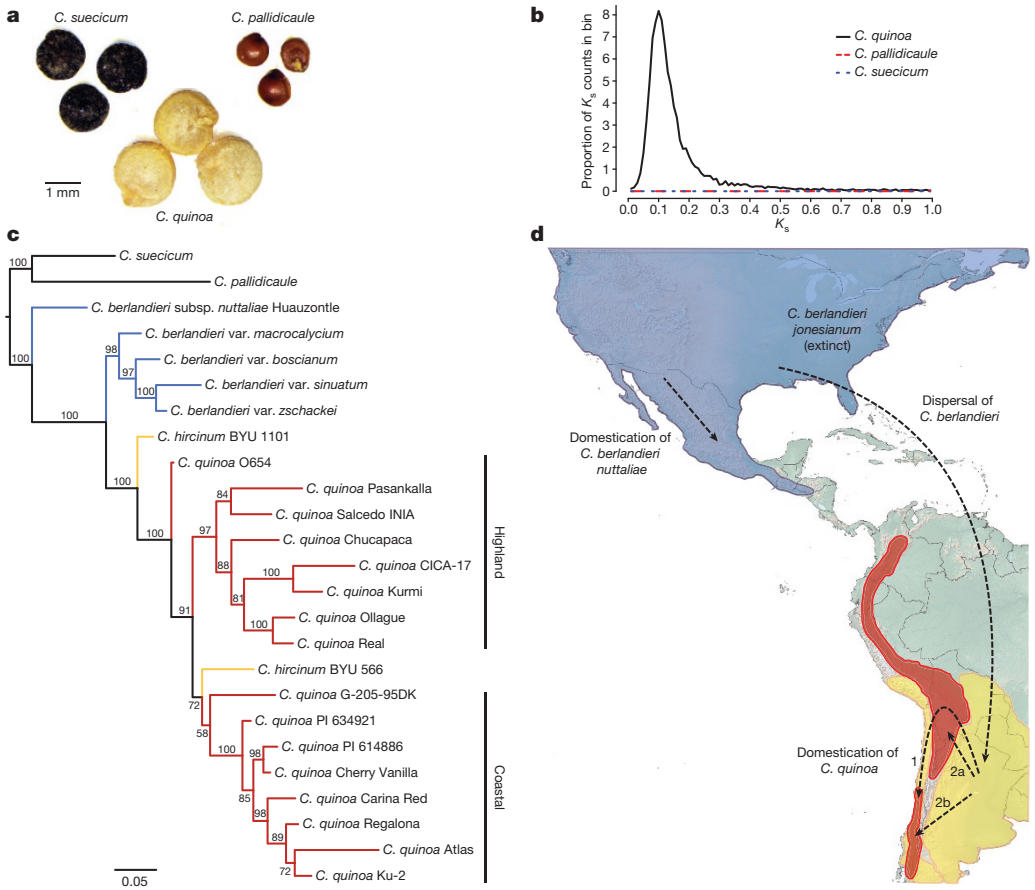


Figure 1 | Evolutionary history of quinoa. a, Seeds of *C. suecicum*, *C. pallidicaule* and quinoa. **b**, The proportion of gene pairs in each species binned according to K_s values. **c**, Maximum likelihood tree generated from 3,132 SNPs. Black branches, diploid species. Coloured branches, tetraploid species: red, quinoa; blue, *C. berlandieri*; yellow, *C. hircinum*. Branch values represent the percentage of 1,000 bootstrap replicates that support the topology. Scale bar represents substitutions per site. **d**, Evolutionary relationships of *Chenopodium* species, showing the hypothesized long-range dispersal of an ancestral *C. berlandieri* to South America, and the eventual domestication of quinoa from *C. hircinum*, either from a single event that gave rise to highland and subsequently coastal quinoa (1), or in two events that gave rise to highland (2a) and coastal (2b) quinoa independently. Blue, red and yellow shading represents the geographic distribution of *C. berlandieri*, quinoa and *C. hircinum*, respectively.

indicate that *Cq*A02 and *Cq*A04 are orthologous to *B. vulgaris* chromosomes 8 and 2 (*Bv*chr8 and *Bv*chr2), respectively, whereas *Cq*B01 appears to be the result of a chromosome fusion. Likewise, *Cq*A07 appears to be the result of a fusion between ancestral chromosomes orthologous to *Bv*chr3 and *Bv*chr7.

Analysis of sub-genome content

We used OrthoMCL[28] to identify clusters of orthologous genes in related species in Amaranthaceae (Extended Data Fig. 4), and specifically investigated the retention and loss of orthologous genes in quinoa and the three diploid species *C. pallidicaule*, *C. suecicum* and *B. vulgaris* (Fig. 3a, Supplementary Table 6). Using sets of genes for which only one copy could be found in quinoa and in each of the diploid species, we found a similar number (1,031 and 849) of genes lost from the A and B sub-genomes, respectively (Fig. 3b). The previously discussed set of 5,807 homoeologous gene pairs in quinoa are present as single copies in each of the diploids (Figs 2b, 3b) and therefore represent a core set of single-copy genes retained in each genome and sub-genome. We also investigated genes retained in multiple copies. We found that quinoa, like *C. rubrum*[29] and *B. vulgaris*[30], contains two genes that are orthologous to the *A. thaliana* gene *FT*, which regulates flowering time. Quinoa contains two homoeologous copies of *FT2* and three of *FT1*, owing to a local tandem duplication (Fig. 3c, Extended Data Fig. 5). *FT* is known to promote flowering in *A. thaliana*, and functional orthologues have been found in other species[31]; however, *B. vulgaris* was found to contain a second *FT* gene that acts antagonistically by repressing flowering before vernalization. Future functional studies will help determine the function of these duplicated genes in quinoa, which, unlike *B. vulgaris*, does not require vernalization.

Mechanisms underlying saponin production

Quinoa seeds contain a mixture of triterpene glycosides called saponins[32]. Although saponins may be beneficial for plant growth (for example, by deterring herbivory[33,34]), they must be removed before human consumption as they are haemolytic and produce a bitter flavour. Because this process is costly, is often water-intensive, and can reduce the nutritional value of the seeds[35], the development of saponin-free lines is a major breeding objective in quinoa[8]. We found that saponins accumulate in the seed pericarp (Fig. 4a, Extended Data Fig. 6) between 20 and 24 days after anthesis (Fig. 4b), eventually accounting for 4% (w/w) of the mature seed mass (Supplementary Information 8.1). We identified and annotated 43 different saponins in the seeds of the reference sample (Supplementary Table 9), adding to the almost 100 different saponins that have been previously identified in different samples of quinoa[32,36].

Naturally occurring sweet quinoa strains that contain very low levels of saponins are present within the quinoa germplasm[37], although the underlying genes regulating the absence of saponins in these lines are unknown. To identify these genes, we performed linkage mapping and bulk segregant analysis (BSA) using two populations segregating for the presence of saponins in the seeds: Kurmi (sweet) × 0654 (bitter), and Atlas (sweet) × Carina Red (bitter). Consistent with reports from other populations[38], segregation ratios in these populations indicate that the presence of seed saponins is controlled by a single gene, with the presence of saponins being dominant (71 bitter and 21 sweet in Kurmi × 0654; 567 bitter and 175 sweet in Atlas × Carina Red). We note that quantitative and qualitative differences exist in the saponins identified in bitter lines (Extended Data Fig. 7, Supplementary Table 8) and that the presence and absence of saponins was correlated with differences in seed coat thickness, with bitter lines having significantly thicker seed coats than sweet lines (Extended Data Fig. 8).

Linkage mapping and BSA in each population identified the same region on *Cq*B16 that distinguishes the bitter and sweet lines (Fig. 4c). Frequencies of the sweet allele in both populations reached 100% for markers located in *Cq*B16 on scaffold 3489. We investigated the genes in a 700-kb window surrounding this region of 100% sweet allele

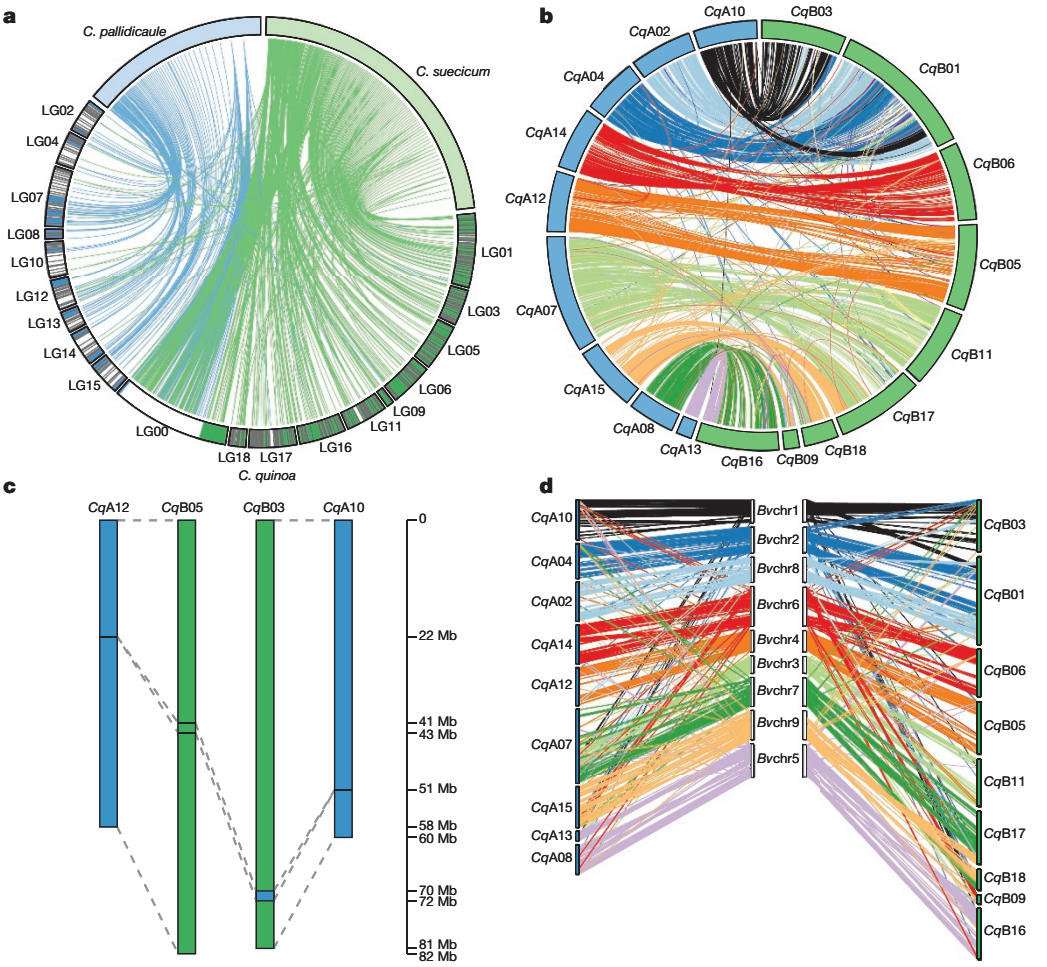


Figure 2 | Identification and characterization of quinoa sub-genomes. a, Blue lines and green lines connect regions of the *C. pallidicaule* and *C. suecicum* genomes, respectively, with their orthologous regions in the quinoa genome based on BLASTN. Quinoa scaffolds are arranged based on their positions in linkage groups, with blue- and green-coloured bars indicating sub-genome assignment based on mapped reads from the two diploid species. Scaffolds that could not be unambiguously assigned to a sub-genome based on read mapping are shown in white. Grey bars separate neighbouring scaffolds. **b**, Homoeologous gene pairs in the A (blue chromosomes) and B (green chromosomes) sub-genomes. **c**, Simplified representation of synteny between *Cq*A12, *Cq*B05, *Cq*B03 and *Cq*A10. Dotted lines connect large-scale syntenic regions between the A (blue) and B (green) sub-genomes. The scale bar indicates approximate positions defining the indicated syntenic blocks. For purposes of visualization, *Cq*B05 and *Cq*A10 were inverted. **d**, Syntenic relationships between *B. vulgaris* and the A and B sub-genomes of quinoa. Colours distinguish quinoa regions syntenic to each *B. vulgaris* chromosome. Blue and green quinoa chromosomes indicate the A and B sub-genomes, respectively.

