nature communications

Article

https://doi.org/10.1038/s41467-025-67371-7

Population-specific pangenome unveils a third *FAD2* gene and solves the peanut mid-oleic fatty acid mystery

Received: 9 February 2025

Accepted: 28 November 2025

Published online: 23 December 2025

Check for updates

Ethan Thompson[1,2,3,7], Walid Korani[3,7], Dongliang Wu[1,2,7], Vanika Garg[4,7], Brandon Tonnis[5], Mingli Wang[5], C. Corley Holbrook[2], Peggy Ozias-Akins[6], Albert K. Culbreath[1], Rajeev K. Varshney[4] ✉, Baozhu Guo[2] ✉ & Josh P. Clevenger[3] ✉

Accumulating evidences have shown that the mid-oleic fatty acid phenotype in peanuts cannot be explained by the traditional two-gene model involving *AhFAD2A* and *AhFAD2B*, which are genes encoding fatty-acid desaturase 2. But the underlying genetic mechanism remains unclear. Here, we present a population-specific pangenome using the eight founder genomes of the PeanutMAGIC population. This graph-based pangenome serves as a comprehensive reference, capturing all segregating haplotypes within the population. We conduct whole genome sequencing for the MAGIC Core, a subset of 310 RILs, for genotyping. Using pangenome-based genotypes, we trace recombination for detailed genomic analysis and phenotypic association. This investigation identifies a unique third gene, named *AhFAD2C*, near *AhFAD2B*. When recombination occurs, *AhFAD2C* segregates from *AhFAD2B*. We reveal the genotype determining mid-oleic fatty acid phenotype. Our findings underscore the limitations of a single-reference genome, which leads to false association and marker discovery. In contrast, a population-specific pangenome provides a more reliable framework for genomic studies. This study reveals insights into the genetic mechanism of peanut oil quality and demonstrates the advantages of population-specific pangenomes.

Peanut (*Arachis hypogaea* L.) is cultivated globally as a sustainable and affordable oil and protein source, yielding 54 million tons annually (http://www.fao.org/faostat, 2020). Peanut seeds contain considerable amounts of oil (~50%, dry weight) and protein (~25%, dry weight) that are bioavailable and beneficial for human health[1–3]. Oleic and linoleic fatty acids are two heart-healthy, major components of the total oil content in peanut seeds. Therefore, peanut seed oil composition is an important quality trait for confectionery and wholesale markets, leading to a focus on oil composition in breeding programs and functional studies[4,5]. High-oleic (>75%) peanut cultivars are desirable for their increased shelf life and improved palatability in whole or partial peanut products, whereas low (<55%) to mid-oleic (55–75%) cultivars, with high-palmitic acid content, are valued for peanut butter-based products. In cultivated peanut, fatty acid desaturase 2 (AhFAD2) is the enzyme responsible for the conversion of oleic acid into linoleic acid, with the functional genes *AhFAD2A* and *AhFAD2B*[6,7]. Loss-of-

[1]Department of Plant Pathology, University of Georgia, Tifton, GA, USA. [2]USDA-ARS Crop Genetics and Breeding Research Unit, Tifton, GA, USA. [3]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. [4]WA State Agricultural Biotechnology Centre, Centre for Crop and Food Innovation, Food Futures Institute, Murdoch University, Murdoch, WA, Australia. [5]USDA-ARS Plant Genetic Resources Conservation Research Unit, Griffin, GA, USA. [6]Department of Horticulture and Institute of Plant Breeding, Genetics and Genomics, University of Georgia, Tifton, GA, USA. [7]These authors contributed equally: Ethan Thompson, Walid Korani, Dongliang Wu, Vanika Garg. ✉e-mail: rajeev.varshney@murdoch.edu.au; baozhu.guo@usda.gov; jclevenger@hudsonalpha.org

function mutations in *AhFAD2A* and *AhFAD2B* have been identified in breeding populations, with selectable markers available for marker-assisted selection (MAS) of high-oleic phenotypes[5,8,9]. The *AhFAD2A* and *AhFAD2B* markers have been utilized in the development of a high-oleic and root-knot nematode (RKN)-resistant variety, 'Tifguard High O/L', suggesting that information of two genes is sufficient to predict and select for high-oleic varieties with other desired traits[5]. However, mid-oleic peanut lines have been observed, but their inheritance could not be explained solely by the previously established two-gene model[10], leading to the hypothesis that a third recessive gene, along with *AhFAD2A* and *AhFAD2B*, is needed to maintain the high-oleic fatty acid trait in peanut[10]. Interestingly, Pandey et al.[11] reported that *AhFAD2B* had a higher phenotypic effect on oleic acid content than *AhFAD2A* using two biparental populations, and the normal distribution of oleic fatty acid in the mapping populations did not fit the two-gene model[11], which may suggest a neighboring functional gene that has not been appropriately mapped. The mystery of mid-oleic has perplexed the peanut community, particularly concerning the confectionery industry's need for consistent quality in products.

Biparental populations have been the primary focus of plant genetics. However, they are limited by their construction from a single cross involving only two genetic donors[12]. Multiparental populations, such as multiparent advanced generation intercross (MAGIC) populations, address these limitations by incorporating multiple hybrid crosses and genetic donors[13]. This approach facilitates a segregating population with equal genetic contribution from all founders and known pedigrees, thereby enhancing diversity and fine recombination. The increased genetic diversity and recombination in MAGIC populations enable finer genetic mapping of traits such as oleic acid content to resolve the mystery that has eluded the peanut community. The PeanutMAGIC population exemplifies the advantages of a multiparental breeding approach in peanut, providing increased genetic variation and recombination for genomic study[13].

The community reference genome of *A. hypogaea* cv. 'Tifrunner' (TR) was a joint effort of the International Peanut Genome Sequencing Consortium[14]. The publication of a reference genome signified a major milestone in the peanut community, due to the allotetraploid (AABB-type genome; $2n = 4x = 40$ chromosomes) nature of this species and the limited subgenome divergence that stemmed from the recent hybridization event from its related progenitors (*Arachis duranensis*, AA; *Arachis ipaensis*, BB). These factors have made genome sequencing and assembly challenging for tetraploid peanut[14,15]. The reference genome has been pivotal for divulging the intricacies and origin of cultivated peanut for crop improvement; however, it does not reveal all the genetic variance within cultivated peanut or a particular mapping population. Recent advancement in sequencing and bioinformatics has been driving a paradigm shift from a single reference-based genome to graph-based pangenome reference in discovery of structural variants and trait mapping[16,17]. Conversely, a species-level pangenome offers a library of all variations present throughout the species, more variants than within a specific population. The founders of the PeanutMAGIC population do not fully represent the global diversity of peanut. Therefore, the use of a species-level pangenome might include variations that are not possible in the population, potentially introducing error and leading to improper genotyping for PeanutMAGIC recombinant inbred lines (RILs).

Here, we present chromosome-level assemblies for all founders of PeanutMAGIC[13] and construct a population-specific pangenome for the identification of all inheritable haplotypes for genotyping. With considerable detail, we can detect recombination points that accumulates throughout population synthesis using pangenome-based markers to identify the cause of the mid-oleic phenotype that puzzles the peanut research community (Supplementary Fig. 1). Additionally, we show the limitations of single-reference genotyping that leads to false discovery and the improvements of pangenome-based studies. We support these finding with long read low coverage sequencing and full transcript RNA.

