Title

The genome of *Chenopodium quinoa*

Authors

David E. Jarvis*, Yung Shwen Ho*, Damien J. Lightfoot*, Sandra M. Schmöckel*, Bo Li*, Theo 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 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

*These authors contributed equally to this work

Summary

Chenopodium quinoa (quinoa) is a highly nutritious grain crop with high abiotic stress tolerance that has been 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 and assembly of two diploids from among the ancestral gene pools of quinoa – which enabled the identification of sub-genomes in quinoa – and reduced-coverage genome sequences for 22 other accessions 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 the inclusion of a premature stop codon, thereby inactivating the protein and leading to the absence of saponins in sweet quinoa accessions. These genomic resources are an important first step towards the genetic improvement of quinoa to help increase global food security in the face of climate change and a growing world population.

Almost 800 million people are undernourished, in part due to limited food production on marginal lands[1]. These poor agricultural environments are often disproportionately affected by climate change, pressure from population growth, and intensification of agriculture, all of which exacerbate the challenge of producing sufficient and nutritious food[2]. Meeting the caloric and nutritional demands of these growing populations will not only require increases in overall food production, but also the development of new crops that can be grown sustainably in agricultural environments that are increasingly susceptible to degradation.

Quinoa *(Chenopodium quinoa* Willd., 2n = 4x = 36) is a highly nutritious crop that is adapted to thrive in a wide range of agroecosystems. The centre of diversity of quinoa is near Lake Titicaca on the border of Bolivia and Peru, where it has adapted to the arid and saline soils of the high plains of the Andean Altiplano (>3,500 meters above sea level). Quinoa was presumably first domesticated more than 7,000 years ago by Pre-Columbian cultures and was known as the "Mother Grain" of Incan Empire[3]. Quinoa was able to spread with pre-Incan cultures due to its adaptability to diverse environments, such as the Mediterranean climates of coastal Chile as well as the intermediate highland valleys of Peru and Ecuador. Abiotic stresses to which it is adapted include soil salinity[4,5] , frost, high UV irradiance, and drought[6] . Recently, quinoa has gained international attention because of the nutritional value of its seeds, which are gluten-free, have a low glycaemic index[7] , and contain an excellent balance of essential amino acids, fibre, lipids, carbohydrates, vitamins, and minerals[8] . Thus, 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 recognised when the United Nations declared 2013 as the International Year of Quinoa (http://www.fao.org/quinoa-2013/what-is-quinoa/nutritional-value/en/), this being one of only three times plants have received such a designation.

Despite its agronomic potential, quinoa is still an underutilised crop[9], with relatively few active breeding programs[10]. Breeding efforts to improve the crop for important agronomic traits, including tolerance to heat and biotic stress, are needed to expand quinoa production worldwide and to relieve stress on Andean production environments strained by intensive cultivation and the effects of climate change. To accelerate the improvement of quinoa, we present here the allotetraploid quinoa genome, which was sequenced and assembled using long-read sequencing technology in combination with optical, chromosome contact, and genetic maps. We demonstrate the utility of the genome sequence by identifying genes that underlie the production of betalain pigments and seed triterpenoid saponin content. Moreover, we sequenced the genomes of additional diploid and tetraploid *Chenopodium* species to characterise 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 quinoa, 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 using single-molecule real-time (SMRT) sequencing technology from Pacific Biosciences (PacBio). After removal of contaminant and chloroplast sequences, the assembly consisted of 4,232 contigs, with a contig N50 of 1.66 megabases (Mb) (Table 1). To further improve this assembly, we generated genome maps using optical mapping technology[11] from BioNano Genomics and chromosome contact from Dovetail Genomics[12]. Scaffolding of the PacBio, BioNano, and Dovetail assemblies produced a final assembly consisting of 3,483 scaffolds, with a scaffold N50 of 3.84 Mb and 90% of the assembled genome contained in 439 scaffolds. 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[13,14]). To combine scaffolds into

pseudomolecules, an existing linkage map from quinoa[15] was integrated with two new linkage maps generated from independent populations. The resulting map (Extended Data Fig. 1) of 6,403 unique markers spanned a total length of 2,034 centimorgans (cM) and consisted of 18 linkage groups (LGs) (Supplementary Table 7), corresponding to the haploid chromosome number of quinoa; we refer to the LGs hereafter as chromosomes, which are numbered based on numbering from a previously published SNP linkage map[15]. The mapped markers came from 565 scaffolds, representing 1.18 Gb (85%) of the total assembly length (Table 1, Supplementary File 1, Supplementary File 2).

Predicted protein-coding and miRNA 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 Iso-Seq approaches. The inclusion of full-length reads from Iso-Seq helped resolve several gene models (Extended Data Fig. 2a). The annotation contains 44,776 gene models (Supplementary Table 2, Extended Data Fig. 2b), including 33,365 genes with annotation edit distance (AED)[16,17] values ≤ 0.3 (Extended Data Fig. 2c). Although the number of annotated genes is greater than that of most sequenced diploid species, it is in line with sequenced tetraploid species[18] and is therefore likely a result of the presence of two sub-genomes in quinoa. Sixty-four percent of the genome was found to be repetitive, including a large proportion of long terminal repeat (LTR) transposable elements (Supplementary Table 1). To assess the completeness of the genome annotation, we evaluated the coverage of a set of 956 highly conserved core eukaryotic genes found in a wide range of plant taxa[19]. We identified 97.3% of the genes in the Plantae BUSCO dataset (Supplementary Table 3), which is suggestive of a complete assembly and annotation. Eighty-seven percent of the single-copy orthologs used in the BUSCO analysis were duplicated in the quinoa genome, which is likely a reflection of the tetraploid nature of the quinoa genome.

The utility of the assembly, linkage maps, and annotation was demonstrated by mapping the betalain locus, which controls stem pigmentation in one of the segregating mapping populations and is often used as a morphological marker in breeding programs. 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 (CqA02), 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[20]. 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)[21], 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). Fine mapping is now required to enable confident identification of the causative mutation and separation of primary and pleiotropic effects.