frequency. Of the 54 annotated genes in this region (Supplementary Data 7), two are similar to genes previously shown to play a role in saponin biosynthesis. Specifically, AUR62017204 and AUR62017206 are neighbouring genes annotated as basic helix–loop–helix (bHLH) transcription factors sharing homology (Extended Data Fig. 9a) with the class IVa bHLH genes that are known to regulate triterpenoid biosynthesis in *Medicago trancatula*[39]. In *M. truncatula*, overexpression of the *triterpene saponin biosynthesis activating regulator 1* (*TSAR1*) and *TSAR2* bHLH transcription factors was recently shown to increase the expression of genes in the triterpenoid biosynthetic pathway, resulting in increased accumulation of triterpene saponins[39]. TSAR1 and TSAR2 were also found to bind to the DNA motif 5′-CACGHG-3′ (where H can be A, C, or T)[39]. In quinoa, we found that AUR62017206 (hereafter *TSAR-like 2*, *TSARL2*) was expressed in root tissue but not in flowers or immature seeds, whereas AUR62017204 (hereafter *TSARL1*) was almost exclusively expressed in seeds, with significantly lower expression levels in sweet lines (Supplementary Data 8). We identified the DNA motif bound by *M. truncatula* TSAR1 and TSAR2 within 2 kb upstream of the start codon in several saponin biosynthetic pathway genes in quinoa (Fig. 4d). Expression levels of these genes and several other genes in the saponin biosynthetic pathway were significantly downregulated in sweet lines (Fig. 4d, Supplementary Data 8). Together, these results suggest that TSARL1 might be a functional TSAR orthologue, although whether this is due to shared ancestry or convergent evolution is unclear.

The *TSARL1* transcript was alternatively spliced in the sweet progeny of Kurmi and 0654. A SNP in the last position of exon 3 (G2078C) co-segregates with the presence of saponins in the Kurmi × 0654 progeny. The G2078C SNP alters the canonical intron/exon splice boundary (Fig. 4e), probably leading to the alternative splicing at an upstream

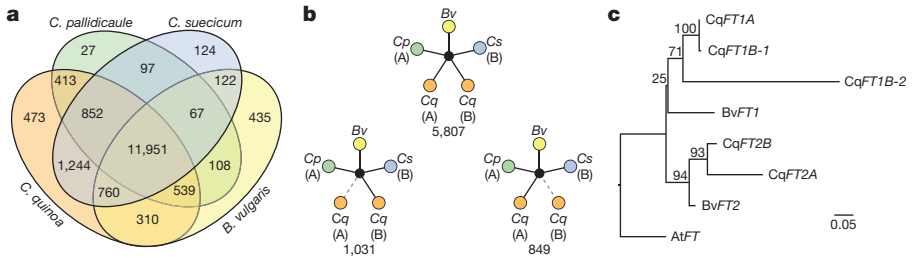


Figure 3 | Sub-genome gene loss and retention. a, The number of orthologous protein-coding gene clusters shared between or unique to quinoa, *C. pallidicaule*, *C. suecicum* and *B. vulgaris*. **b**, The number of gene sets for which each gene has been retained as a single copy in each genome/sub-genome (middle), or lost from the quinoa A (left) or B (right) sub-genome. **c**, Maximum likelihood tree of flowering locus T (*FT*) sequences, indicating the presence of two sets of orthologues in quinoa and *B. vulgaris*. The tree is rooted on the branch containing *FT* from *A. thaliana*. Branch values represent the percentage of 1,000 bootstrap replicates that support the topology. Scale bar represents substitutions per site.

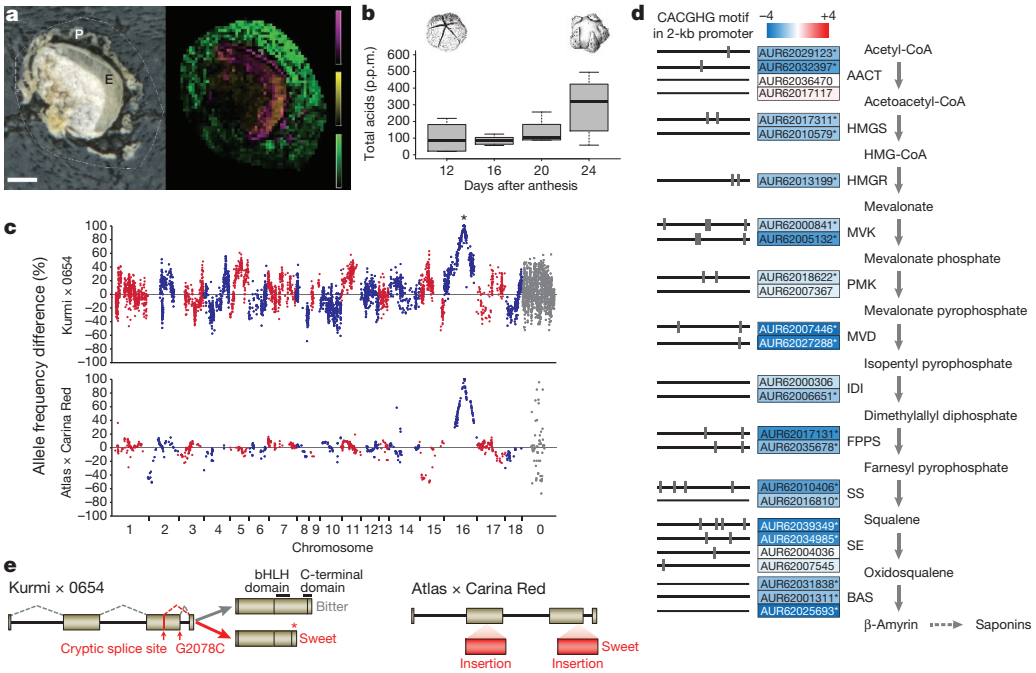


Figure 4 | Candidate gene underlying saponin production. **a**, Imaging mass spectrometry visualization of selected masses, including saponins in the pericarp of a quinoa seed. Purple gradient bar, tentative phosphatidylcholine-(34:1), ($[M+Na]^+$ $m/z = 782.5610$, calc. 782.5670, 7.7 p.p.m. error); yellow gradient bar, tentative triacylglycerol-(54:6), ($[M+K]^+$ $m/z = 917.6971$, calc. 917.6995, 2.6 p.p.m. error); green gradient bar indicates a representative saponin phytolaccagenic acid with sugar chains hexose–pentose–hexose ($[M+K]^+$ $m/z = 1173.5114$, calc. 1173.509, −2.0 p.p.m. error). Coloured bars represent the ion signal intensity scaled from 0% (bottom) to 50% (top) of maximum signal. Scale bars, 500 μm. **b**, Accumulation of saponins as measured by total acids during seed development. Illustrations represent fruit development at 12 and 24 days after anthesis. $n = 6$, 5, 5 and 5, respectively, for 12, 16, 20 and 24 days after anthesis. **c**, The percentage difference in allele frequency of sweet progeny compared to bitter progeny in the Kurmi × 0654 (top) and Atlas × Carina Red (bottom) populations. Alternating red and blue dots indicate positions of markers along alternating chromosomes, with unmapped markers in chromosome 0 shown in grey. Asterisk above the top panel indicates the approximate position of *TSARL1*. **d**, The saponin biosynthetic pathway, showing enzymes that catalyse each step of the pathway and the quinoa gene ID for genes encoding each enzyme. Boxes surrounding each gene ID are coloured according to their fold change in expression ($\log_2$) in sweet lines compared to bitter lines of Kurmi × 0654. Horizontal lines to the left of each gene ID represent the 2-kb region upstream of the start codon of each gene, with tick marks indicating the positions of motifs putatively recognized by TSARL1. **e**, Gene models of *TSARL1* in bitter and sweet lines. Red asterisk, premature stop codon in sweet lines of Kurmi × 0654.

cryptic splice site in the sweet lines (Fig. 4e). This alternative splicing of *TSARL1* results in a premature stop codon (Extended Data Fig. 9b) and a truncated protein that modelling predicts to be compromised in its ability to form homodimers and/or to bind DNA (Extended Data Fig. 9b–d, Supplementary Information 8.8), which are both necessary for regulation of transcription. All bitter strains in our re-sequencing pool share the same allele (G) found in the bitter progeny of Kurmi and 0654, whereas all sweet strains (Chucapaca, G205-95DK, Salcedo INIA, as well as the mapping population parents Kurmi and Atlas) but one (Pasankalla) contain the same allele (G2078C) as the sweet progeny. Notably, however, although the G2078C allele is present in the sequenced Atlas line, none of the sweet progeny in the Atlas × Carina Red population were found to have the G2078C allele. Additional sequencing of individual plants of the Atlas variety revealed a low level of heterogeneity within the variety for the *TSARL1* gene, with some plants containing the G2078C allele and others containing sequence insertions (Supplementary Information 7.2.4). Thus, it is likely that the Atlas plant used in the cross with Carina Red—which, importantly, was not the same plant used for sequencing—possessed a mutation other than the G2078C allele. Indeed, we found strong evidence of insertions in and around the *TSARL1* gene in all the sweet progeny of the Atlas × Carina Red population (Fig. 4e, Extended Data Fig. 10). In particular, two exonic insertions in *TSARL1* in the sweet progeny probably inactivate the gene and result in a sweet phenotype. The identification of multiple, independent mutations in *TSARL1* that co-segregate with the sweet phenotype strongly suggests that this gene regulates the presence and absence of saponins in quinoa seeds. The early steps in the saponin biosynthetic pathway that are presumably regulated by TSARL1 are shared by other pathways, including sterol biosynthesis. Inactivation of these other pathways could be detrimental to normal plant growth, although we see no noticeable phenotypic differences between bitter and sweet siblings. It is possible that precursors needed for sterol biosynthesis are provided by the methylerythritol 4-phosphate (MEP) pathway in the plastid[40,41]. Future functional studies will help clarify these issues.