Results

Genome assemblies of PeanutMAGIC founders and pangenome construction

The PeanutMAGIC founders, "Georgia-13M" (13M)[18], "SunOleic 97R" (97R)[19], "GT-C20" (C20)[20], "Florida-07" (F07)[21], "NC94022" (NC)[22], and "TifNV-High O/L" (TNV)[23], were sequenced using two cells of PacBio Sequel II, and "GP-NC WS16" (WS16)[24] was sequenced using one cell of PacBio Revio. Four of the founders (13M, 97R, C20, and F07) were scaffolded using high-throughput chromosome conformation capture (Hi-C) data to generate chromosome-length assemblies (Supplementary Data 1). Three founders (NC, TNV, and WS16) were scaffolded using RagTag[25], with the eighth founder (TR) as the reference (Supplementary Data 2). The genome size of the founders ranged from 2490 Mb (F07) to 2629 Mb (TNV) with contig N50 ranging from 21 Mb (F07) to 60 Mb (TNV) (Supplementary Table 1). The contig N50 for TR[14] was 1 Mb, highlighting the improved assembly quality achieved for the founder genomes. The variance of contig N50 for sequenced founders may stem from sequence quality differences and scaffolding efficiency between genotypes. The completeness of the assemblies was estimated through Benchmarking Universal Single-Copy Orthologs (BUSCO)[26], ranging from 95.2 (F07) to 97.6 (TR and WS16) (Supplementary Table 2). The duplicated BUSCO values among the founders ranged from 89.3 (13M and F07) to 93.3 (TR), indicating most of the allopolyploid gene duplications were maintained from the polyploidization event that led to present day peanut (Fig. 1a).

A graph-based pangenome was constructed using the Minigraph-Cactus pipeline with TR as the reference genome. This population-specific pangenome of PeanutMAGIC can be directly utilized with TR-based annotations[27,28]. The resulting acyclic, directed graph enables the resolution of founder haplotypes as traversed paths, with bubbles representing sites of variation, thereby preserving the founder origin of genetic variation. Within the PeanutMAGIC pangenome, 2,762,166 variant sites were identified, consisting of 1,606,159 SNPs and 1,156,007 non-SNP variants, including indels, inversions, and complex variants. Previously, only 138,151 markers were identified for PeanutMAGIC RILs, indicating that there are nearly 20 times more variants present in the population compared to the observed markers used to evaluate the population[13]. To examine similarities within the allotetraploid genome, syntenic blocks were identified and show parallel sequences for homeologous and non-homeologous regions (Fig. 1b). The high collinearity between and within the A and B subgenomes supports previous studies and exemplifies the difficulties of generating markers in the past[29–32]. Furthermore, syntenic block connections among paralogs underscore the repetitiveness of genes within subgenomes, further complicating marker identification and genome assembly. The total number of variants per chromosome ranged from 34,463 (Chr.08) to 669,133 (Chr.09) (Fig. 1c). The increased variation on Chr.09 is due to the introgression from the wild diploid peanut *Arachis cardenasii* by the founder TNV[13]. The difference between the number of variants on homeologous chromosomes ranged from 989 (628 SNPs; 361 non-SNP; Chr.04 and Chr.14) to 153,392 (30,288 SNP; 123,104 non-SNP; Chr.01 and Chr.11), excluding Chr.09 and Chr.19, suggesting different levels of conservation between homeologs (Fig. 1c, and Supplementary Table 3).

In the case of a population-specific pangenome, other published *A. hypogaea* genomes are available that can be leveraged to study genomic variations at a species level, as in other crops like *Vitis vinifera* or *Brassica oleracea*[33,34], or at the genus level, such as *Citrullus* or *Cicer*[35,36]. As a reference library, a population-specific pangenome should detail heritable variance within the population without creating a new source of reference bias from variants that are in other genomes that are not in the pedigree. Furthermore, the risk of overfitting the

https://doi.org/10.1038/s41467-025-67371-7

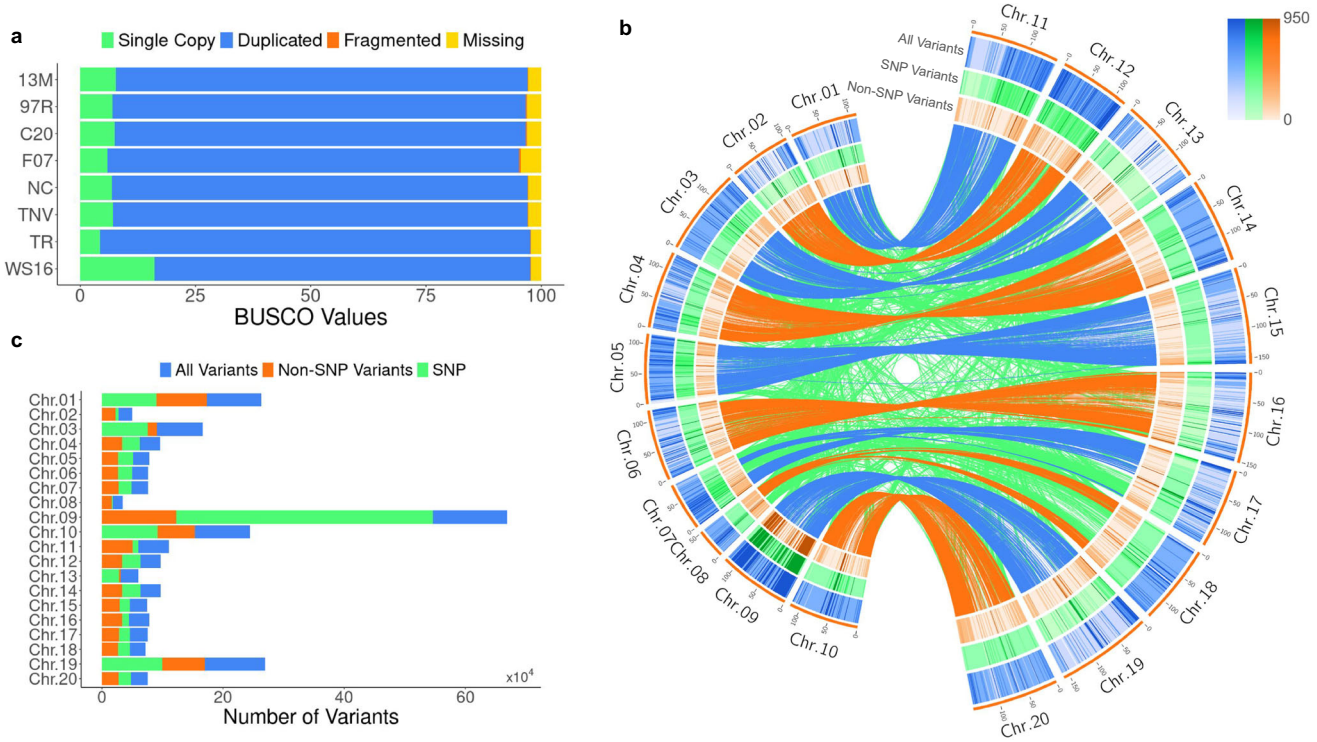


Fig. 1 | PeanutMAGIC pangenomic variation. a BUSCO values of the eight PeanutMAGIC founder assemblies, 'Georgia-13M' (13M), 'SunOleic 97R' (97R), 'GT-C20' (C20), 'Florida-07' (F07), 'NC94022' (NC), 'TifNV-High O/L' (TNV), 'Tifrunner' (TR), and 'GP-NC WS16' (WS16), showing high duplicated BUSCO values from similar subgenomes. **b** Circos plot representing A (Chr.01-Chr.10, left side) and B (Chr.11-Chr.20, right side) subgenomes of cultivated peanut with outer orange bars as scales in Mb. The distribution of all variant sites within the PeanutMAGIC pangenome is represented as a blue heatmap (per 1 Mb; 0-950 sites), single nucleotide polymorphism (SNP) variant locations are represented as a green heatmap (per 1 Mb; 0–950 sites), and all non-SNP variants are represented as an orange heatmap (per 1 Mb; 0–950 sites). Blue and orange lines represent homeologous syntenic blocks between A and B subgenomes and green lines represent non-homeologous syntenic blocks. **c** Number of PeanutMAGIC pangenome variants per chromosome. Blue bars represent all variants, orange bars represent non-SNP variants, and green bars represent SNP variants. Note: parts of the longer bars are overshadowed by smaller bars. Source data are provided as a Source Data file.