Evolutionary history of quinoa

Quinoa is an allotetraploid that resulted from the hybridisation of ancestral A- and B-genome diploid species[22]. Single-gene sequencing studies previously identified pools of North American and Eurasian diploids, respectively, as candidate sources of the A and B sub-genomes[23–25], with hybridisation occurring somewhere in North America, in a scenario similar to that of cotton[26]. To better understand genome structure and evolution in quinoa, we sequenced, assembled, and annotated the A-genome diploid *C. pallidicaule* and the B-genome diploid *C. suecicum*[24] (Fig. 1a, Table 1). We note that *C.*

pallidicaule (commonly called cañahua or kañiwa) is itself a regionally important and highly nutritious alternative crop species native to the Andean Altiplano. A high proportion of orthologous gene pairs in quinoa showed similar rates of synonymous substitutions per synonymous site (Ks), indicative of a whole-genome duplication event (Fig. 1b). This most likely represents the hybridisation 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*[27] and for core eukaryotes[28], we estimate the tetraploidisation to have occurred 3.3 – 6.3 million years ago.

Multiple tetraploid species – or ecotypes of the biological species complex – have arisen from the ancestral tetraploid following hybridisation, including *C. berlandieri* and *C. hircinum*, which, together with quinoa, are commonly referred to as the allotetraploid goosefoot complex (ATGC). Long-range dispersal of the ancestral weedy ATGC eventually resulted in the appearance and domestication of quinoa in the Lake Titicaca Basin over 7,000 years ago[3,29]. Five ecotypes of quinoa have been previously described (Altiplano, Coastal, Salares, Inter-Andean Valley, and Yungas)[30], although quinoa is most commonly grouped simply into highland and coastal types[29]. In addition to quinoa, the ATGC was the source of at least three other cultigens: Mesoamerican vegetable *huauzontle* and pseudocereal *chia roja* (*C. berlandieri* ssp. *nuttaliae*[31]); and the extinct staple pseudocereal of the North American Hopewell Culture, *C. berlandieri* ssp. *jonesianum*[32]. The evolutionary relationships among the ATGC taxa, including quinoa, remain unclear[29]. To begin to resolve these issues, we re-sequenced 15 additional quinoa accessions. Together with the reference accession, PI 614886, these 16 sequenced quinoa accessions represent four major cultivated ecotypes (Supplementary File 5). We also sequenced five accessions of *C. berlandieri* representing four of its major ecotypes and one accession each of *C. hircinum* from the Pacific and Atlantic Andean watersheds (Supplementary File 5). Phylogenetic analysis of these taxa suggests that *C. hircinum* and quinoa have a parent-derivative relationship, and that North American *C. berlandieri* is the basal member of the ATGC with Mesoamerican *huauzontle* as an independent domesticate (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* accession 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 of *C. hircinum* with highland and coastal quinoa ecotypes, as well as provide germplasm for breeding broadly adapted coastal quinoa cultivars for warm-season production. The single-nucleotide polymorphisms (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

The availability of the *C. pallidicaule* and *C. suecicum* genome sequences enabled the identification of chromosomes belonging to the A and B sub-genomes of quinoa. 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 scaffolds (totalling 202.6 and 646.3 Mb) were assigned to the A and B sub-genomes, respectively (Fig. 2a, Supplementary File 6). A mini-satellite repeat (18-24J) previously shown to be more abundant in the B sub-genome of quinoa and related species[33] is over-represented in scaffolds assigned to the B sub-genome (Supplementary File 6). Nine chromosomes were assigned to each sub-genome based on the sub-genome assignments of scaffolds contained in each chromosome (chromosomes hereafter designated as CqA or CqB, followed by the chromosome number). On a chromosome level, the B sub-genome accounted for a larger percentage of

both the genetic (1087 cM) and physical (696 Mb) sizes of the quinoa genome than the A sub-genome (946 cM, 488 Mb). This result was not unexpected, given the differences in the estimated genome sizes of *C. pallidicaule* (452 Mb) and *C. suecicum* (815 Mb) were predicted based on k-mer analyses.

A total of 5,807 orthologous gene pairs for which one gene was phylogenetically more closely related to an ortholog from *C. pallidicaule* and the other was phylogenetically more closely related to an ortholog from *C. suecicum* were used to elucidate homoeologous relationships among the quinoa chromosomes. Visualization of the chromosomal locations of these homoeologous gene pairs revealed a high degree of synteny between the A and B sub-genomes (Fig. 2b). Interestingly, a small number of gene pairs (5.7%) mapped within the same sub-genome, suggesting that recombination and chromosomal rearrangement 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 CqB05 and CqB03. The genes in the region of ~54-56 Mb of CqB03 are phylogenetically more similar to the A-genome diploid *C. suecicum* and therefore likely originated from the A sub-genome chromosome CqA12 (Fig. 2c). In addition to exchanges between sub-genomes, we also see evidence of large chromosomal rearrangements within sub-genomes, complicating the assignment of homoeologous chromosome pairs. For example, ortholog analysis clearly identifies CqB05 and CqA12 as homoeologous, although the same analysis is much more complicated with other chromosomes, including between CqB01, CqA02, and CqA04, and between CqA07, CqB11, and CqB17 (Fig. 2b). To clarify these relationships, we identified syntenic regions between chromosomes of the diploid *B. vulgaris*[34] (*n* = 9) and the A and B sub-genome chromosomes of quinoa (Fig. 2d). These results indicate that CqA02 and CqA04 are orthologous to *B. vulgaris* chromosomes 8 and 2 (Bvchr8 and Bvchr2), respectively, whereas CqB01 appears to be the result of a chromosome fusion. Likewise, CqA07 appears to be the result of a fusion between ancestral chromosomes orthologous to Bvchr3 and Bvchr7.

Analysis of sub-genome content

Duplicated genes are often lost following genome duplication events[35]. We used OrthoMCL[36] to identify clusters of orthologous genes in related species in the 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). To identify genes lost from either the A or B sub-genome of quinoa, we identified orthologous sets of genes for which only one copy could be found in quinoa and in each of the diploid species. Of these, we found a similar number (1,031 and 849) of genes lost from the A and B sub-genomes, respectively (Fig. 3b). To identify duplicated genes retained as single-copy in each of the quinoa sub-genomes, we identified sets of genes for which one copy was found in each of the diploids and two copies were found in quinoa. Of these, 5,807 sets contained one quinoa gene assigned to each of the A and B sub-genomes (Fig. 3b, Fig. 2b), thus representing a core set of single-copy genes retained in each genome and sub-genome. We also investigated genes retained in multiple copies. Notably, we found that quinoa, like *B. vulgaris*[37], contains two genes that are orthologous to the *Arabidopsis thaliana FT* flowering-time regulator gene, with quinoa containing two homoeologous copies of each ortholog due to its tetraploid nature (Fig. 3c, Extended Data Fig. 5). FT is known to promote flowering in *A. thaliana*, and functional orthologs have been found in other species[38]; however, *B. vulgaris*, which requires vernalisation to induce the transition to flowering, was found to contain a second *FT* gene that acts antagonistically by repressing flowering and whose expression is reduced following vernalisation[37]. In quinoa, one pair of homoeologous *FT* genes was found to be orthologous to the *FT* gene that promotes flowering in *B. vulgaris*, and a second pair of homoeologous genes was found to be orthologous to the gene that represses flowering in *B. vulgaris*. The identification of two pairs of *FT* orthologs in quinoa, as well as the previous identification of two *FT* genes in *C. rubrum*[39], suggests that the common ancestor of quinoa and *B. vulgaris* also had two

FT genes. Given the lack of a vernalisation requirement to induce flowering in quinoa, it is possible that both *FT* genes have retained the normal *FT* function in quinoa, while the second *FT* gene has acquired a repressive function in *B. vulgaris* following its divergence from quinoa. Future functional studies will help address these questions, which are important for many aspects determining yield and yield maintenance under abiotic stress.