Conclusions

As an emerging international crop, quinoa has great potential to enhance global food security. The high-quality reference genome assembly presented here will accelerate improvements of quinoa. Major breeding objectives for quinoa improvement include the development of shorter plants with fewer branches and more compact seed heads, increased heat and biotic stress tolerance, and the introgression of the sweet phenotype into commercial varieties. The identification of the likely causative mutation underlying the sweet phenotype not only provides insights into triterpenoid saponin biosynthesis, but also enables accelerated breeding of sweet commercial varieties using marker-assisted selection. The diversity present in the primary gene pool of quinoa, which we have begun to characterize, will also help direct future breeding strategies. The resources presented here also help to make quinoa a useful model for studying polyploid genome evolution and mechanisms of abiotic stress tolerance, in particular salinity tolerance.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 June 2016; accepted 8 January 2017.
Published online 8 February; corrected online 15 February 2017
(see full-text HTML version for details).

1. Risi, C. & Galwey, N. W. The *Chenopodium* grains of the Andes: Inca crops for modern agriculture. *Adv. Appl. Biol.* **10,** 145–216 (1984).
2. Adolf, V. I., Shabala, S., Andersen, M. N., Razzaghi, F. & Jacobsen, S.-E. Varietal differences of quinoa's tolerance to saline conditions. *Plant Soil* **357,** 117–129 (2012).
3. Hariadi, Y., Marandon, K., Tian, Y., Jacobsen, S.-E. & Shabala, S. Ionic and osmotic relations in quinoa (*Chenopodium quinoa* Willd.) plants grown at various salinity levels. *J. Exp. Bot.* **62,** 185–193 (2011).
4. Jacobsen, S.-E., Mujica, A. & Jensen, C. R. The resistance of quinoa (*Chenopodium quinoa* Willd.) to adverse abiotic factors. *Food Rev. Int.* **19,** 99–109 (2003).
5. Gordillo-Bastidas, E., Díaz-Rizzolo, D. A., Roura, E., Massanés, T. & Gomis, R. Quinoa (*Chenopodium quinoa* Willd.), from nutritional value to potential health benefits: an integrative review. *J. Nutr. Food Sci.* **6,** 497 (2016).
6. Vega-Gálvez, A. *et al.* Nutrition facts and functional potential of quinoa (*Chenopodium quinoa* Willd.), an ancient Andean grain: a review. *J. Sci. Food Agric.* **90,** 2541–2547 (2010).
7. Massawe, F., Mayes, S. & Cheng, A. Crop diversity: an unexploited treasure trove for food security. *Trends Plant Sci.* **21,** 365–368 (2016).
8. Zurita-Silva, A., Fuentes, F., Zamora, P., Jacobsen, S.-E. & Schwember, A. R. Breeding quinoa (*Chenopodium quinoa* Willd.): potential and perspectives. *Mol. Breed.* **34,** 13–30 (2014).
9. Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30,** 771–776 (2012).
10. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res.* **26,** 342–350 (2016).
11. Palomino, G., Hernández, L. T. & Torres, E. C. Nuclear genome size and chromosome analysis in *Chenopodium quinoa* and *C. berlandieri* subsp. *nuttalliae*. *Euphytica* **164,** 221–230 (2008).
12. Kolano, B., Siwinska, D., Pando, L. G., Szymanowska-Pulka, J. & Maluszynska, J. Genome size variation in *Chenopodium quinoa* (Chenopodiaceae). *Plant Syst. Evol.* **298,** 251–255 (2012).
13. Maughan, P. J. *et al.* Single nucleotide polymorphism identification, characterization, and linkage mapping in quinoa. *Plant Genome* **5,** 114–125 (2012).
14. Yasui, Y. *et al.* Draft genome sequence of an inbred line of *Chenopodium quinoa*, an allotetraploid crop with great environmental adaptability and outstanding nutritional properties. *DNA Res.* **23,** 535–546 (2016).
15. Li, F. *et al.* Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* **33,** 524–530 (2015).
16. Eilbeck, K., Moore, B., Holt, C. & Yandell, M. Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics* **10,** 67 (2009).
17. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13,** 329–342 (2012).
18. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210–3212 (2015).
19. Kolano, B. *et al.* Molecular and cytogenetic evidence for an allotetraploid origin of *Chenopodium quinoa* and *C. berlandieri* (Amaranthaceae). *Mol. Phylogenet. Evol.* **100,** 109–123 (2016).
20. Brown, D. C., Cepeda-Cornejo, V., Maughan, P. J. & Jellen, E. N. Characterization of the granule-bound starch synthase I gene in *Chenopodium*. *Plant Genome* **8,** 1 (2014).
21. Štorchová, H., Drabešová, J., Cháb, D., Kolář, J. & Jellen, E. N. The introns in flowering locus T-like (*FTL*) genes are useful markers for tracking paternity in tetraploid *Chenopodium quinoa* Willd. *Genet. Resour. Crop Evol.* **62,** 913–925 (2015).
22. Walsh, B. M., Adhikary, D., Maughan, P. J., Emshwiller, E. & Jellen, E. N. *Chenopodium* polyploidy inferences from salt overly sensitive 1 (*SOS1*) data. *Am. J. Bot.* **102,** 533–543 (2015).
23. Koch, M. A., Haubold, B. & Mitchell-Olds, T. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol. Biol. Evol.* **17,** 1483–1498 (2000).
24. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290,** 1151–1155 (2000).
25. Wilson, H. D. Quinua biosystematics II: free-living populations. *Econ. Bot.* **42,** 478–494 (1988).
26. Kolano, B. *et al.* Chromosomal localization of two novel repetitive sequences isolated from the *Chenopodium quinoa* Willd. genome. *Genome* **54,** 710–717 (2011).
27. Dohm, J. C. *et al.* The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature* **505,** 546–549 (2014).
28. Li, L., Stoeckert, C. J. Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).
29. Cháb, D., Kolár, J., Olson, M. S. & Storchová, H. Two flowering locus T (*FT*) homologs in *Chenopodium rubrum* differ in expression patterns. *Planta* **228,** 929–940 (2008).
30. Pin, P. A. *et al.* An antagonistic pair of *FT* homologs mediates the control of flowering time in sugar beet. *Science* **330,** 1397–1400 (2010).
31. Pin, P. A. & Nilsson, O. The multifaceted roles of FLOWERING LOCUS T in plant development. *Plant Cell Environ.* **35,** 1742–1755 (2012).
32. Kuljanabhagavad, T., Thongphasuk, P., Chamulitrat, W. & Wink, M. Triterpene saponins from *Chenopodium quinoa* Willd. *Phytochemistry* **69,** 1919–1926 (2008).
33. de Geyter, E., Lambert, E., Geelen, D. & Smagghe, G. Novel advances with plant saponins as natural insecticides to control pest insects. *Pers. Technol.* **1,** 96–105 (2007).
34. Kuljanabhagavad, T. & Wink, M. Biological activities and chemistry of saponins from *Chenopodium quinoa* Willd. *Phytochem. Rev.* **8,** 473–490 (2009).
35. Konishi, Y., Hirano, S., Tsuboi, H. & Wada, M. Distribution of minerals in quinoa (*Chenopodium quinoa* Willd.) seeds. *Biosci. Biotechnol. Biochem.* **68,** 231–234 (2004).
36. Madl, T., Sterk, H., Mittelbach, M. & Rechberger, G. N. Tandem mass spectrometric analysis of a complex triterpene saponin mixture of *Chenopodium quinoa*. *J. Am. Soc. Mass Spectrom.* **17,** 795–806 (2006).
37. Mastebroek, D. H., Limburg, H., Gilles, T. & Marvin, H. J. P. Occurrence of sapogenins in leaves and seeds of quinoa (*Chenopodium quinoa* Willd). *J. Sci. Food Agric.* **80,** 152–156 (2000).
38. Ward, S. M. A recessive allele inhibiting saponin synthesis in two lines of Bolivian quinoa (*Chenopodium quinoa* Willd.). *J. Hered.* **92,** 83–86 (2001).
39. Mertens, J. *et al.* The bHLH transcription factors TSAR1 and TSAR2 regulate triterpene saponin biosynthesis in *Medicago truncatula*. *Plant Physiol.* **170,** 194–210 (2016).
40. Laule, O. *et al.* Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **100,** 6866–6871 (2003).
41. Rodríguez-Concepción, M. & Boronat, A. Breaking new ground in the regulation of the early steps of plant isoprenoid biosynthesis. *Curr. Opin. Plant Biol.* **25,** 17–22 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements Research reported in this publication was supported by the King Abdullah University of Science and Technology (KAUST), by USDA/NIFA-REEIS grant #2012-51300-20100 (WSU/BYU), by NSF grant #1339412, and by a grant from the German Research Foundation, DFG (grant no. JU205/24-1). We thank T. Ramaraj from the National Center for Genomic Resources (Santa Fe) for his bioinformatics assistance, O. Mohammed Eid Alharbi at the KAUST Imaging and Characterization Laboratory for generating SEM images, K. Zemmouri at the KAUST Greenhouse for plant care, and H. Ho (Heno) Hwang and I. D. Gromicho at the KAUST Academic Writing Department for assistance with quinoa illustrations and photographs. We also thank H. Štorchová, F. Morton, D. Bertero, F. Fuentes and D. Brenner of USDA-ARS-NPGS for their assistance in providing seeds. Metabolite imaging was conducted at Metabolomics Australia (School of BioSciences, The University of Melbourne, Australia), a NCRIS initiative under Bioplatforms Australia Pty Ltd.

Author Contributions M.T. and D.E.J. conceived the project. M.T. supervised the research. M.T., D.E.J., Y.S.H., D.J.L., S.M.S., P.J.M., E.N.J., E.N.v.L., C.G.v.d.L. and T.G. conceived and designed the experiments and managed particular components of the project. Y.S.H. led the bioinformatics analyses. Y.S.H., D.J.L., D.E.J. and B.L. did most of the compilation of the genome scaffolds and the genomic analyses. Y.S.H. annotated the genome and undertook the analysis of repetitive elements. E.N.v.L. did the final genetic mapping. E.N.J. and P.J.M. provided germplasm, oversaw sequencing of the diploid genomes, oversaw the BioNano mapping, led the comparative genomics work and shared their knowledge of quinoa. S.M.S. oversaw all saponin-related analyses. C.T.M., A.R.S., T.B., E.S., K.M., E.N.J. and P.J.M. prepared materials and undertook sequencing activities. K.M., H.O. and S.N. oversaw all phylogenetic analyses. Y.H.W. and G.G. analysed the microRNAs. N.S., N.M.K., R.R.R., X.G. and S.A.-B. did saponin analyses. N.D. and C.J. analysed the genes related to flowering time. S.T.A. and M.A.M. did the computational structure–function analysis. B.A.B. and U.R. did the metabolomics imaging. All authors contributed to the writing of the paper. D.E.J., Y.S.H., D.J.L., S.M.S., B.L. and M.T. organised the manuscript. D.E.J. and M.T. coordinated the project. D.E.J., Y.S.H., D.J.L., S.M.S. and B.L. contributed equally.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.T. (mark.tester@kaust.edu.sa).