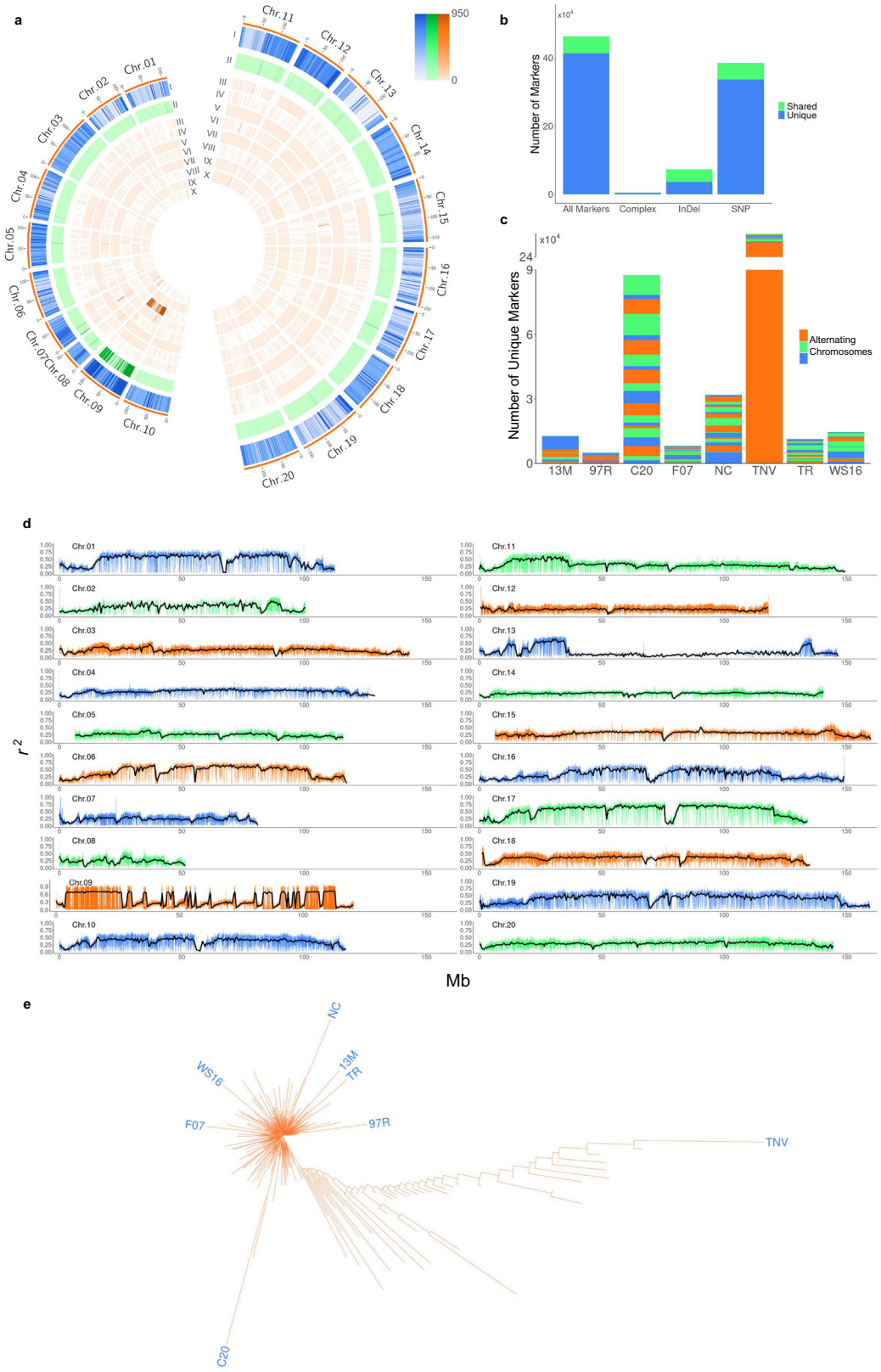
pangenome to the population should be minimal due to all offspring evaluated are from the genetic sources in the PeanutMAGIC pangenome[13]. To understand the differences between a population-specific pangenome and a species-level pangenome for use as a genotyping reference, we constructed a separate extended pangenome that includes genomes of three peanut genotypes, "Shitouqi"[37], "Fuhuasheng"[38] from China, and "Bailey II" from North Carolina[39] along with the PeanutMAGIC founder genomes to serve as a species-level pangenome representative. The extended pangenome has a total of 10,721,659 variants, adding 7,959,493 new variants (2.88 times of PeanutMAGIC pangenome total variants). These variants span throughout the genomes and are not strictly in defined regions that can be simply identified (Supplementary Figs. 2 and 3). The excessive variation in the extended pangenome that is not included in the population could increase genotyping errors if used for genotype calling, making species-level pangenomes improper for mapping population applications. Therefore, the PeanutMAGIC pangenome can provide a comprehensive library of inherited haplotypes that is specific to the PeanutMAGIC population. This resource serves as an unbiased reference for accurate genotyping, trait mapping, and identification of candidate variations associated with phenotypes of interest.

PeanutMAGIC pangenome markers reveal chromosomal recombination patterns

The PeanutMAGIC population has 3187 RILs, and the MAGIC Core is a subset of 310 RILs to represent the entire PeanutMAGIC population. The MAGIC Core subset has been sequenced using Illumina technology with an average coverage of 1×[13]. A total of 463,273 markers, including SNP, InDel, and complex variants (sites with more than two variants including InDels and multi-allelic SNPs) were identified for the MAGIC Core using three criteria: a variant is present in at least 25% of the MAGIC Core, in the PeanutMAGIC pangenome, and in at least one founder linear genome (Supplementary Fig. 4). Of the pangenomic markers, founder-specific markers preserve the origin of a locus and enable accurate tracking of recombination points throughout the breeding scheme. A total of 413,488 unique markers (variant possessed by only one founder) were identified, including 336,941 SNPs, 36,268 InDels, and 4697 complex sites (Fig. 2a). The high number of SNP markers, compared to InDel and complex markers, could be attributed to the limitations of low coverage sequencing for marker selection (Fig. 2b). In the case where structural variants lacked consistent and sufficient coverage, differentiating them across the population becomes challenging. Structural variants often contain repetitive sequences from neighboring regions, leading to ambiguous read mapping to both present and absent variants. This ambiguity renders structural variants less informative for marker selection. The number of founder-specific markers ranged from 5023 in 97R to 241,971 in TNV (Fig. 2c). A majority of TNV specific markers (240,955; 99.58%) are located on Chr.09, which harbors the diploid *A. cardenasii* introgression[23]. The founder C20 has 87,611 markers, highlighting its unique variants derived from a different subspecies and germplasm collection originating in China[20] (Fig. 2c).

The first 6.3 Mb of Chr.05 and Chr.15 lack markers due to the homeologous recombinant nature of these regions[14]. To illustrate this, we aligned the founder TR to the diploid progenitors *A. duranensis* (AA) and *A. ipaensis* (BB) (Supplementary Fig. 5). The alleles within these locations can move, in blocks, between maternal Chr.05 and

https://doi.org/10.1038/s41467-025-67371-7



paternal Chr.05 in addition to maternal or paternal Chr.15. This complexity of tetrasomic inheritance makes it challenging to assign a single location for a marker site or to pass through Mendelian segregation filtering. To accurately determine where the marker resides and track recombination between Chr.05 and Chr.15 was challenging and difficult to describe with current genotyping methods; therefore, these regions were excluded from marker calling to not perpetuate error from genotyping. This issue is consistent between the single-reference-based markers and pangenome-based markers and are not present in either marker set[13].