Mechanisms underlying saponin production

Quinoa seeds contain a mixture of triterpene glycosides called saponins[40]. Although saponins may be beneficial for plant growth – for example, by deterring herbivory[41,42] – they are also haemolytic and produce a bitter flavour in quinoa seeds and must therefore be removed before human consumption. Because this process is costly, is often water-intensive, and can reduce the nutritional value of the seeds[43], the development of saponin-free lines is a major breeding objective in quinoa[10].

Using imaging mass spectrometry (MS)[44], we show that saponins accumulate in the seed pericarp (Fig. 4a, 4b, Extended Data Fig. 6), which is maternal tissue derived from the ovary wall. We found that saponins accumulate in the seed early in seed development, between 20 and 24 days after anthesis (Fig. 4c), eventually accounting for 4% (w/w) of the mature seed mass (Supplementary Information 8.1.).

Triterpenoid saponins are synthesised as part of the mevalonic acid-dependent pathway, which also produces phytosterols[45]. They are comprised of an aglycone backbone to which various sugar side chains can be attached, including glucose, galactose, and arabinose. Almost 100 different saponins have been identified in different accessions of quinoa[40,46]. Using liquid chromatography–MS (LC-MS), we identified and annotated 47 different saponins in the reference accession (Supplementary Table 9).

Naturally-occurring sweet quinoa accessions that contain very low levels of saponins are present within the quinoa germplasm[47], 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[48], segregation ratios in these populations indicated that the presence and absence 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 qualitative differences exist in the types of 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).

To identify the gene controlling the absence of saponins in these populations, complementary genotyping analyses were performed, including RNA-Seq using RNA extracted from clusters of flowers and seeds of the parents and of 14 bitter and 15 sweet homozygous F3 lines (Kurmi × 0654 population; Extended Data Fig. 7), and DNA-Seq performed using DNA extracted from leaves of the parents and 94 sweet F2 lines (Atlas × Carina Red population). Linkage mapping and BSA in each population identified the same region on CqB16 that distinguishes the bitter and sweet lines (Fig. 5a). Frequencies of the sweet allele in both populations reached 100% for markers located in CqB16 on Scaffold 3489. We investigated the genes in a 700-kb window surrounding this region of 100% sweet allele frequency. Of the 54 annotated genes in this region (Supplementary File 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 triterpene saponin biosynthesis in

Medicago trancatula[49]. In *M. truncatula*, overexpression of the TRITERPENE SAPONIN BIOSYNTHESIS ACTIVATING REGULATOR 1 (TSAR1) and TSAR 2 bHLH transcription factors was recently shown to increase expression of genes in the saponin biosynthetic pathway, resulting in increased accumulation of triterpene saponins[49]. The overexpression of *TSAR1* and *TSAR2* elicited different patterns of transactivation of downstream saponin biosynthetic genes in *M. truncatula*, suggesting distinct functionalities within the triterpenoid biosynthetic pathway, including the accumulation of different classes of saponins[49]. Similarly, our RNA-Seq data also shows distinct expression patterns for these two genes. Specifically, AUR62017206 (hereafter *C. quinoa TSAR-LIKE2, CqTSARL2*) was expressed in root tissue but not in flowers or immature seeds, whereas AUR62017204 (hereafter *C. quinoa TSAR-LIKE1, CqTSARL1*) was almost exclusively expressed in seeds, with significantly lower expression levels in sweet lines (Supplementary File 8). Together, these results suggest that CqTSARL1 might be a functional TSAR ortholog, although whether this is due to shared ancestry or convergent evolution is unclear.

In *M. truncatula*, TSAR1 and TSAR2 were found to bind to the DNA motif 5'-CACGHG-3' (where H can be A, C, or T)[49]. We identified this motif within 2 kb upstream of the start codon in several saponin biosynthetic pathway genes in quinoa, including genes encoding hydroxymethylglutaryl (HMG) CoA synthase, HMG CoA reductase, mevalonate kinase, mevalonate-5-pyrophosphate decarboxylase, isopentenyl diphosphate isomerase, farnesyl pyrophosphate synthase, squalene synthase, and squalene epoxidase (Fig. 5b). Expression levels of these genes – as well as several other genes in the saponin biosynthetic pathway – were significantly downregulated in sweet lines (Fig. 5b, Supplementary File 8).

Analysis of RNA-Seq reads in the sweet progeny of Kurmi and 0654 revealed that the *CqTSARL1* transcript was alternatively spliced in exon 3 due to a SNP (G2078C) in the last position of the exon, thereby altering the canonical intron/exon splice boundary (Fig. 5c). The SNP co-segregates with the presence of saponins in our Kurmi × 0654 F3 population, and is likely the causative mechanism for the alternative splicing at an upstream cryptic splice site in the sweet lines (Fig. 5c). This alternative splicing of *CqTSARL1* 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-9d, Supplementary Information 8.8.), which are both necessary for regulation of transcription. All bitter accessions in our re-sequencing pool share the same allele (G) found in the bitter progeny of Kurmi and 0654, whereas all sweet accessions (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. Interestingly, however, although the G2078C allele is present in the sequenced Atlas line, none of the sweet progeny in the Atlas x 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 *CqTSARL1* gene, with some plants containing the G20178C 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 different mutation other than the G2078C allele. Indeed, we found strong evidence of insertions in and around the *CqTSARL1* gene in all the sweet progeny of the Atlas × Carina Red population (Fig. 5c, Extended Data Fig. 10). In particular, two exonic insertions in *CqTSARL1* in the sweet progeny are likely to inactivate the gene and result in a sweet phenotype. The identification of multiple, independent mutations in *CqTSARL1* that co-segregate with the sweet phenotype strongly suggests that this gene regulates the presence and absence of saponins in quinoa seeds. PCR markers designed from sequences surrounding *CqTSARL1* on CqB16 are perfectly linked to the presence and absence of saponins, and can now be used in marker-assisted selection (MAS) to accelerate the development of sweet commercial quinoa varieties.