Reviewer Information *Nature* thanks S. Bak, A. Paterson and N. Stein for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Quinoa sequencing and assembly. We sequenced *Chenopodium quinoa* Willd. (quinoa) accession PI 614886 (BioSample accession code SAMN04338310; also known as NSL 106399 and QQ74). DNA was extracted from leaf and flower tissue of a single plant, 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). *De novo* assembly was conducted using the smrtmake assembly pipeline (https://github.com/PacificBiosciences/smrtmake) and the Celera Assembler, and the draft assembly was polished using the quiver algorithm.

DNA was also sequenced using an Illumina HiSeq 2000 machine. For this, 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 reads were processed with Trimmomatic (v0.33)[42], and reads <75 nucleotides in length after trimming were removed from further analysis. The remaining high-quality reads were assembled with Velvet (v1.2.10)[43] using a *k*-mer of 75.

Integrating BioNano optical maps with the PacBio assembly. High-molecular-weight DNA was isolated and labelled from leaf tissue of three-week old quinoa plants according to standard BioNano protocols, using the single-stranded nicking endonuclease Nt.BspQI. 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. PacBio-BioNano hybrid scaffolds were identified using IrysView's hybrid scaffold alignment subprogram.

Chicago library preparation and sequencing. Using the same DNA prepared for PacBio sequencing, a Chicago library was prepared as described previously[10]. The library was sequenced on an Illumina HiSeq 2500.

Scaffolding the PacBio and BioNano assemblies with HiRise. Chicago sequence data (in FASTQ format) was used to scaffold the PacBio-BioNano hybrid assembly 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.

Kurmi × 0654 population mapping and genetic marker analysis. A population was developed by crossing Kurmi (green, sweet) and 0654 (red, bitter). Homozygous high- and low-saponin F_2 lines were identified by planting 12 F_3 seeds derived from each F_2 line, harvesting F_4 seed from these F_3 plants, and then performing foam tests on the F_4 seed. Phenotyping was validated using gas chromatography/mass spectrometry (GC/MS). RNA was extracted from inflorescences containing a mixture of flowers and seeds at various stages of development from the parents and 45 individual F_3 progeny. RNA extraction and Illumina sequencing were performed as described above. Sequencing reads from all lines were trimmed using Trimmomatic and mapped to the reference assembly using TopHat[44], and SNPs were called using SAMtools mpileup (v1.1)[45].

For linkage mapping, markers were assigned to linkage groups on the basis of the grouping by JoinMap v4.1. 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.

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].

Atlas × Carina Red population mapping and genetic marker analysis. A second mapping population was developed by crossing Atlas (sweet) and Carina Red (bitter). Bitter and sweet F_2 lines were identified by performing foam and taste tests on the F_3 seed. DNA sequencing was performed with DNA from the parents and 94 sweet F_2 lines, as described above, and sequencing reads were mapped to the reference assembly using BWA. SNPs were called in the parents and in a merged file containing all combined F_2 lines.

Genotype calls were generated for the 94 F_2 genotypes by summing up read counts over a sliding window of 500 variants, at all variant positions for which the parents were homozygous and polymorphic. Over each 500-variant stretch, all reads with Atlas alleles were summed, and all reads with the Carina Red allele were summed. Markers were assigned to linkage groups using JoinMap, with regression mapping used to obtain the genetic maps per linkage group.

Integrated linkage map. The Kurmi × 0654 and Atlas × Carina Red maps were integrated with the previously published quinoa linkage map[13], 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 values >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.

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.

Sequencing and assembly of *C. pallidicaule* and *C. suecicum*. DNA was extracted from *C. pallidicaule* (PI 478407) and *C. suecicum* (BYU 1480) and was sent to the Beijing Genomic Institute (BGI, Hong Kong) 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 × 100-bp reads for each library. The generated reads were trimmed using the quality-based trimming tool Sickle (https://github.com/najoshi/sickle). The trimmed reads were then assembled using the ALLPATHS-LG assembler[47], and GapCloser v1.12[48] was used to resolve N spacers and gap lengths produced by the ALLPATHS-LG assembler.

Genome annotation. 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]. RepeatMasker[50] was used to discover and identify repeats within the respective genomes.

AUGUSTUS[51] was used for *ab initio* gene prediction, using model training based on coding sequences from *Amaranthus hypochondriacus*, *Beta vulgaris*, *Spinacia oleracea* and *Arabidopsis thaliana*. RNA-seq and isoform sequencing reads generated from RNA of different tissues were mapped onto the reference genome using Bowtie 2 (ref. 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 isoform sequencing) 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. Genes were characterized 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.

Re-sequencing. The following quinoa accessions were chosen for DNA re-sequencing: 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. All sequencing was performed with an Illumina HiSeq 2000 machine, using either 125-bp (Atlas and Carina Red) or 100-bp (all other accessions) paired-end libraries. Reads were trimmed using Trimmomatic and mapped to the reference assembly using BWA (v0.7.10)[55]. Read alignments were manipulated with SAMtools, and the mpileup function of SAMtools was used to call SNPs.

Identification of orthologous genes. Orthologous and paralogous gene clusters were identified using OrthoMCL[28]. Recommended settings were used for all-against-all BLASTP comparisons (Blast+ v2.3.0[56]) and OrthoMCL analyses. Custom Perl scripts were used to process OrthoMCL outputs for visualization with InteractiVenn[57].

Phylogenetic inference. Using OrthoMCL, orthologous gene sets containing two copies in quinoa and one copy each in *C. pallidicaule*, *C. suecicum*, and

B. vulgaris were identified. 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].

In addition, the genomic variants of all 25 sequenced taxa (Supplementary Data 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 genome (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 which was used as an input into SNPhylo[62] to construct the phylogenetic relationship using maximum likelihood and 1,000 bootstrap iterations.

To identify *FT* homologues, the protein sequence from the *A. thaliana* flowering time gene *FT* was used as a BLAST query. Filtering for hits with an *E* value $<1\times e^{-3}$ and with RNA-seq evidence resulted in the identification of four quinoa proteins. One quinoa protein (AUR62013052) appeared to be comprised of two tandem repeats which were separated for the purposes of phylogenetic analysis. For the construction of the phylogenetic tree, protein sequences from these five quinoa *FT* homologues were aligned using Clustal Omega[63] along with two *B. vulgaris* (gene models: *Bv*FT1-miuf.t1, *Bv*FT2-eewx.t1) and one *A. thaliana* (AT1G65480.1) homologue. 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 neighbour 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 visualized using the CoGE[65] GEvo tool and the Multi-Genome Synteny Viewer[66].

The alignment of bHLH domains was performed with Clustal Omega[63], using sequences from Mertens *et al.*[39]. The phylogeny was inferred using the maximum likelihood method based on the JTT matrix-based model[67]. Initial trees for the heuristic search were obtained automatically by applying Neighbour-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. All positions containing gaps and missing data were eliminated.

Distinguishing and analysing 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. For every base in the quinoa genome, the depth coverage of properly paired reads 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 5× coverage from both diploids. 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. 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 8 kb 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).

Sub-genome synteny was analysed by plotting the positions of homoeologous pairs of A- and B-sub-genome pairs within the context of the 18 chromosomes using Circos. Synteny between the sub-genomes and *B. vulgaris* was assessed by first creating pseudomolecules by concatenating scaffolds which were known to be ordered and oriented within each of the nine chromosomes. Syntenic regions between these *B. vulgaris* chromosomes and those of quinoa were then identified using the recommended settings of the CoGe SynMap tool[70] and visualized using MCScanX[71] and VGSC[72]. For the purposes of visualization, quinoa chromosomes *Cq*B05, *Cq*A08, *Cq*B11, *Cq*A15 and *Cq*B16 were inverted.

Saponin analyses. Quinoa seeds were embedded in a 2% carboxymethylcellulose solution and frozen above liquid nitrogen. Sections of 50 μm thickness were obtained using a Reichert-Jung Frigocut 2800N, modified to use a Feather C35 blade holder and blades at −20 °C using a modified Kawamoto method[73]. A 2,5-dihydroxybenzoic acid (Sigma-Aldrich) matrix (40 mg ml^{-1} in 70% methanol) was applied using a HTX TM-Sprayer (HTX Technologies LLC) with attached LC20-AD HPLC pump (Shimadzu Scientific Instruments). Sections were vacuum dried in a desiccator before analysis. The optical image was generated using an Epson 4400 Flatbed Scanner at 4,800 d.p.i. For mass spectrometric analyses, a Bruker SolariX XR with 7T magnet was used. Images were generated using Bruker Compass FlexImaging 4.1. Data were normalized to the TIC, and brightness optimization was employed to enhance visualization 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 at 400 *m/z*. Lipids were putatively assigned by searching the LipidMaps database[74] (http://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.

Quinoa flowers were marked at anthesis, and seeds were sampled at 12, 16, 20 and 24 days after anthesis. A pool of five 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. Derivatized solution was analysed using single quadrupole GC/MS system (Agilent 7890 GC/5975C MSD) equipped with EI source at ionisation energy of 70 eV. Chromatography separation was performed using DB-5MS fused silica capillary column (30m × 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.

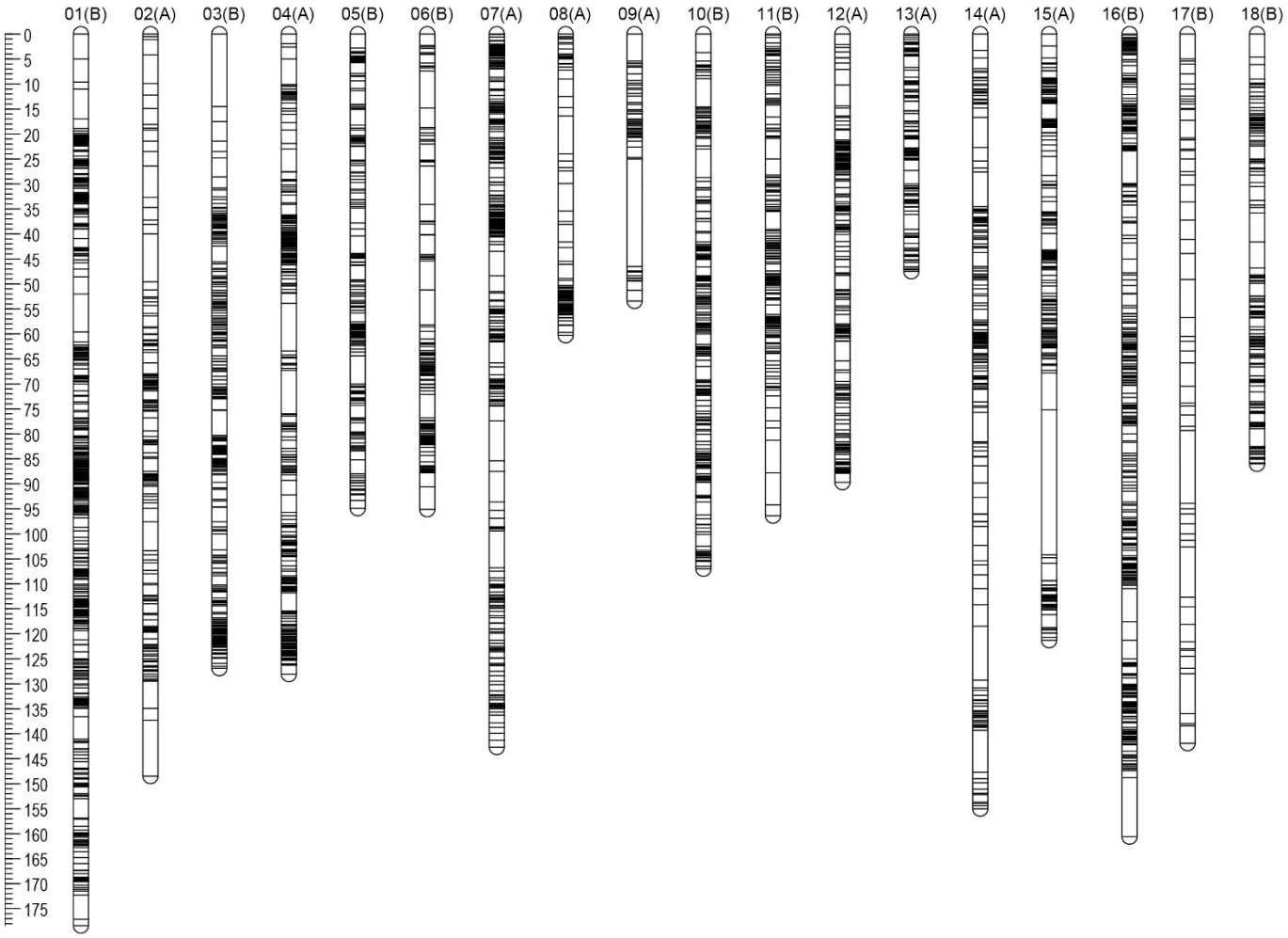
Specific, individual saponins were identified in quinoa using a preparation of 20 mg of seeds performed according a modified protocol from Giavalisco *et al.*[75]. Samples were measured with a Waters ACQUITY Reversed Phase Ultra Performance Liquid Chromatography (RP-UPLC) coupled to a Thermo-Fisher Exactive 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).