To visualize recombination patterns within the PeanutMAGIC population, we analyzed r^2 values in a sliding window of 300 markers (Fig. 2d). The level of recombination varied both on the macro and micro scale across the genome. The chromosomes Chr.02, Chr.03,

Fig. 2 | PeanutMAGIC pangenomic markers. a Circos plot representing A (Chr.01-Chr.10, left side) and B (Chr.11-Chr.20, right side) subgenomes of cultivated peanut with outer orange bars representing scales in Mb. I, Heatmap of all variant sites within the PeanutMAGIC pangenome, per 1 Mb. II, Heatmap of all pangenome markers retained for analysis, per 1 Mb. Heat map of founder-specific markers: 'Georgia-13M' (13M, III), 'SunOleic 97R' (97R, IV), 'GT-C20' (C20, V), 'Florida-07' (F07, VI), 'NC94022' (NC, VII), 'TifNV-High O/L' (TNV, VIII), 'Tifrunner' (TR, XI), 'GP-NC WS16' (WS16, X), per 1 Mb. **b** The number of all markers, complex markers, InDel markers, and SNP markers. Blue represents the number of markers unique to a single founder and green represents the number of markers that are not exclusive to one founder as a percentage of the total bar. **c** The number of markers unique to each founder. Alternating color bands represent the number of markers per chromosome from Chr.01 (bottom) to Chr.20 (top). **d** r^2 values across chromosomes of the PeanutMAGIC Core population. Values were calculated through a sliding window of 300 markers. Individual markers are represented by blue, green or orange lines. The black line represents 0.5 Mb averages of r^2 values. **e** Neighbor Joining phylogenetic tree of the PeanutMAGIC Core using PeanutMAGIC pangenomic markers. Founders are highlighted in blue. Source data are provided as a Source Data file.

Chr.04, Chr.05, Chr.07, Chr.08, Chr.12, Chr.14, Chr.15, Chr.18, and Chr.20 exhibited consistent recombination levels across their respective structures ($r^2 < 0.5$), suggesting the detectable alleles on these chromosomes segregated normally. In contrast, Chr.01, Chr.06, Chr.09, Chr.10, Chr.11, Chr.13, Chr.16, Chr.17, and Chr.19 contain regions with elevated LD ($r^2 > 0.5$), indicating that alleles in these areas do not observe independent assortment (Fig. 2d). The difference in recombination patterns may be the product of chromosome sequence divergence and structural variation facilitating regions that are more compatible than others, particularly from founders with increased genetic diversity (NC, C20, TNV).

Upon further examination, Chr.01, Chr.10, Chr.17, and Chr.19 exhibit large, distinct pericentromeric blocks with limited recombination, compared to the telomeric regions of the respective chromosomes. Chr.01 has a unique pericentromeric region derived from the founder NC (Supplementary Fig. 6). C20 is the donor for the unique pericentromeric regions of Chr.10 and Chr.17 (Supplementary Figs. 7 and 8). Interestingly, a different number of RILs possess the C20-based pericentromere on Chr.10 (27 RILs, 8.71%) compared to Chr.17 (54 RILs, 17.42%). This difference allows two factors to influence the r^2 values for the respective chromosome: the number of lines with the distinct pericentromere and the level of recombination within the pericentromeric region. The unique pericentromere of Chr.19 originates from the founder 13M, a cultivar that possesses considerably less founder-specific alleles than NC and C20 but maintains a unique pericentromere (Supplementary Fig. 9). These findings provide insights into recombination levels and inheritance patterns of pericentromeric regions of cultivated peanut chromosomes.

The chromosomes Chr.09, Chr.11, and Chr.13 displayed specific regions resistant to recombination (Supplementary Figs. 10–12). On Chr.09, the regions originated from the well-known *A. cardenasii* introgression, which spans the majority of Chr.09 on TNV. These regions recombine within the population as micro-fragments rather than large fragments, contrary from previous characterization in biparental populations[40,41] and in PeanutMAGIC[13], suggesting pangenome-based markers improve the ability to track such variants that have been previously undetectable (Supplementary Fig. 10). The elevated LD region on Chr.11 was derived from a block of alleles from the founder NC, suggesting reduced compatibility for recombination in that region compared to other NC alleles toward the middle and end of the chromosome, likely due to increased divergence in this block (Supplementary Fig. 11). Chr.13 contained three regions with elevated LD: one at the top, one at the bottom, and a central block. The top and bottom regions stemmed from C20 alleles that tended to segregate as a block. The central block harbored little marker diversity, suggesting that the elevated LD in this region is due to an inability to detect recombination rather than a true reduction in recombination (Supplementary Fig. 12).

The hindered ability to track recombination was also observed in homeologous chromosomes Chr.06 and Chr.16, where the pericentromeric regions showed limited marker variation. The lack of marker variation artificially inflated LD values due to the scarcity of markers able to trace recombination within these regions (Supplementary Fig. 13). The regions with low marker variance likely reflect existing variation that cannot be accurately tracked, possibly due to increased repetitiveness or structural variants. In low coverage sequencing, such variants can result in the common allele being predominantly called in a block, creating areas with limited trackable variation, artificially inflating LD estimates (Supplementary Fig. 13). In contrast, regions of low genomic complexity such as those on Chr.13, provide limited information to accurately estimate recombination. This is due to an expanded physical window size between markers, which reduces the linkage signal and results in lower r^2 values (Supplementary Fig. 12). Such regions hinder genomic analyzes and association studies due to their limited capacity to provide reliable recombination estimates. For regions with sufficient marker density and marker variation, recombination can be accurately tracked to identify areas of high LD or regions that segregate normally (e.g., Chr.11, Supplementary Fig. 11). Segregant founder-specific markers enable the ability to identify the origin of a particular genomic region, provided the region is distinct. Thus, a population-specific pangenome provides valuable insights to distinguish between regions of genuinely low complexity, inhibited recombination, and those merely lacking data.

A phylogenetic tree of the MAGIC Core shows an even distribution of founders and lines, except for the TNV clade (Fig. 2e). This divergence stems from the inheritance and recombination of sizable *A. cardenasii* introgression blocks on Chr.09 from the donor TNV. Recombination of this introgression has been a point of interest to identify causal resistance genes for peanut RKN[5,40,41]. Genomic regions on Chr.09 have been previously identified with consistent high LD using single-reference-based markers[13]. However, with pangenome-based markers, variable LD patterns are observed in the introgressed region throughout the PeanutMAGIC population (Supplementary Fig. 14). This variability highlights the capability of pangenome-based markers to capture recombination events with greater precision. By detecting micro-fragments of the introgression throughout the population, these markers provide insights into introgression inheritance and recombination processes that single-reference-based markers cannot resolve (Supplementary Figs. 10 and 14). Such enhanced resolution could be pivotal in the ongoing efforts to identify the elusive RKN resistance genes[42].

PeanutMAGIC pangenomic markers facilitate the identification of a third *AhFAD2* gene associated with high-oleic fatty acid content

An unbiased reference that integrates a comprehensive set of segregating variants improves marker retention and information within a population, thereby empowering association studies. To assess this hypothesis, peanut oleic acid phenotypes were selected due to their minimal environmentally influenced nature[43,44], their status as the most studied yet unresolved peanut trait[10], and their importance for crop quality[4].