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 characterise, 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.

Methods

Quinoa sequencing and assembly

We sequenced *Chenopodium quinoa* Willd. (quinoa) accession PI 614886 (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, Republic of Korea). *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)[50], 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)[51] 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[12]. 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[12]. 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.

Linkage mapping and genetic marker analyses

Kurmi × 0654 population. A population was developed by crossing Kurmi (green, sweet) and 0654 (red, bitter). 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). RNA was extracted from inflorescences containing a mixture of flowers and seeds at various stages of development from the parents and 45 individual F3 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[52], and SNPs were called using SAMtools mpileup (v1.1)[53].

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[54].

Atlas x Carina Red population. A second mapping population was developed by crossing Atlas (sweet) and Carina Red (bitter). Bitter and sweet F2 lines were identified by performing foam and taste tests on the F3 seed. DNA sequencing was performed with DNA from the parents and 94 sweet F2 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 F2 lines

Genotype calls were generated for the 94 F2 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[15], 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.

Chromosome pseudomolecules. Pseudomolecules were assembled by concatenating scaffolds based on their order and orientation as determined from the integrated linkage map, with 2,000 N's inserted

between each scaffold. Scaffolds that mapped to multiple linkage groups were assigned to the linkage group to which the greatest number of markers from that scaffold were mapped. A custom Perl script was used to convert the coordinates from the scaffold-level annotation.

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, 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). The trimmed reads were then assembled using the ALLPATHS-LG assembler[55], and GapCloser v1.12[56] 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[57]. RepeatMasker[58] was used to discover and identify repeats within the respective genomes.

AUGUSTUS[59] 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 Iso-Seq reads generated from RNA of different tissues were mapped onto the reference genome using Bowtie 2[60] and GMAP[61], respectively. Hints with locations of potential intron-exon boundaries were generated from the alignment files with the software package BAM2hints in the MAKER package[62]. 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. Genes were 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.

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)[63]. 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[36]. Recommended settings were used for all-against-all BLASTP comparisons (Blast+ v2.3.0[64]) and OrthoMCL analyses. Custom Perl scripts were utilised to process OrthoMCL outputs for visualisation with InteractiVenn[65].

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[66]. The 7,433 alignments were converted into PHYLIP format files by the seqret command in the EMBOSS package[67]. Individual gene trees were then constructed using the maximum likelihood method using proml in PHYLIP[68].

In addition, 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[69] and SAMtools, and the qualified SNPs were combined into a single VCF file which was used as an input into SNPhylo[70] to construct the phylogenetic relationship using maximum likelihood and 1,000 bootstrap iterations.

To identify *FT* homologs, the protein sequence from *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. For the construction of the phylogenetic tree, protein sequences from these four quinoa FT homologs were aligned using MAFFT[66] (progressive method G-INS-1) along with two *B. vulgaris* (gene models: BvFT1-miuf.t1, BvFT2-eewx.t1), one *A. thaliana* (AT1G65480.1), and one *O. sativa* (AFK31087.1) homolog. Phylogenetic analysis was performed with MEGA[71] (v6.06). The JTT+G model was selected as the best fitting model. The phylogenetic tree was estimated using the maximum likelihood method with 1,000 bootstrap replicates. The syntenic relationships between the coding sequences of the chromosomal regions surrounding these *FT* genes were visualised using the GEvo tool of CoGE[72].

The alignment of bHLH domains was performed with Clustal Omega[73], using sequences from Mertens *et al*[49]. The phylogeny was inferred using the maximum likelihood method based on the JTT matrix-based model[74]. 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, 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[75]. A custom Perl script was used to calculate the percentage of each scaffold with more than 5X 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[76]. 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).

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 contcatenating *s*caffolds 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[77] and visualized using MCScanX[78] and VGSC[79]. For the purposes of visualization, quinoa chromosomes CqB05, CqA08, CqB11, CqA15, and CqB16 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[80]. 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, Carrboro, NC, USA) with attached LC20-AD HPLC pump (Shimadzu Scientific Instruments, Ermington, NSW, Australia). 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. 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[81] (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 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. 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. 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.

Specific, individual saponins were identified in quinoa using a preparation of 20 mg of seeds performed according a modified protocol from Giavalisco *et al*.[82]. 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[83]was used to produce homology models for the bHLH region of AUR62017204, AUR62017206 and AUR62010677. RaptorX[84] was used for prediction of secondary structure and disorder. QUARK[85] was used for *ab initio* modelling of the C-terminal domain, and the DALI server[86] was used for 3D homology searches of this region. Models were manually inspected and evaluated using the Pymol program (pymol.org).

References

1. FAO, IFAD & WFP. *The State of Food Insecurity in the World 2015. Meeting the 2015 international hunger targets: taking stock of uneven progress*. Rome, FAO (2015).
2. Pretty, J. Agricultural sustainability: concepts, principles and evidence. *Phis. Trans. R. Soc. B.* **363,** 447–465 (2008).
3. Risi, C. & Galwey, N. W. The Chenopodium grains of the Andes: Inca crops for modern agriculture. *Adv. Appl. Biol.* **10,** 145–216 (1984).
4. 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).
5. 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).
6. 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).
7. 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).
8. Vega-Gálvez, A. *et al.* Nutrition facts and functional potential of quinoa (*Chenopodium quinoa* Willd.), an ancient Andean grain: a review. *J. Sci. Food Agr.* **90,** 2541–2547 (2010).
9. Massawe, F., Mayes, S. & Cheng, A. Crop diversity: an unexploited treasure trove for food security. *Trends Plant Sci.* **21,** 365–368 (2016).
10. Zurita-Silva, A., Fuentes, F., Zamora, P., Jacobsen, S.-E. & Schwember, A. R. Breeding quinoa (*Chenopodium quinoa* Willd.): potential and perspectives. *Mol. Breeding* **34,** 13–30 (2014).
11. Lam, E. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30,** 771–776 (2012).
12. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26,** 342–350 (2016).
13. Palomino, G., Hernández, L. T. & Torres, E. d. l. C. Nuclear genome size and chromosome analysis in *Chenopodium quinoa* and *C. berlandieri* subsp. *nuttalliae*. *Euphytica* **164,** 221–230 (2008).
14. 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).
15. Maughan, P. J. *et al.* Single nucleotide polymorphism identification, characterization, and linkage mapping in quinoa. *Plant Genome* **5,** 114–125 (2012).
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. 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).
19. 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).
20. Gandía-Herrero, F & García-Carmona, F. Biosynthesis of betalains: yellow and violet plant pigments. *Trends Plant Sci.* **18,** 334–343 (2013).
21. 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).