SwissModel[76] was used to produce homology models for the bHLH region of AUR62017204, AUR62017206 and AUR62010677. RaptorX[77] was used for prediction of secondary structure and disorder. QUARK[78] was used for *ab initio* modelling of the C-terminal domain, and the DALI server[79] was used for 3D homology searches of this region. Models were manually inspected and evaluated using the PyMOL program (http://pymol.org).

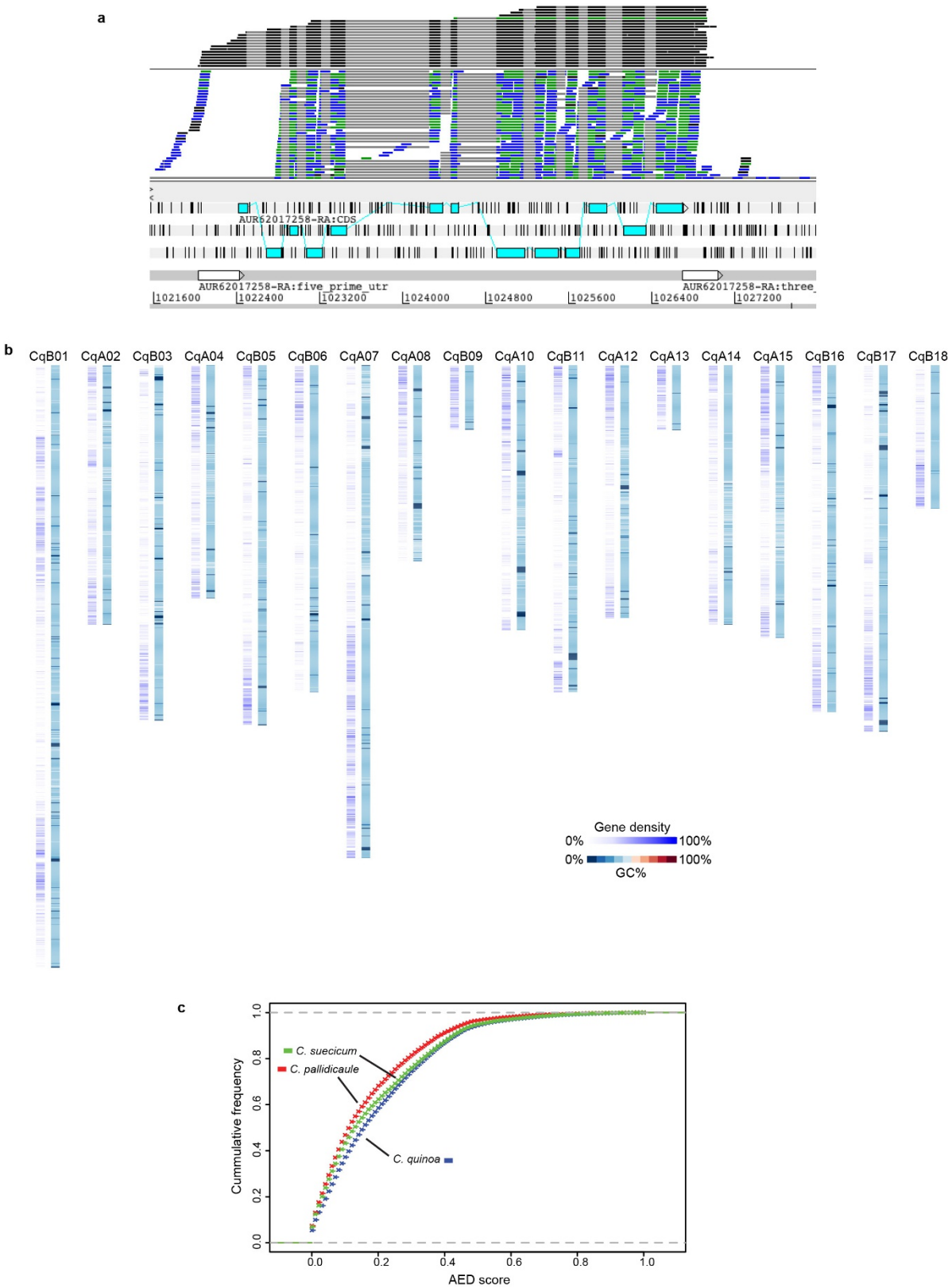
Data availability statement. The genome assemblies and sequence data for *C. quinoa*, *C. pallidicaule* and *C. suecicum* were deposited at NCBI under BioProject codes PRJNA306026, PRJNA326220 and PRJNA326219, respectively. Additional accessions numbers for deposited data can be found in Supplementary Data 9. The quinoa genome can also be accessed at http://www.cbrc.kaust.edu.sa/chenopodiumdb/ and on the Phytozome database (http://www.phytozome.net/).

42. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
43. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
44. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25,** 1105–1111 (2009).
45. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
46. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
47. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA* **108,** 1513–1518 (2011).
48. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1,** 18 (2012).
49. Smit, A. F. A. & Hubley, R. *RepeatModeler Open-1.0*. http://www.repeatmasker.org (2008–2015).
50. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0*. http://www.repeatmasker.org (2013–2015).
51. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24,** 637–644 (2008).
52. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
53. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21,** 1859–1875 (2005).
54. Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18,** 188–196 (2008).

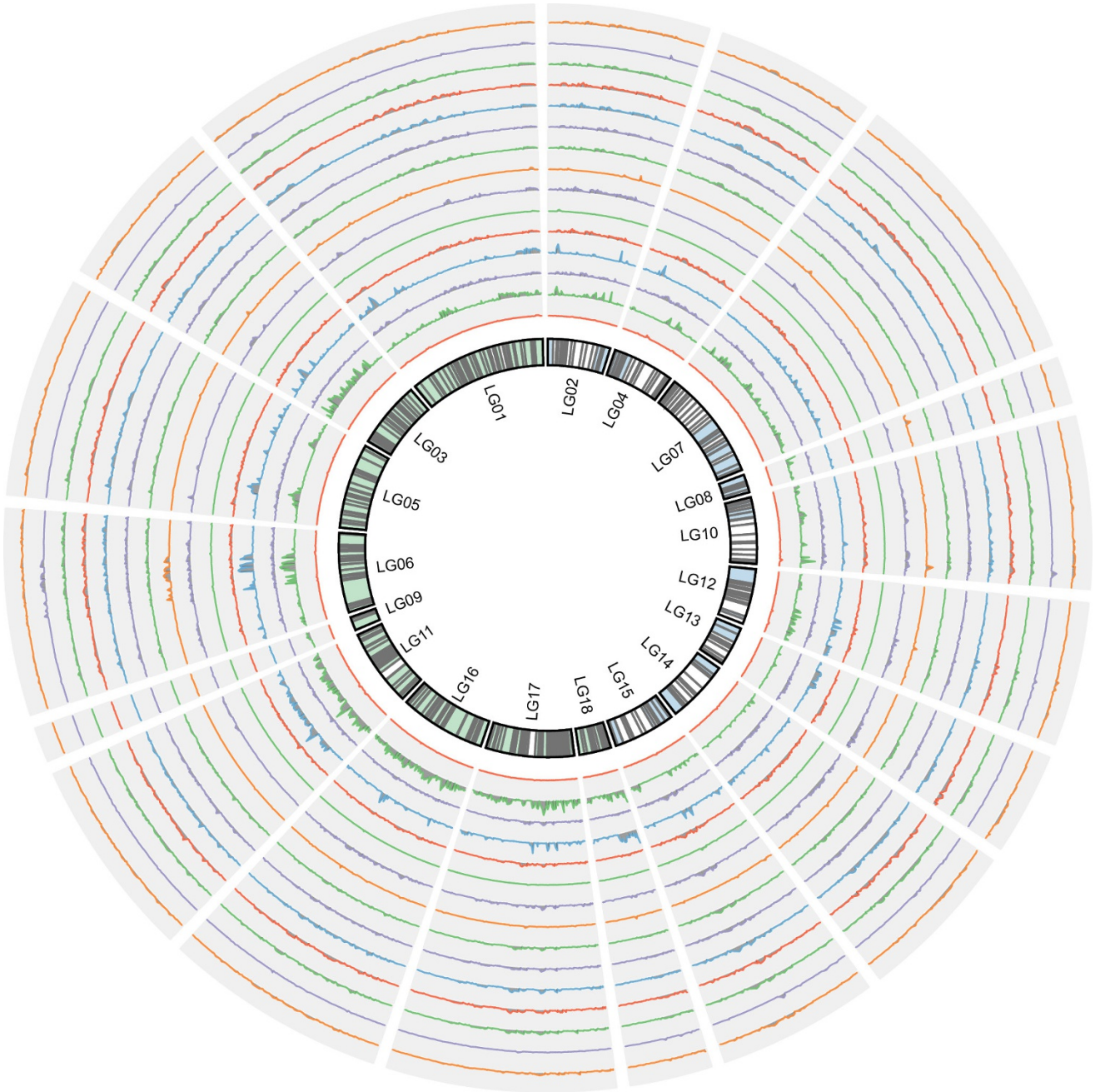
55. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
56. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
57. Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P. & Minghim, R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* **16,** 169 (2015).
58. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
59. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277 (2000).
60. Felsenstein, J. *PHYLIP (Phylogeny Inference Package) Version 3.6a3*. http://evolution.genetics.washington.edu/phylip.html
61. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
62. Lee, T.-H., Guo, H., Wang, X., Kim, C. & Paterson, A. H. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* **15,** 162 (2014).
63. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).
64. 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).
65. Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* **53,** 661–673 (2008).
66. Revanna, K. V. *et al.* A web-based multi-genome synteny viewer for customized data. *BMC Bioinformatics* **13,** 190 (2012).
67. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8,** 275–282 (1992).
68. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
69. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19,** 1639–1645 (2009).
70. Lyons, E., Pedersen, B., Kane, J. & Freeling, M. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Trop. Plant Biol.* **1,** 181–190 (2008).
71. Wang, Y. *et al. MCScanX*: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40,** e49 (2012).
72. Xu, Y. *et al.* VGSC: a web-based vector graph toolkit of genome synteny and collinearity. *BioMed Res. Int.* **2016,** 7823429 (2016).
73. Kawamoto, T. Use of a new adhesive film for the preparation of multi-purpose fresh-frozen sections from hard tissues, whole-animals, insects and plants. *Arch. Histol. Cytol.* **66,** 123–143 (2003).
74. Sud, M. *et al.* LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **35,** D527–D532 (2007).
75. Giavalisco, P., Köhl, K., Hummel, J., Seiwert, B. & Willmitzer, L. ^{13}C isotope-labeled metabolomes allowing for improved compound annotation and relative quantification in liquid chromatography-mass spectrometry-based metabolomic research. *Anal. Chem.* **81,** 6546–6551 (2009).
76. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22,** 195–201 (2006).
77. Källberg, M., Margaryan, G., Wang, S., Ma, J. & Xu, J. RaptorX server: a resource for template-based protein structure modeling. In *Protein Structure Prediction* (ed. Kihara, K.) 17–27 (Springer, New York, 2014).
78. Xu, D. & Zhang, Y. *Ab initio* protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80,** 1715–1735 (2012).
79. Holm, L. & Rosenström, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38,** W545–W549 (2010).