To understand the phenotypic distributions of oleic acid content in the MAGIC Core, seeds from the population were subjected to gas chromatography (GC) for oleic acid quantification. Of the eight founders, four (13M, 97R, F07, TNV) exhibited high-oleic fatty acid phenotypes (>75%), one founder (WS16) displayed mid-oleic (55–75%), and three founders (C20, NC, TR) expressed low oleic content phenotypes (<55%). Within the MAGIC Core, a normal distribution

https://doi.org/10.1038/s41467-025-67371-7

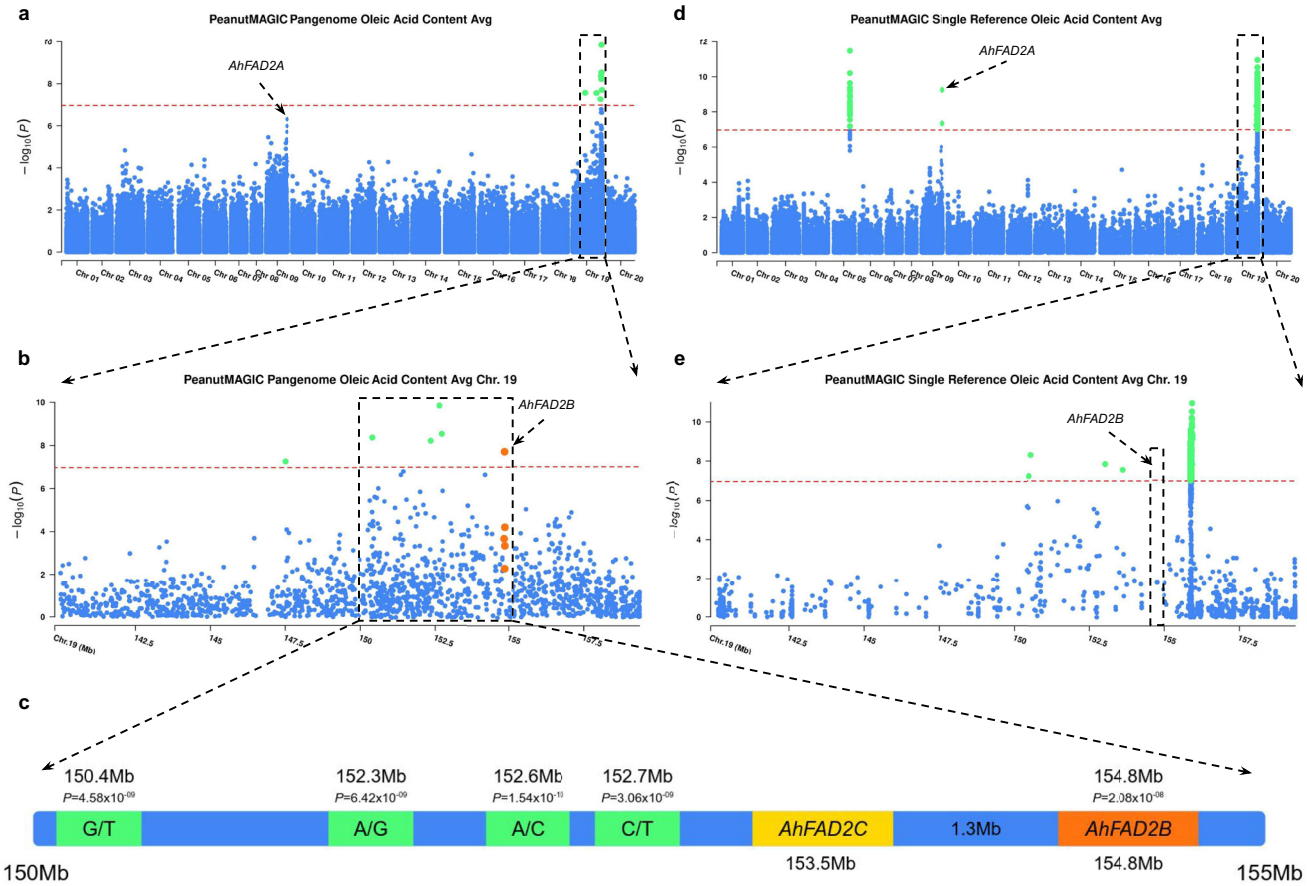


Fig. 3 | Oleic acid content association with PeanutMAGIC pangenomic and single reference markers. a Manhattan plot showing the association signals of pangenome-based markers with oleic acid content from general linear model genome wide association study. A Bonferroni-corrected *P* value of 0.05 was used as a significant threshold ($P = 1.1 \times 10^{-7}$), represented by a horizontal dashed red line. **b** Local association plot of Chr.19 from (**a**) showing multiple significant locations. The orange highlight represents the location of the functional gene *AhFAD2B*. Regions upstream of a known functional gene have higher association values. **c** Genetic view of the associated region showing significant markers and genes of interest. Markers are highlighted in green and denoted above the genome bar with *P* values and physical locations. The known functional gene *AhFAD2B* is highlighted orange. Between the associated markers and a known functional gene, a third *FAD2* gene was identified and termed *AhFAD2C*, highlighted in yellow. **d** Manhattan plot showing PeanutMAGIC Core single-reference marker signals associated with oleic acid content from general linear model genome wide association study, with an additional signal identified on Chr.05. **e** Local association plot on Chr.19 from (**d**) showing a strong association downstream of the known *AhFAD2B* location, with no signal observed at the physical position of *AhFAD2B*. Source data are provided as a Source Data file.

was observed for low and mid-oleic content, with a distinct peak for high-oleic phenotypes (Supplementary Fig. 15). The normal distributions of low and mid-oleic content phenotypes suggest the involvement of more than two genes influencing oleic acid content, as reported by Branch et al.[10].

A genome wide association study (GWAS) using PeanutMAGIC pangenome-based markers identified two chromosomes, Chr.09 and Chr.19, associated with oleic acid content (Fig. 3a). A local association plot of Chr.19 pinpointed the known functional gene *AhFAD2B* at the correct location of 154.8 Mb ($P = 2.08 \times 10^{-08}$) (Fig. 3b). Interestingly, upstream of *AhFAD2B*, four markers with higher significance values ($P = 6.42 \times 10^{-09}$, 4.58×10^{-09}, 3.06×10^{-09}, 1.54×10^{-10}) suggest the presence of another functional gene upstream of *AhFAD2B*. Using the TR protein annotations[14,45], we identified a third *AhFAD2* gene, 1.3 Mb upstream of *AhFAD2B*, neighboring the associated markers (Fig. 3c). This gene, named as *AhFAD2C*, lies near *AhFAD2B*, making segregation challenging in small mapping populations, especially considering the presence of a third gene, *AhFAD2A*, on Chr.09. Additionally, the proximity of the two genes explained the increased association signal in this study and previous findings on Chr.19 compared to Chr.09 using biparental population[11].

To compare with the pangenome-based study, we conducted GWAS using single-reference-based markers for oleic acid content. This analysis included a total of 138,151 markers, 3.35 times less than pangenome-based markers[13]. This study identified three chromosomes associated with oleic acid content: Chr.05, Chr.09, and Chr.19 (Fig. 3d). Chr.05 produced a likely false-positive signal, which was not detected with the PeanutMAGIC pangenome markers and had higher association values than known functional genes. The region identified on Chr.05 spans 3 Kb (90,216,715–90,219,762) with no annotated genes in the region and no genes related to fatty acid biosynthesis were found near this region, suggesting that the signal is a false-positive (Supplementary Data 3). Furthermore, visualization of the PeanutMAGIC pangenome shows no variants in the region, although 30 single-reference markers were called (Supplementary Fig. 16). The identification of these markers in the region may stem from how single-reference markers are called, where the reads are mapped to the reference and sequences that are similar can be interpreted as a polymorphism instead of mapping to a different location that is present in a different founder. These locations can be identified in the pangenome and are mapped directly. Additionally, the use of other association models with single-reference-based markers also fail to

identify both Chr.09 and Chr.19, suggesting a genotyping error instead of a model issue (Supplementary Fig. 17). Upon examination of the local association plot of Chr.19 for single-reference-based markers, the identified signal was 1.1 Mb downstream (155.9 Mb) of *AhFAD2B* (154.8 Mb), demonstrating potential inaccuracies with single-reference-markers (Fig. 3e). The inaccurate signals on Chr.05 and Chr.19 undermine the legitimacy of the significant markers in the region containing *AhFAD2C*, thus could be overlooked in a single-reference study. These challenges exemplify some of the difficulties in understanding the genetic underpinnings of oleic acid content utilizing a single-reference framework and highlight the improved accuracy and precision of genomic associations when using a population-specific pangenome over a single-reference genome.