22. 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).
23. 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).
24. 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).
25. 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).
26. Wendel, J. F., Brubaker, C., Alvarez, I., Cronn, R. & Stewart, J. M. Evolution and natural history of the cotton genus. In *Genetics and Genomics of Cotton* (ed. Paterson, A. H.) 3–22 (Springer, New York, 2009).
27. 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).
28. Lynch, M. & Conery, J.S. The evolutionary fate and consequences of duplicate genes. *Science* **10,** 1151–1155 (2000).
29. Wilson, H. D. Quinua biosystematics II: free-living populations. *Econ. Bot.* **42,** 478–494 (1988).
30. Tapia, M.E., Mujica, A. & Canahua, A. Origen y distribución geográfica y sistemas de producción de la quinua (*Chenopodium quinoa* Wild). *Publicación Universidad Nacional Técnica del Altiplano (Perú)* **1,** 1–5 (1980).
31. Wilson, H. D. & Heiser Jr., C. B. The origin and evolutionary relationships of 'Huauzontle' (*Chenopodium nuttalliae* Safford), domesticated chenopod of Mexico. *Amer. J. Bot.* **66,** 198–206 (1979).
32. Smith, B.D. *Chenopodium berlandieri* ssp. *jonesianum*: evidence for a Hopewellian domesticate from Ash Cave, Ohio. *Southeastern Archaeology* **4,** 107–133 (1985).
33. Kolano, B *et al.* Chromosomal localization of two novel repetitive sequences isolated from the *Chenopodium quinoa* Willd. genome. *Genome* **54,** 710–717 (2011).
34. Dohm, J. C. *et al.* The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature* **505,** 546–549 (2014).
35. Sankoff, D., Zheng, C. & Zhu, Q. The collapse of gene complement following whole genome duplication. *BMC Genomics* **11,** 313 (2010).
36. Li, L., Stoeckert Jr., C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).
37. 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).
38. Pin, P.A. & Nilsson, O. The multifaceted roles of FLOWERING LOCUS T in plant development. *Plant Cell Environ.* **35,** 1742–1755 (2012).
39. Cháb, D., Kolář, J., Olson, M. S. & Štorchová, H. Two *FLOWERING LOCUS T* (*FT*) homologs in *Chenopodium rubrum* differ in expression patterns. *Planta* **228**, 929–940 (2008).
40. Kuljanabhagavad, T., Thongphasuk, P., Chamulitrat, W. & Wink, M. Triterpene saponins from *Chenopodium quinoa* Willd. *Phytochemistry* **69,** 1919–1926 (2008).
41. de Geyter, E., Lambert, E., Geelen, D. & Smagghe, G. Novel advances with plant saponins as natural insecticides to control pest insects. *Pest Technol.* **1,** 96–105 (2007).
42. Kuljanabhagavad, T. & Wink, M. Biological activities and chemistry of saponins from *Chenopodium quinoa* Willd. *Phytochem. Rev.* **8,** 473–490 (2009).
43. Konishi, Y., Hirano, S., Tsuboi, H. & Wada, M. Distribution of minerals in quinoa (*Chenopodium quinoa* Willd.) seeds. *Biosci. Biotechnol. Biochem.* **68,** 231–234 (2004).

44. Boughton, B. A., Thinagaran, D., Sarabia, D., Bacic, A. & Roessner, U. Mass spectrometry imaging for plant biology: a review. *Phytochem. Rev.* **15,** 445–488 (2015).
45. Augustin, J. M., Kuzina, V., Andersen, S. B. & Bak, S. Molecular activities, biosynthesis and evolution of triterpenoid saponins. *Phytochemistry* **72,** 435–457 (2011).
46. 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,** 765–806 (2006).
47. 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 Agr.* **80,** 152–156 (2000).
48. Ward, S. M. A recessive allele inhibiting saponin synthesis in two lines of Bolivian quinoa (*Chenopodium quinoa* Willd.). *J. Hered.* **92,** 83–86 (2001).
49. Mertens, J. *et al.* The bHLH transcription factors TSAR1 and TSAR2 regulate triterpene saponin biosynthesis in *Medicago truncatula*. *Plant Physiol.* **170,** 194–210 (2016).
50. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
51. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
52. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
53. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
54. 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).
55. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 1513–1518 (2011).
56. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1,** 18 (2012).
57. Smit, A. F. A. & Hubley, R. *RepeatModeler Open-1.0*. <http://www.repeatmasker.org> (2008-2015).
58. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0*. <http://www.repeatmasker.org> (2013-2015).
59. 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).
60. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
61. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21,** 1859–1875 (2005).
62. Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18,** 188–196 (2008).
63. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
64. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
65. 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).
66. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).

67. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277 (2000).
68. Felsenstein, J. *PHYLIP (Phylogeny Inference Package) Version 3.6a3*. Distributed by the author: http://evolution.genetics.washington.edu/phylip.html
69. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
70. 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).
71. 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).
72. Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* **53,** 661–673 (2008).
73. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).
74. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Method. Biochem. Anal.* **8,** 275–282 (1992).
75. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
76. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19,** 1639–1645 (2009).
77. 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).
78. Wang, Y. *et al. MCScanX*: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucl. Acids Res.* **40,** e49 (2012).
79. Xu, Y. *et al.* VGSC: a web-based vector graph toolkit of genome synteny and collinearity. *Biomed. Res. Int.* **2016,** 7823429 (2016).
80. 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).
81. Sud, M. *et al.* LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **35,** D527–D532 (2007).
82. 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).
83. 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).
84. 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).
85. Xu, D. & Zhang, Y. *Ab initio* protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80,** 1715–1735 (2012).
86. Holm, L & Rosenström, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38,** W545–W549 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

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), and by a grant from the German Research Foundation, DFG (grant No. JU205/24-1). We thank Thiruvarangan Ramaraj from the National Center for Genomic Resources (Santa Fe, NM) for his bioinformatics assistance, Ohoud Mohammed Eid Alharbi at the KAUST Imaging and Characterization Lab for generating SEM images, Khadija Zemmouri at the KAUST Greenhouse for plant care, and Hyun Ho (Heno) Hwang and Ivan D. Gromicho at the KAUST Academic Writing Department for assistance with quinoa illustrations and photographs. We also thank Helena Štorchová, Frank Morton, Daniel Bertero, Francisco Fuentes and David Brenner of USDA-ARS-NPGS for their assistance in providing seed.