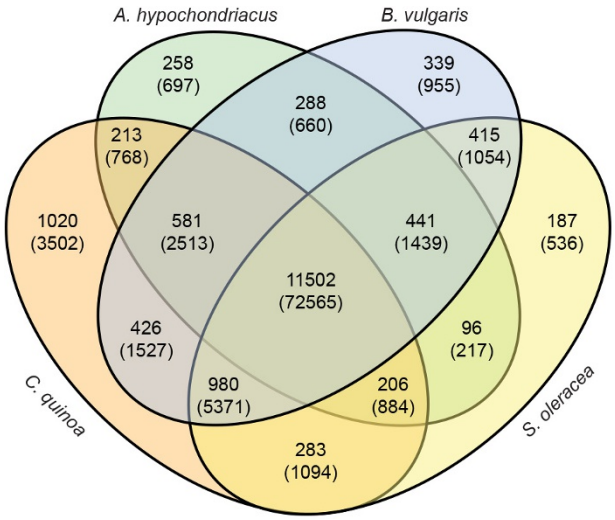
Extended Data Figure 1 | Quinoa linkage map. Linkage map generated by integrating maps from three independent quinoa populations. Black bands in each linkage group represent mapped markers. Letters next to each linkage group name indicate putative assignments of the linkage group to the A or B sub-genome. Numbers of each linkage group correspond to the numbering of chromosomes (for example, LG01(B) is equivalent to *Cq*B01). Scale bar, cM.



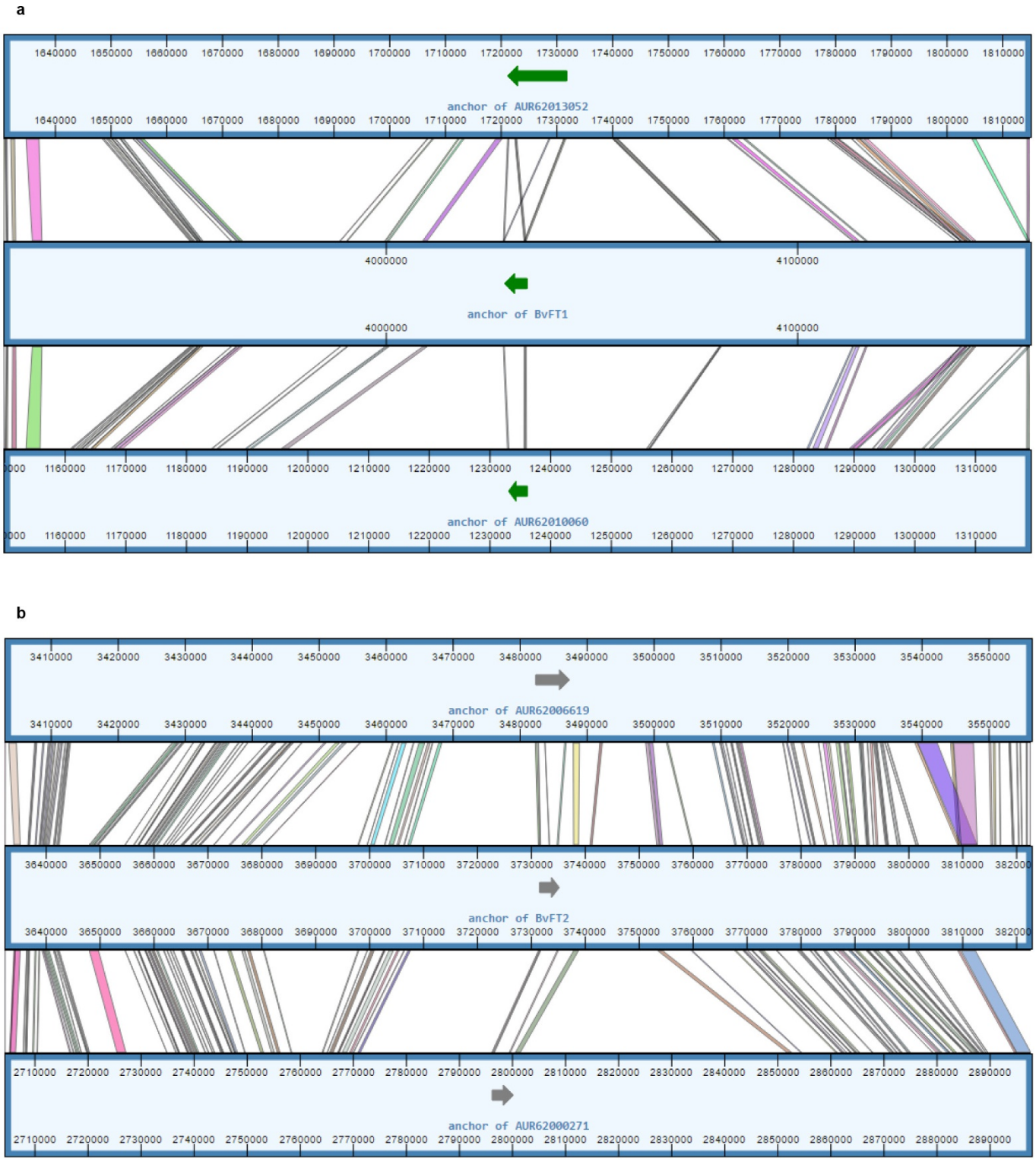
Extended Data Figure 2 | Quinoa genome annotation. a, Representative gene model showing mapped RNA sequencing reads generated using Illumina or isoform sequencing technologies. The top and middle panels show isoform sequencing and RNA-seq reads, respectively, that have been mapped to the chromosomal location containing the AUR62017258 gene model, which is shown on the bottom panel. Light grey lines in the top two panels indicate regions where reads were split to indicate introns positions. Full-length isoform sequencing reads were able to span the 5′ untranslated region, all exons, and the 3′ untranslated region in a single read. **b**, Gene density and GC% in 100-kb windows in quinoa chromosomes. **c**, The frequency of annotation edit distance (AED) scores for the assemblies of quinoa (blue), *C. pallidicaule* (red) and *C. suecicum* (green).



Extended Data Figure 3 | Distribution of SNPs in the quinoa genome. Frequency of SNPs in the sequenced quinoa accessions, relative to the reference quinoa genome assembly, in a 1-Mb window size. *y* axis scale is from 0 to 10,000 SNPs. The innermost track shows scaffolds arranged according to their placement in the linkage groups, with scaffolds coloured according to sub-genome assignment based on mapping sequencing reads from *C. pallidicaule* and *C. suecicum*, as in Fig. 2a. From inside to outside, the remaining tracks show SNPs in PI 634921, Atlas, CICA-17, Carina Red, Cherry Vanilla, Chucapaca, G-205-95DK, Ku-2, Kurmi, 0654, Ollague, Pasankalla, Real, Regalona and Salcedo INIA.

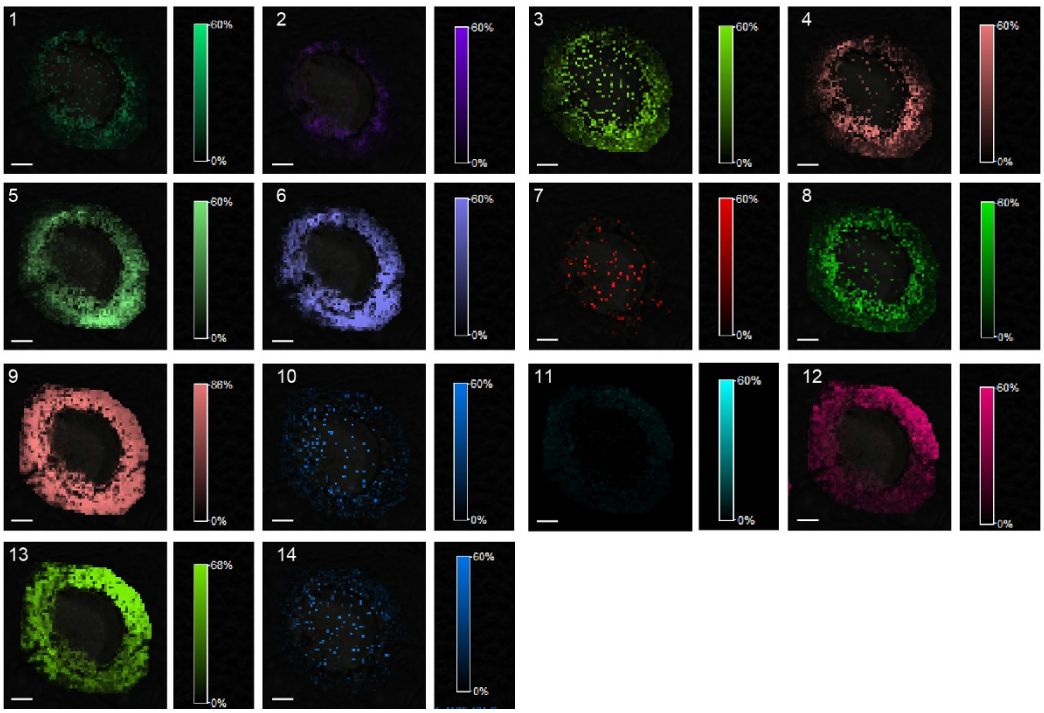


Extended Data Figure 4 | Orthologous genes in quinoa, *Amaranthus hypochondriacus*, *Beta vulgaris* and *Spinacia oleracea*. Venn diagram representing the number of protein-coding gene clusters shared between, or distinct to, the indicated species. The number in parentheses indicates the total number of genes contained within the associated clusters.