Comparison of biparental and multiparental oleic acid mapping using population-specific pangenome-based markers

To validate these findings, we analyzed a biparental population derived from SunOleic 97R (97R) and NC94022 (NC), referred to as the 'S' population, which has been sequenced at 5× coverage[46,47]. We constructed an 'S' population-specific pangenome and generated markers using the same approach to ensure that the differences between the populations were not due to methodological discrepancies. For the 'S' population, 113,056 population-specific pangenomic marker sites were identified for association with oleic acid content. Similar to the MAGIC Core study, two regions, Chr.09 and Chr.19, were associated with oleic acid content (Fig. 4a). A closer examination of Chr.19 revealed a genomic block associated with *AhFAD2B* and *AhFAD2C* (Fig. 4b). This region exhibited elevated LD ($r^2 > 0.3$) in the 'S' population, suggesting that it generally segregates as a unit (Fig. 4c). In contrast, the MAGIC Core displayed considerably lower LD ($r^2 < 0.2$) in this region to facilitate independent observation of *AhFAD2B* and *AhFAD2C* (Fig. 4d). The findings of the biparental population suggest that only *AhFAD2A* and *AhFAD2B* may be required for the oleic acid phenotype, as minimal recombination occurred between *AhFAD2B* and *AhFAD2C* in the lines examined in this study. This finding highlights the difficulty of previous biparental population studies to resolve the genetic underpinnings of oleic acid and exemplifies the utility of multiparental populations for genomic studies.

The utility of *AhFAD2C* in phenotypic prediction and candidate functionality

The limited findings from biparental mapping populations can be exacerbated when applied as molecular markers in predictive breeding programs, with a greater number of recombinants. This issue was particularly evident in mid-oleic lines derived from a high-oleic population, which puzzled the research community[10]. To illustrate this issue, we performed PCR-based genotyping on the PeanutMAGIC population to evaluate if tracking only *AhFAD2A* and *AhFAD2B* is sufficient for predicting oleic acid content, a common practice in MAS for peanut breeding[5]. DNA was extracted from young leaf tissue of the MAGIC Core and genotyped using *AhFAD2A* and *AhFAD2B* PCR-based selectable markers previously employed in peanut breeding[5]. Interestingly, two founders, TR and WS16, shared the same genotypes for both *AhFAD2A* and *AhFAD2B*, yet TR had low-oleic acid content while WS16 had mid-oleic acid content (Fig. 5a). The founder WS16 was found to possess a mutant genotype of *AhFAD2C*, whereas TR carried a wildtype genotype of *AhFAD2C*. Additionally, six MAGIC Core lines were classified as mutants for both *AhFAD2A* and *AhFAD2B*, thus were identified as high-oleic lines using MAS but exhibited mid-oleic acid content phenotypes. The six lines would be excluded from future use in a breeding program due to the interpretation of contamination without information of *AhFAD2C* showing that recombination occurred between mutant *AhFAD2B* and mutant *AhFAD2C* with a wildtype *AhFAD2C* (Fig. 5a). These findings indicate the necessity of including the third gene, *AhFAD2C*, in MAS application for peanut breeding to predict high-oleic phenotypes and to evaluate line purity.

We further examined *AhFAD2C* to identify potential causal variations. Visualization of the PeanutMAGIC pangenome revealed only two variations near *AhFAD2C*, one SNP located 907 bp upstream of the start codon and one 3-bp InDel situated 182 bp upstream of the SNP (Fig. 5b). Sequence alignment to the founder linear genomes revealed four possible haplotypes for this region: C:A, C:ATTA, T:A, and T:ATTA (Supplementary Table 4). All high-oleic founders (13M, 97R, F07, TNV) and the mid-oleic founder (WS16) possess the C:ATTA haplotype. The low oleic founders exhibited distinct haplotypes: TR had the T:A haplotype, NC had the T:ATTA haplotype, and C20 had the C:A haplotype. To call variants for this specific location throughout the MAGIC Core, we utilized personalized pangenomes[48] to accurately call the SNP and InDel site upstream of *AhFAD2C*. This approach allows for the utility of PeanutMAGIC pangenome subgraphs as a calling reference based on *k*-mer counts instead of the whole pangenome, reducing the potential for false variant calls. This yielded calls for the C/T SNP and the A/ATTA InDel variations upstream of *AhFAD2C* (Supplementary Fig. 18). We found this region to have parallels with other regions on Chr.19, making consistent calling for this site challenging and impractical in a single reference framework without a library of possible haplotypes from the pangenome to reduce ambiguous mapping and interpretation error (Supplementary Fig. 19). We performed an analysis of variance (ANOVA) on the three *AhFAD2* genes to determine their significance for oleic acid content. Results showed the Pr(>F) values of $<2.0 \times 10^{-16}$ for *AhFAD2A* and *AhFAD2B*, and 1.91×10^{-07} for *AhFAD2C*, indicating that all three genes significantly affect oleic acid content in the MAGIC Core population (Supplementary Table 5).

While *AhFAD2A* and *AhFAD2B* each have two alleles, *AhFAD2C* contains four haplotypes that can segregate within the PeanutMAGIC population. To determine which variations had the greatest impact on oleic acid content, a Student's *t*-test was conducted among the different haplotypes (Fig. 5c). The largest single-site difference was observed between C:ATTA and T:ATTA, with a *p*-value of 4.6×10^{-05}. Another notable difference was between C:ATTA and C:A, with a *p*-value of 1.3×10^{-03}. These results suggest that the primary causal variant is the C/T SNP, as no significant change was observed with the addition of the InDel (from T:ATTA to T:A), and the SNP change (C:ATTA to T:ATTA) had a greater impact than the InDel change (C:ATTA to C:A). These analyzes were repeated in the biparental 'S' population, yielding consistent results and confirming the significance of *AhFAD2C* for oleic acid content, with the C/T SNP identified as the primary variation upstream of *AhFAD2C* (Supplementary Table 6, and Supplementary Fig. 20). Additionally, a post-hoc Tukey's honest significant difference (HSD) test was conducted for significant ANOVA factors in both the MAGIC Core and 'S' populations. We found that all significant ANOVA factors had significant differences in the MAGIC Core and 'S' populations, consistent with both the Student's *t*-test and ANOVA analyzes (Supplementary Tables 7 and 8). Of the *AhFAD2C* haplotype comparisons in the PeanutMAGIC, T:ATTA-C:ATTA, C:ATTA-C:A, and T:A-C:ATTA had significant differences (p adj < 0.05), consistent with the Student's *t*-test analysis. The ANOVA, Student's *t*-test, and post-hoc Tukey's HSD analyzes of the MAGIC Core and 'S' populations support that *AhFAD2A*, *AhFAD2B* and *AhFAD2C* significantly contribute to oleic acid content, and the primary causal variation for *AhFAD2C* is the C/T SNP found in both the MAGIC Core and 'S' populations. These findings underscore the value of using a population-specific pangenome to improve the understanding of the genetic basis of important traits.