Author contributions

MT and DEJ conceived the project. MT supervised the research. MT, DEJ, YSH, DJL, SMS, PJM, ENJ, ENvL, CGvdL and TG conceived and designed the experiments and managed particular components of the project. YSH led the bioinformatics analyses. YSH, DJL, DEJ and BL did most compilation of the genome scaffolds and the genomic analyses. YSH annotated the genome and undertook the analysis of repetitive elements. ENvL did the final genetic mapping. ENJ and PJM provided germplasm, oversaw sequencing of the diploid genomes, oversaw the BioNano mapping, led the comparative genomics work and generously shared their deep knowledge of quinoa. SMS oversaw all saponin-related analyses. CTM, ARS, TB, ES, KM, ENJ and PJM prepared materials and undertook sequencing activities. KM, HO and SN oversaw all phylogenetic analyses. YHW and GG analysed the microRNAs. NS, NMK, RRR, XG and SA-B did saponin analyses. ND and CJ analysed the genes related to flowering time. STA and MAM did the computational structure-function analysis. BAB and UR did the metabolomics imaging. All authors contributed to the writing of the paper. DEJ, YSH, DJL, SMS, BL and MT organized the manuscript. DEJ and MT coordinated the project. DEJ, YSH, DJL, SMS and BL contributed equally.

Author information

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 File 9. The quinoa genome can also be accessed at http://www.cbrc.kaust.edu.sa/chenopodiumdb/ and on the Phytozome database (www.phytozome.net/). Reprints and permission 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 MT (mark.tester@kaust.edu.sa).

Tables

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	1,385,456,844
Longest scaffold (bp)	2,949,784	1,614,553	11,561,360	11,561,360	23,816,425	133,230,289*
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	64,593,919*
L50 scaffold count	243	1,285	-	177	105	10*
N90 scaffold length (bp)	55,204	27,807	-	157,165	249,904	26,142,578*
L90 scaffold count	1,215	5,075	-	800	439	17*
Missing bases (%)	2.52	7.49	0.00	4.53	4.56	4.56

*Based on scaffolds assigned to LGs

Figure legends

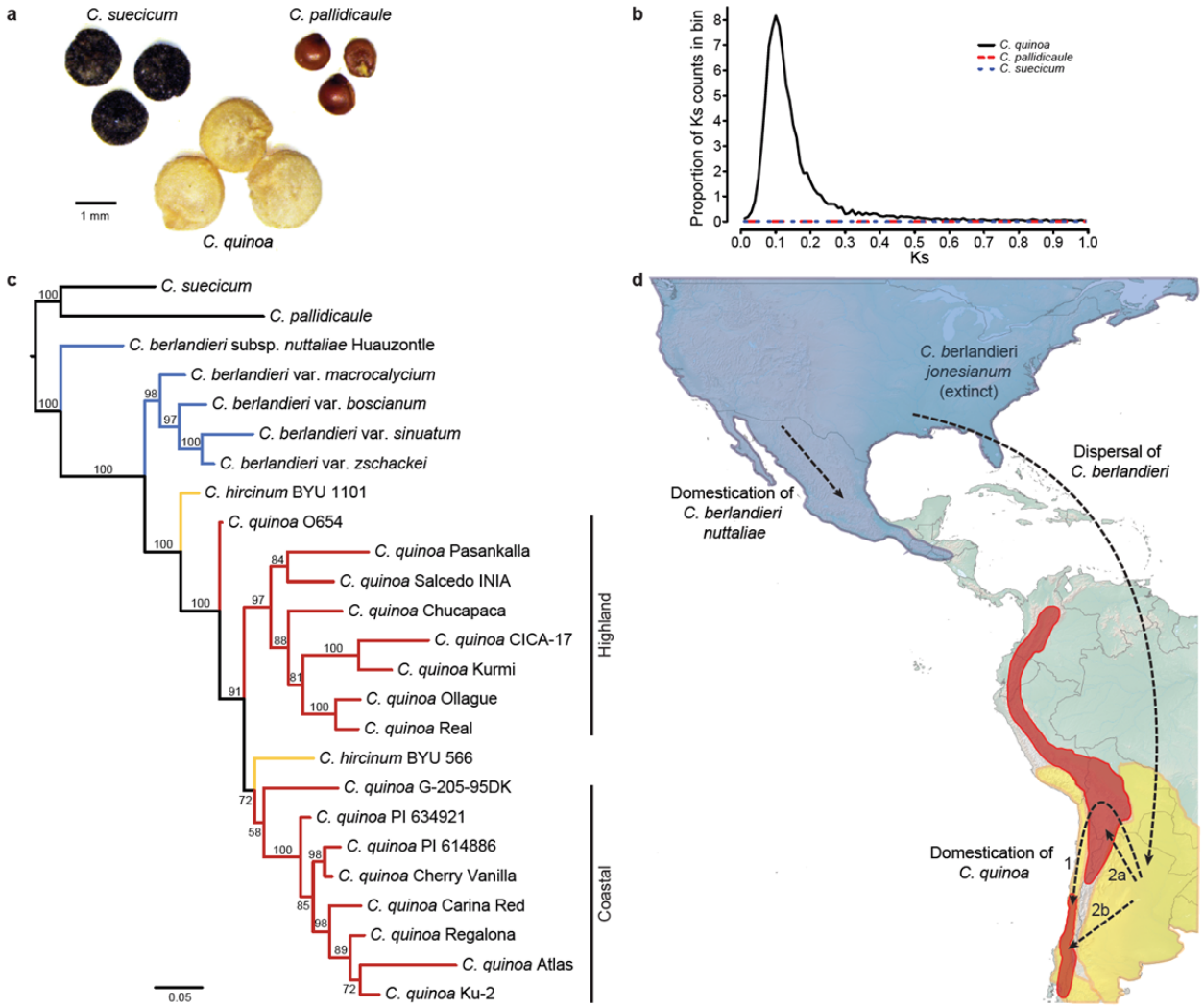

Figure 1 | Evolutionary history of quinoa. a, Seeds of *C. suecicum, C. pallidicaule*, and quinoa. **b**, The proportion of gene pairs in each species binned according to Ks values. **c**, Maximum likelihood tree generated from 3,132 SNPs. Black branches, diploid species. Coloured branches, tetraploid species: red, quinoa accessions; blue, *C. berlandieri* accessions; yellow, *C. hircinum* accessions. 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 hypothesised 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 (1) that gave rise to highland and subsequently coastal quinoa, or in two events that gave rise to highland (2a) and coastal quinoa (2b) independently. Blue, red, and yellow shading represents the geographic distribution of *C. berlandieri*, quinoa, and *C. hircinum*, respectively.