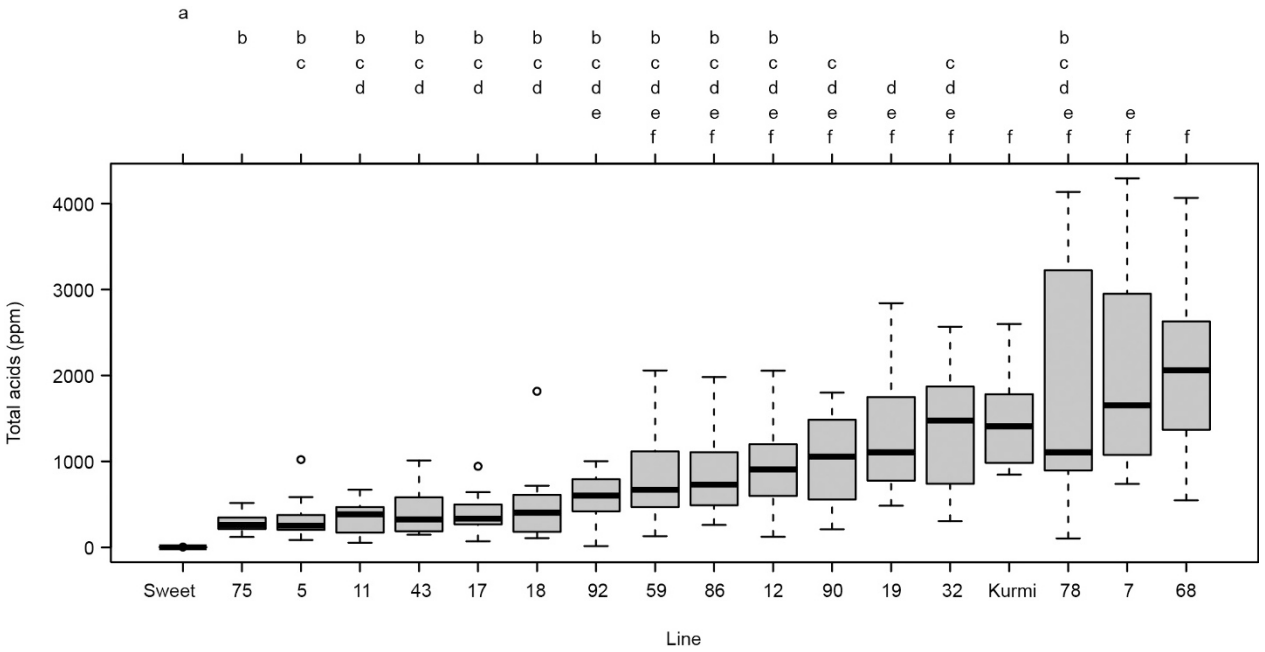


Extended Data Figure 5 | Syntenic relationships supporting the orthologous relationships between quinoa and *Beta vulgaris FT* genes. a, Chromosomal regions containing Bv*FT1* and the quinoa genes AUR62013052 (Cq*FT1B*) and AUR62010060 (Cq*FT1A*). The annotation for AUR62013052 appears to contain two duplicate genes, which we have separated and designated as Cq*FT1B-1* and Cq*FT1B-2*. **b**, Chromosomal regions containing Bv*FT2* and the quinoa genes AUR62000271 (Cq*FT2A*) and AUR62006619 (Cq*FT2B*). Green and grey arrows represent *FT1* and *FT2* genes, respectively, and coloured lines connect orthologous coding sequences surrounding the *FT* genes.

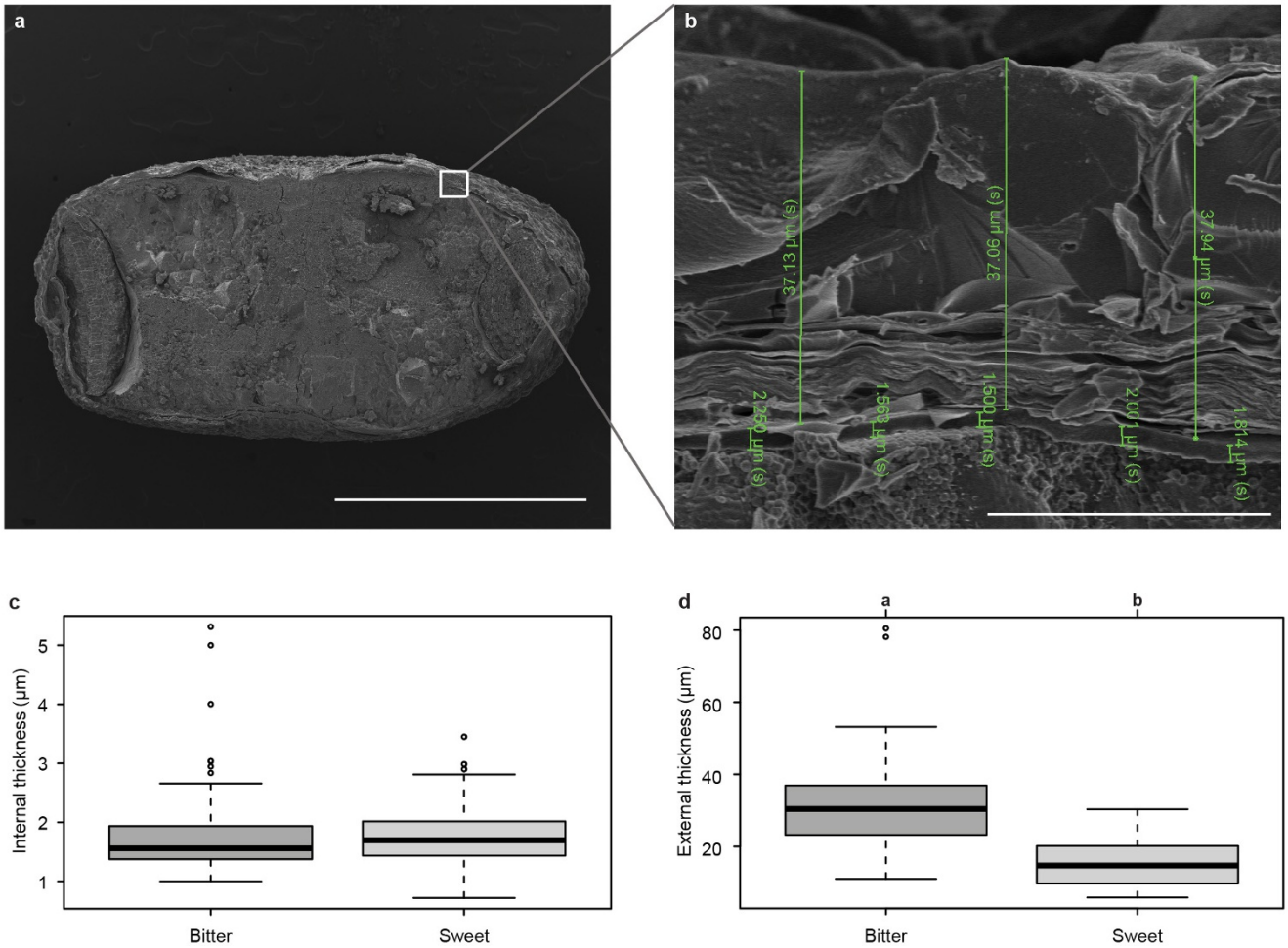
ID	Saponin[a]	Formula	Calculated	Observed	Error (ppm)
1	PA(unknown) +Na	[C36H56O10+Na]+	671.37657	671.37080	-8.6
2	PA(unknown) +K	[C36H56O10+K]+	687.35051	687.34390	-9.6
3	AG487(Pent) +Na	[C41H64O14+Na]+	803.41883	803.41090	-9.9
4	AG487(Pent) +K	[C41H64O14+K]+	819.39277	819.38620	-8.0
5	PA(Pent) +Na	[C42H66O15+Na]+	833.42939	833.42941	0.0
6	PA(Pent) +K	[C42H66O15+K]+	849.40333	849.40366	0.4
7	OA(Pent-Hex)[b] +Na	[C47H76O17+Na]+	935.49747	935.49550	-2.1
8	OA(Pent-Hex)[b] +K	[C47H76O17+K]+	951.47141	951.47800	6.9
9	PA(Hex-Pent)x2 +Na	[C48H76O20+Na]+	995.48222	995.48150	-0.7
10	AG515(HexA-Hex-Pent) +Na	[C47H68O22+Na]+	1007.40945	1007.40000	-9.4
11	Hed(Hex-Hex-Pent) +K	[C53H86O23+K]+	1129.51915	1129.51100	-7.2
12	PA((Hex-Pent-Hex)/(Hex-Hex-Pent)[b] +Na	[C54H86O25+Na]+	1157.53504	1157.53340	-1.4
13	PA((Hex-Pent-Hex)/(Hex-Hex-Pent)[b] +K	[C54H86O25+K]+	1173.50898	1173.50820	-0.7
14	AG489(Hex-Hex-HexA-) +K	[C54H86O26+K]+	1173.52995	1173.52580	-3.5