The significant variations of *AhFAD2C* in the putative promoter region and the lack of variations within the coding sequence, may suggest that the function of this variant is a change in expression levels. To test this hypothesis, we performed PacBio Kinnex long read RNA sequencing on developing seed tissue for three founders with the

https://doi.org/10.1038/s41467-025-67371-7

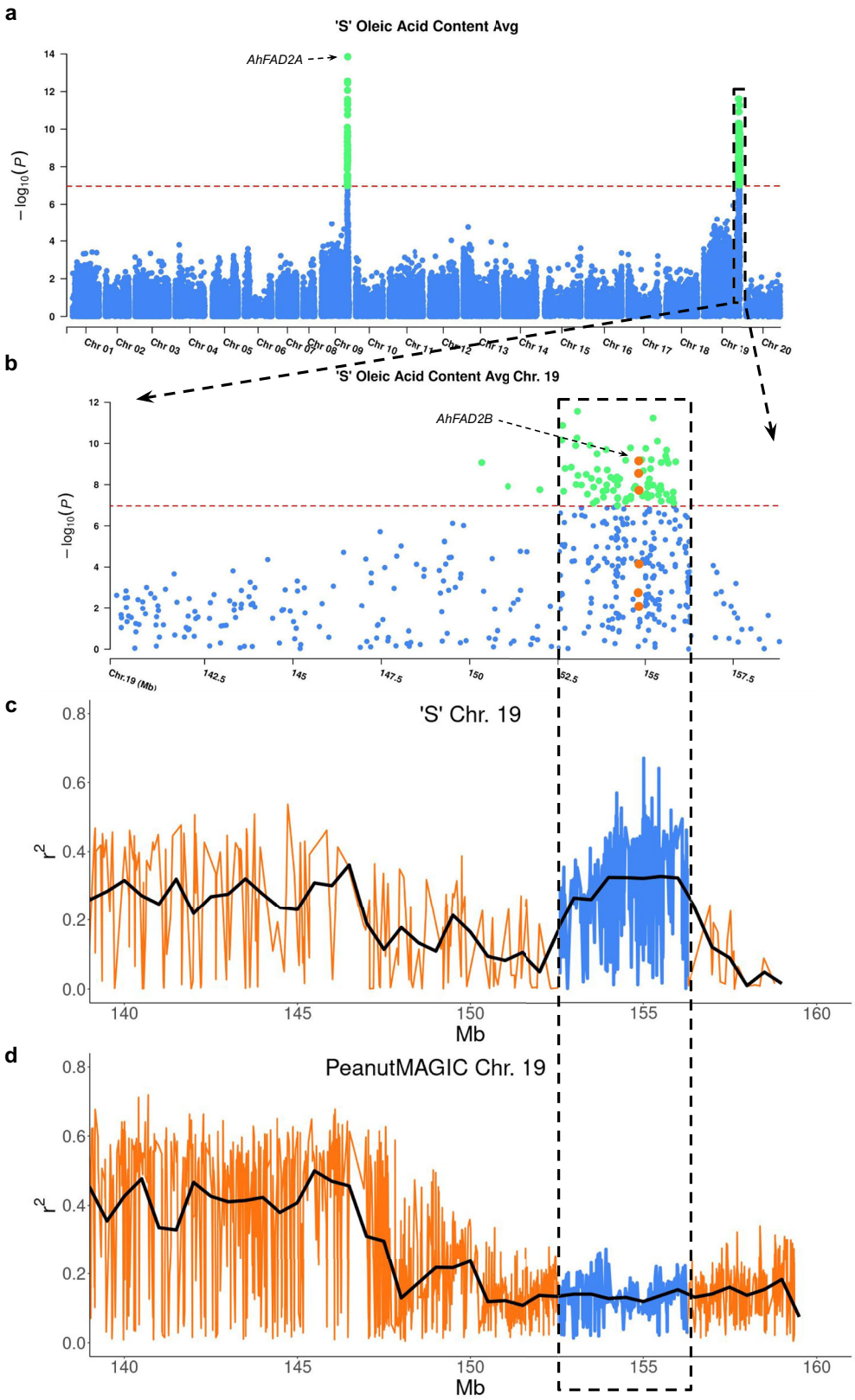


Fig. 4 | The 'S' biparental population pangenome oleic acid content association. a Manhattan plot showing 'S' pangenome marker signals associated with oleic acid content from general linear model genome wide association study. A Bonferroni-corrected *P* value of 0.05 was used as a significant threshold ($P = 1.1 \times 10^{-7}$), represented by a horizontal dashed red line. **b** Local association plot on Chr.19 from (**a**) showing a consistent associated region. The orange highlight represents the location of a functional gene *AhFAD2B*. **c** The 'S' pangenome marker r^2 values of the associated region, indicating increased linkage disequilibrium. **d** PeanutMAGIC pangenome marker r^2 values of the associated region, indicating lower linkage disequilibrium. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-025-67371-7