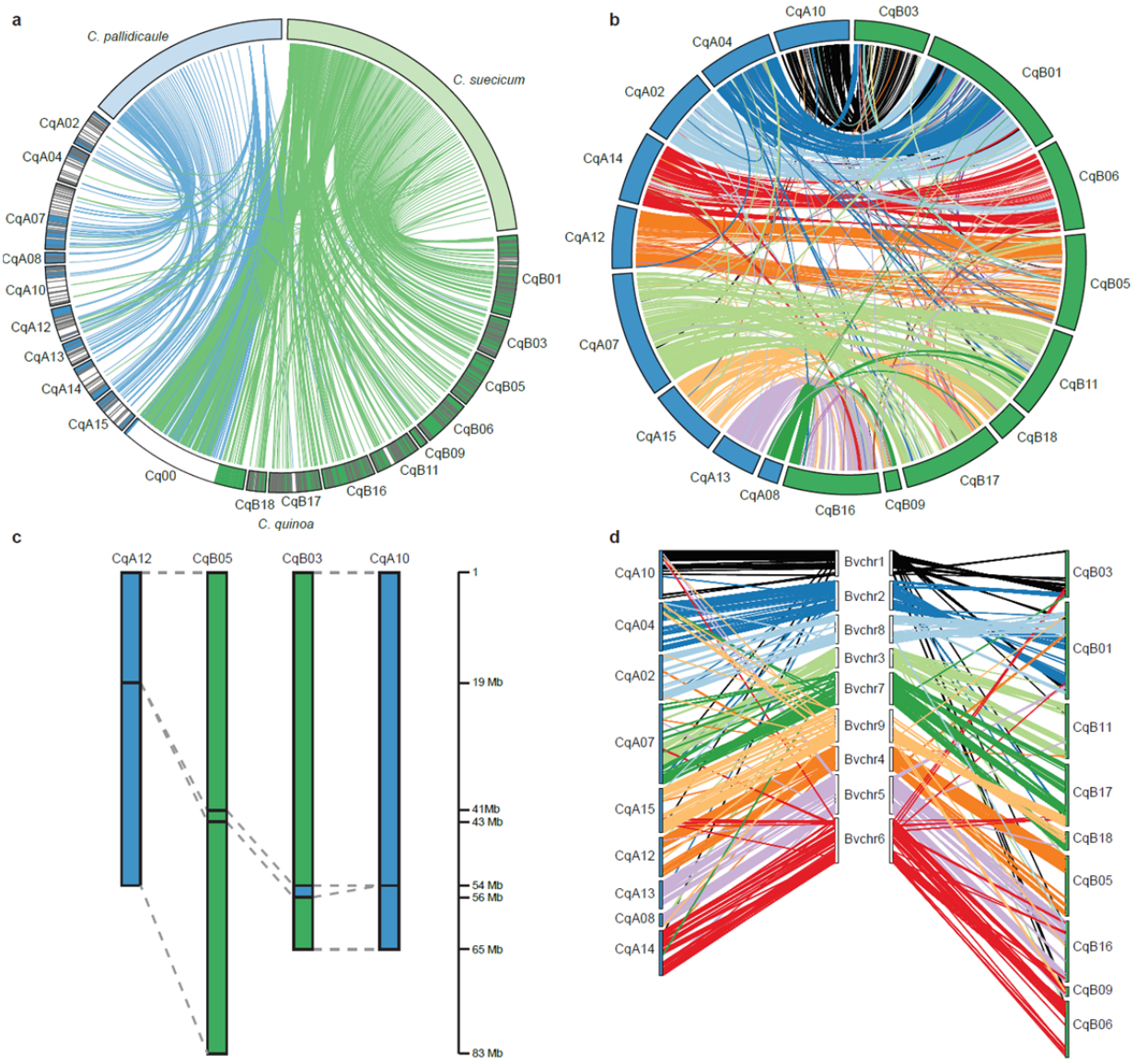

Figure 2 | Identification and characterisation of quinoa sub-genomes. a, Blue lines and green lines connect regions of the *C. pallidicaule* and *C. suecicum* genomes, respectively, with their orthologous regions in the quinoa genome based on BLASTN. Quinoa scaffolds are arranged into chromosomes, with 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 CqA12, CqB05, CqB03, and CqA10, highlighting the homeologous regions between the B sub-genome chromosomes. 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, CqB05 and Cq10A were inverted. **d**, Syntenic relationships between *B. vulgaris* (Bv) 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.

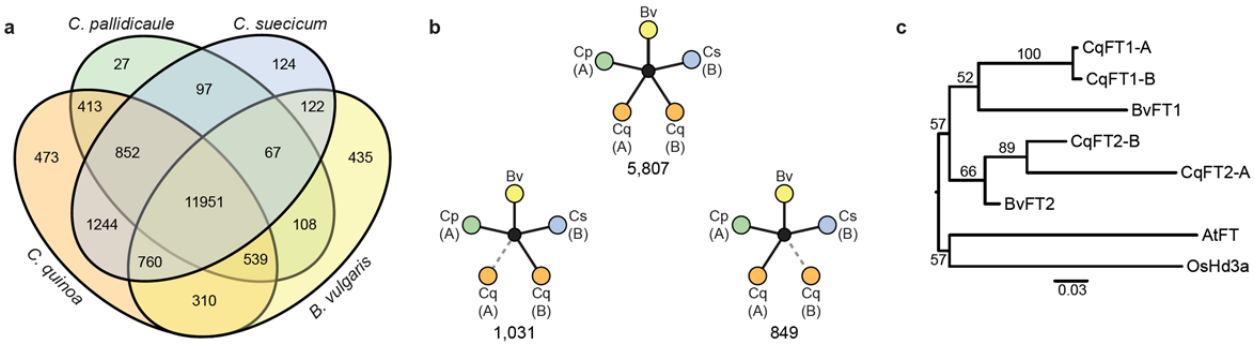

Figure 3 | Sub-genome gene loss and retention. a, The number of orthologous protein-coding gene clusters shared between or unique to quinoa, *C. pallidicaule*, *C. suecicum*, and *B. vulgaris*. **b**, The number of gene sets for which each gene has been retained as a single copy in each genome/sub-genome (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 orthologs in quinoa (Cq) and *B. vulgaris* (Bv). The tree is rooted on the branch containing orthologous FT sequences from *Arabidopsis thaliana* (AtFT) and *Oryza sativa* (OsHd3a). Branch values represent the percentage of 1,000 bootstrap replicates that support the topology. Scale bar represents substitutions per site.