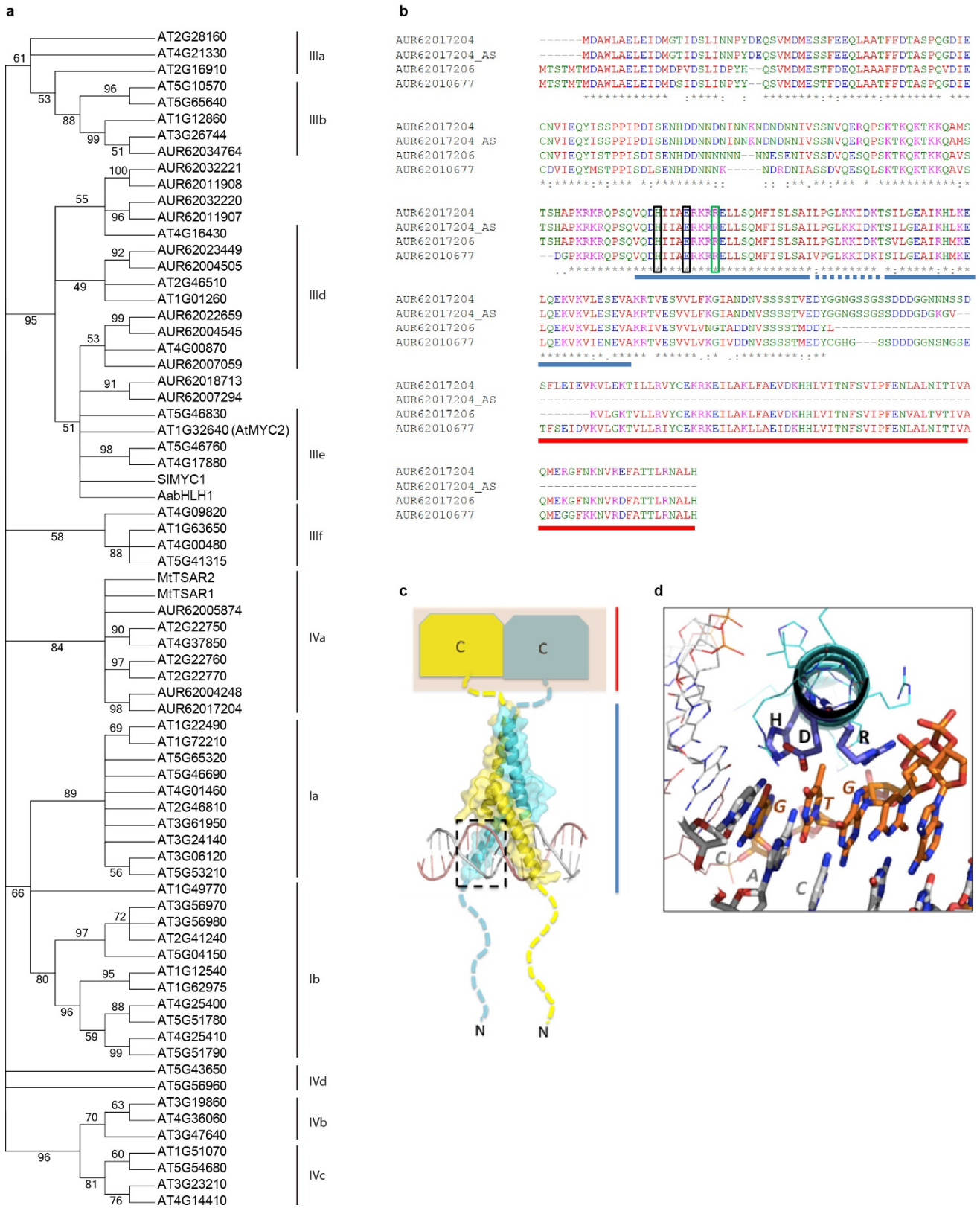
Extended Data Figure 6 | Imaging mass spectrometry visualization of saponins in the pericarp of a quinoa seed. Table lists saponin annotation. [a]PA, phytolaccagenic acid; Hed, hederagenin; SA, serjanic acid; OA, oleanolic acid; AG489, AG515, AG487 refer to new aglycones with a specific *m/z*. Pen, pentose; Hex, hexose; HexA, corresponding sugar acid. [b]Pairs of saponins with the same chemical formula but different retention times. Image label indicates saponin ID. Scale bars, 500 μm.



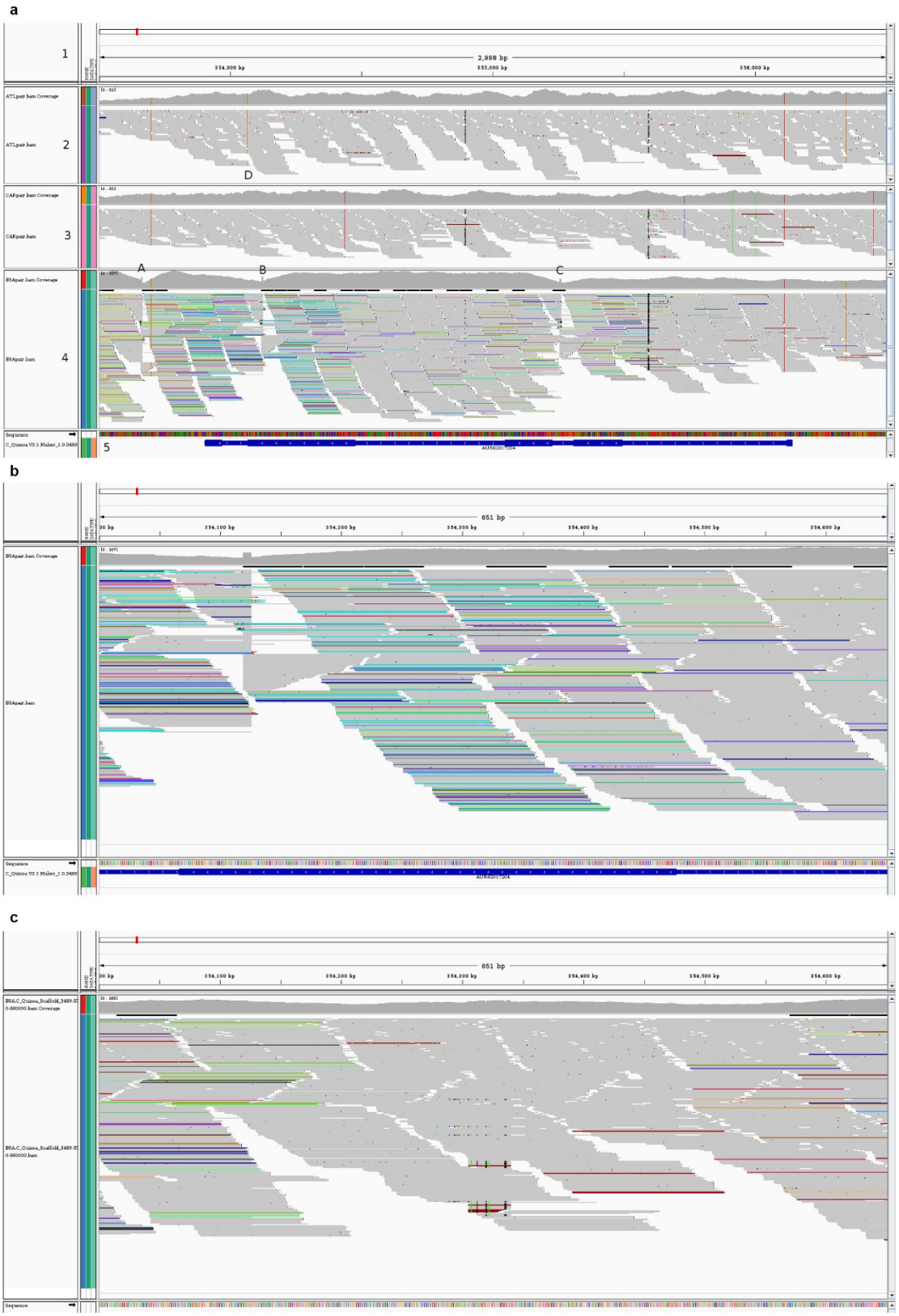
Extended Data Figure 7 | Saponin content in 0654 × Kurmi population. Twelve individual seeds of 17 bitter lines, and 12 pooled seeds from each of 16 sweet lines were analysed for derivatized saponins using gas chromatography/mass spectrometry. Data for sweet lines, including parent 0654, were consolidated into one box plot (Sweet). Box plots show median values (solid horizontal lines), 25th and 75th percentile values (box), 90th percentile values (whiskers) and outlier values (open circles). Quantification was performed using standards of oleanolic acid. Letters above each box plot represent statistically significant ($P < 0.05$) differences between groups based on Games Howell post hoc test.



Extended Data Figure 8 | Variation in the thickness of seed coat layers in bitter and sweet lines. a, Representative scanning electron microscopy image of a quinoa seed cross-section, showing an example of a region (white box) from which measurements were taken. Scale bar, 1 mm. **b**, Representative scanning electron microscopy image showing measurements of inner and outer seed coat layers. Scale bar, 30 µm **c**, Thickness of the internal seed coat layer in bitter and sweet lines. **d**, Thickness of the external seed coat layer in bitter and sweet lines. Letters above each box plot represent statistically significant ($P < 0.001$) differences between groups based on ANOVA. Box plots show median values (solid horizontal lines), 25th and 75th percentile values (box), 90th percentile values (whiskers) and outlier values (open circles). $n = 91$, 160, 54, and 129 for internal bitter, internal sweet, external bitter, and external sweet, respectively.



Extended Data Figure 9 | Prediction of TSARL1 function using phylogenetics and computational modelling. a, Maximum likelihood tree of select bHLH peptide sequences from quinoa, *Arabidopsis thaliana* and *Medicago truncatula*, showing the close evolutionary relationship between the quinoa bHLH TSARL1 (AUR62017204) and the *M. truncatula* bHLHs TSAR1 and TSAR2 (MEDTR7G080780 and MEDTR4G066460, respectively). Branch values represent the percentage of 500 bootstrap replicates that support the topology. Subclades of the bHLH family, as defined in *A. thaliana*, are indicated on the right. **b**, Sequence alignment of TSARL bHLH sequences. Underlined blue, bHLH homology domain (solid line, helix; dashed line, loop). Underlined red, C-terminal domain. Boxed, residues that confer specificity to E box DNA binding. Green boxed Arg, residue that selects for the central CG dinucleotide. AUR62017204_AS designates the alternatively spliced protein. **c**, Schematic drawing of full-length TSARL1 (AUR62017204). Dashed lines indicate regions predicted to be flexible. Boxed region shows C-terminal domain lost in alternative splice variant. **d**, Zoomed-in view of the black dashed box in **c**, showing TSARL1 specifically binding the CACGHG motif.



Extended Data Figure 10 | Structural variation in *TSARL1* in the sweet progeny of Atlas and Carina Red. a, Screenshot from Integrative Genomics Viewer (IGV) showing read alignment results in a 3-kb region (track 1, top) around *TSARL1* (track 5, bottom). Shown are alignments for Atlas (track 2), Carina Red (track 3), and the merged F_2 sweet lines (track 4). In tracks 2–4, the top portion shows coverage, and the bottom portion shows individual reads. The F_2 merged data (track 4) shows evidence of three structural variants (labelled A, B and C) relative to the reference, whereas Atlas (track 2) only shows evidence of the G2078C SNP (labelled D). **b**, **c**, Screenshots from IGV showing before (**a**) and after (**b**) local re-assembly of the reference sequence to include the B insertion site shown in panel **a**, illustrating the effect on read mapping from the merged F_2 sweet lines. The dips in coverage and discordantly mapped reads (indicated by colours assigned to the reads) around this insertion site are resolved after re-assembly.

CORRECTIONS & AMENDMENTS

CORRIGENDUM

doi:10.1038/nature22384

Corrigendum: The genome of *Chenopodium quinoa*

David E. Jarvis, Yung Shwen Ho, Damien J. Lightfoot, Sandra M. Schmöckel, Bo Li, Theo J. A. Borm, Hajime Ohyanagi, Katsuhiko Mineta, Craig T. Michell, Noha Saber, Najeh M. Kharbatia, Ryan R. Rupper, Aaron R. Sharp, Nadine Dally, Berin A. Boughton, Yong H. Woo, Ge Gao, Elio G. W. M. Schijlen, Xiujie Guo, Afaque A. Momin, Sónia Negrão, Salim Al-Babili, Christoph Gehring, Ute Roessner, Christian Jung, Kevin Murphy, Stefan T. Arold, Takashi Gojobori, C. Gerard van der Linden, Eibertus N. van Loo, Eric N. Jellen, Peter J. Maughan & Mark Tester

Nature **542,** 307–312 (2017); doi:10.1038/nature21370

The Acknowledgements section of this Article should have included the following sentence: "Metabolite imaging was conducted at Metabolomics Australia (School of BioSciences, The University of Melbourne, Australia), a NCRIS initiative under Bioplatforms Australia Pty Ltd.". The original Article has been corrected online.