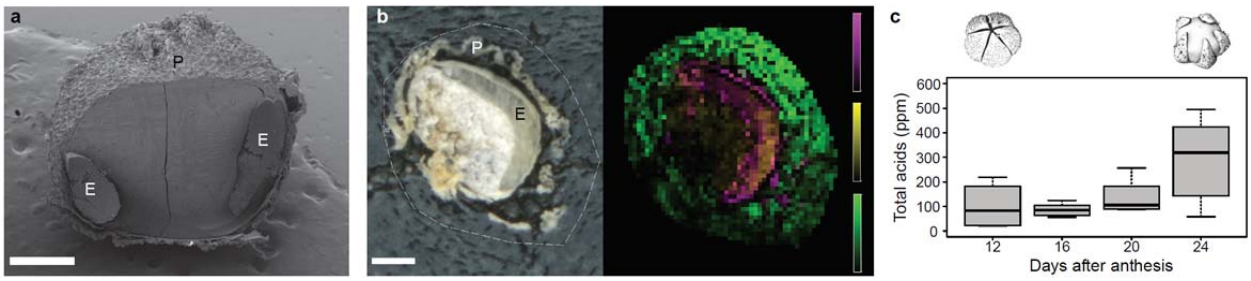

Figure 4 | **Saponins in the seeds of the reference quinoa accession PI 614886. a**, SEM image of a quinoa seed cross-section, showing the pericarp (P) and embryo (E). **b**, Imaging MS visualization of selected masses, including saponins in the pericarp of a quinoa seed. Purple gradient bar, tentative phosphatidylcholine-(34:1), ($[M+Na]^+$ *m/z* = 782.5610, calc. 782.5670, 7.7 ppm error); yellow gradient bar, tentative triacylglycerol-(54:6), ($[M+K]^+$ *m/z* = 917.6971, calc. 917.6995, 2.6 ppm 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 ppm error). Coloured bars represent the ion signal intensity scaled from 0% (bottom) to 50% (top) of maximum signal. Scale bars in **a** and **b**, 500 µm. **c**, Accumulation of saponins as measured by total acids during seed development. Illustrations represent fruit development at 12 and 24 days after anthesis.

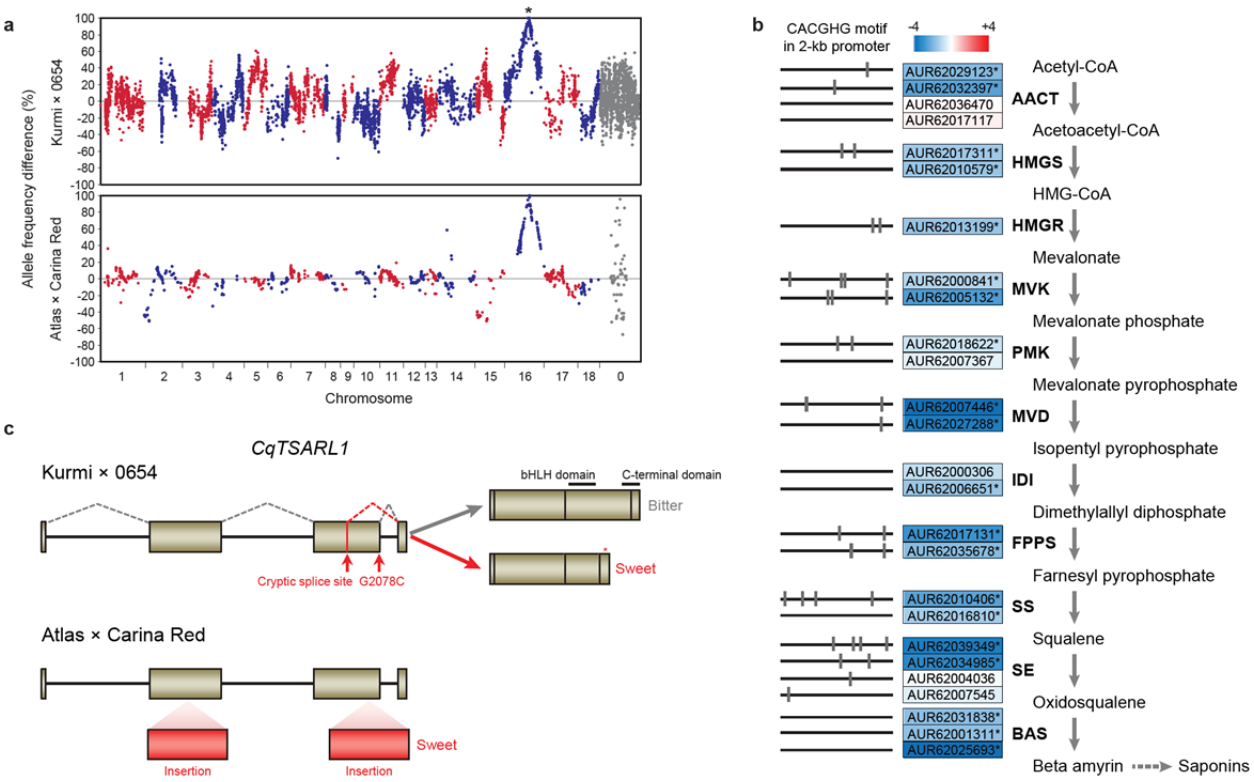

Figure 5 | Candidate gene underlying saponin production. a, Mapping of the saponin production locus. The percentage difference in allele frequency of sweet progeny compared to bitter progeny in the Kurmi × 0654 (top) and Atlas × Carina Red (bottom) populations. Alternating red and blue dots indicate 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 *CqTSARL1*. **b**, 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 CqTSARL1. **c**, Gene models of *CqTSARL1* in bitter and sweet lines. In bitter lines of both populations, normal splicing of the four exons produces a full-length transcript that contains the bHLH and conserved C-terminal domains (indicated). In sweet lines of Kurmi × 0654, alternative splicing occurs at a cryptic splice site in the third exon, likely because of the G20178C mutation at the last position of the exon. This results in an alternative transcript that lacks the C-terminal domain and contains a premature stop codon (red asterisk). In Atlas × Carina Red, sweet lines contain insertions in the second and third exons, which consequently likely disrupt the protein function.