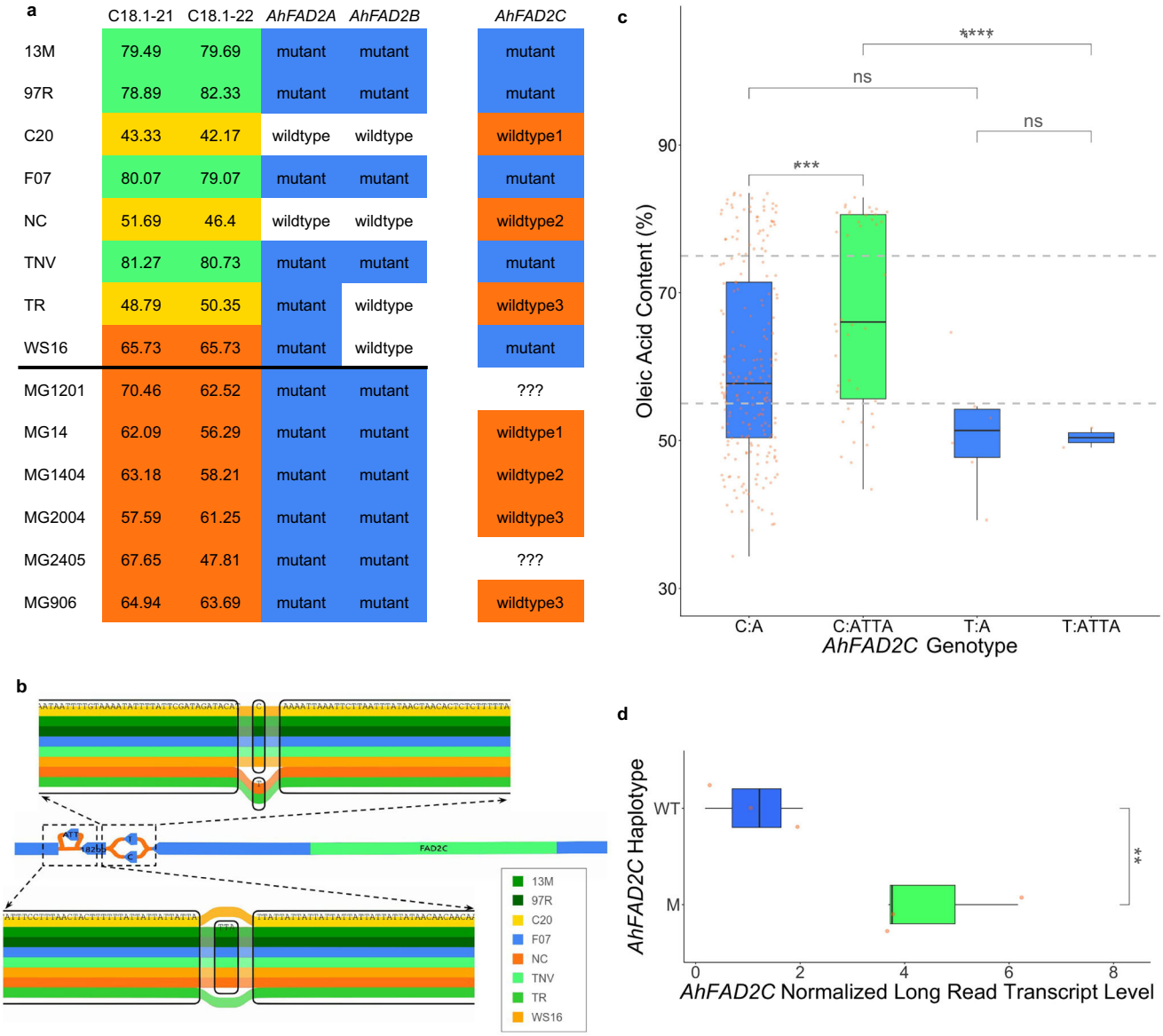


Fig. 5 | Functional variation of *AhFAD2C*. a Oleic acid phenotypes for the years 2021 and 2022 compared to high-throughput marker genotyping for *AhFAD2A* and *AhFAD2B* loci. The double mutant *AhFAD2A* and *AhFAD2B* calls for high-oleic (>75%) lines are not consistent in the peanut MAGIC Core lines. Additionally, mutant *AhFAD2A* and wildtype *AhFAD2B* do not result in consistent oleic content between the founders TR (low-oleic) and WS16 (mid-oleic). The *AhFAD2C* C:ATTA haplotype was indicated as Mutant whereas C:A, T:ATTA, T:A were marked as wildtype 1–3, respectively. **b** Pangenomic visualization of *AhFAD2C* highlighting promoter region variation and haplotypes across the founder lines using Sequence Tube Map (Copyright (c) 2018 Wolfgang Beyer, licensed under the MIT License; https://github.com/vgteam/sequenceTubeMap; no changes were made). **c** *AhFAD2C* MAGIC Core haplotypes compared to average oleic acid concentration. Individual phenotypes are represented as orange points. The *AhFAD2C* haplotype that all high-oleic founders possess is represented by a green bar. Error bars represent the standard error of the mean. Statistical significance was calculated using a two-sided Student's *t*-test.'***' represents a *P* value of 1.3×10^{-03} and '****' represents a *P* value of 4.6×10^{-05}. 'ns' represents no statistical significance, $n = 310$. Gray dashed lines represent 55% and 75% thresholds for mid and high-oleic phenotypes, respectively. **d** Long read RNA data of developing seed tissue. Transcript counts were normalized based on total transcripts per line. The three wildtype *AhFAD2C* variants (C20, NC, and TR) are represented as a blue boxplot and three mutant haplotypes for *AhFAD2C* (13M, TNV, and WS16) are represented as a green boxplot. Statistical significance was calculated using a two-sided Student's *t*-test. '**' represents a *P* value of 3.2×10^{-02}, $n = 6$. For each box plot in (**d**, **e**), the center line represents the median, the box bounds represent the interquartile range (IQR; 25th to 75th percentile), and the whiskers extend to the most extreme data point within that range. Source data are provided as a Source Data file.

C:ATTA haplotype (13M, TNV, and WS16) and the three low-oleic haplotypes (C20, C:A; TR, T:A; and NC, T:ATTA). In total, we generated over 95 million complete transcripts to evaluate expression levels and splice site information of FAD2 genes during seed development. We found that *AhFAD2C* is expressed in developing seed tissue and is differentially expressed in the C:ATTA haplotype compared to the low-oleic haplotypes. This suggests the variants can adjust transcription levels between the *AhFAD2C* haplotypes (Fig. 5d). This data supports the identification of a unique differentially expressed gene associated with oleic acid content using population-specific pangenome-based studies.

We found higher transcript levels in the high-oleic haplotype compared to the low-oleic haplotypes (Fig. 5d), which is counter-intuitive based on the mutant versions of *AhFAD2A* and *AhFAD2B* are understood as inhibitors of function. Upon further examination of the transcripts, it was found that all *AhFAD2C* transcripts from all

genotypes retained the previously annotated introns, where *AhFAD2B* transcripts were found to have the annotated splice sites (Supplementary Fig. 21). The consistent *AhFAD2C* introns within transcripts suggest that they are present during translation. However, within these regions, there are several stop codon signatures that create small open reading frames (3–160 amino acids) that likely do not create functional protein product (Supplementary Fig. 22). Additionally, we found relatively low levels of *AhFAD2C* compared to other annotated *AhFAD2* genes that may suggest their functional efficiency is high (Supplementary Table 9). Based on the increased transcript levels in C:ATTA haplotype compared to low-oleic haplotypes, retention of annotated introns that encode several stop codons, and lower overall transcript levels compared to annotated *AhFAD2* genes, we hypothesize that *AhFAD2C* may function as a negative regulator of other *AhFAD2* genes for high oleic acid phenotypes. Albeit the dataset of the long-read RNA is composed of a small number of genotype observations without replicates. We propose this mechanism for the function of *AhFAD2C*; however, future studies will be needed to clarify the function and potentially used to refine higher oleic acid phenotypes.

We further explored the variance of *AhFAD2C* with low coverage, long read whole genome sequencing of 94 MAGIC Core lines to verify the genotypes of the low coverage, short-read whole-genome sequencing. The long-read data was aligned with the PeanutMAGIC pangenome to generate calls for *AhFAD2C*. Of the six lines that were called mutant for *AhFAD2A* and *AhFAD2B*, four lines had information of the region, and two lines failed to have coverage for *AhFAD2C* variants (Supplementary Table 10). One line, MG1404 was genotyped with the long read data as mutant for *AhFAD2A*, *AhFAD2B*, and *AhFAD2C*, suggesting that this genotype is high-oleic. Previously, MG1404 was genotyped as mutant for *AhFAD2A* and *AhFAD2B* with wildtype *AhFAD2C* and exhibited a mid-oleic phenotype (Fig. 5a). Because of the change in the *AhFAD2C* genotype, we performed high-throughput seed phenotyping to determine if the mid-oleic line MG1404 was contaminated with high-oleic seed. We tested 839 MG1404 seeds and identified 226 seeds (26.94%) were high-oleic, suggesting that this line is no longer a pure line (Supplementary Fig. 23). The impurity of MG1404 exemplifies the persistent challenges which peanut breeders and growers face to ensure consistent quality for confectionary products. This contamination issue would have been undetected using only *AhFAD2A* and *AhFAD2B* genotyping, demonstrating the necessity of *AhFAD2C* to predict high-oleic phenotypes and to evaluate purity without phenotypic investment.

Discussion

Segregating populations are essential resources for trait mapping and genetic improvement in breeding. However, two major challenges must be addressed to effectively utilize these populations: increased recombination events and the accuracy of genotyping calling. The PeanutMAGIC population addresses these issues through fine recombination of multiple genetic donors and a comprehensive library of inheritable haplotypes. We found that using the extended pangenome at the species-level with additional genomes considerably increased the number of non-representative variants, potentially leading to more marker calling errors. The population-specific pangenome resulted in high-quality genotyping calling and enabled accurate and precise association.

In this study, we used pangenome-based markers to track the founder origin of chromosomal segments within a multiparent population. We also traced the founder origins for genomic regions that were resistant to recombination, offering unprecedented insight into the recombination of cultivated peanut. We highlighted the benefits of population-specific pangenome-based genotyping compared to single-reference genotyping for trait association studies. By relying on population-specific reference, we were able to reduce false signals made by single-reference association and offered increased resolution to identify *AhFAD2C*, near *AhFAD2B*, a crucial answer to the mystery of mid-oleic acid phenotypes in high-oleic breeding populations, which puzzled the community. This was validated by using a biparental 'S' population and highlighted the difficulties of biparental population studies to resolve closely linked putative functional genes. Although the mechanism behind *AhFAD2C* is inferred, future studies can validate the functionality and role of *AhFAD2C* to facilitate high-oleic phenotypes using robust differential transcriptomics, genetic transformation, and mutational studies. These findings highlight the power of a population-specific pangenome to capture genetic variation and recombination patterns that would be limited in a single-reference framework. We anticipate the use of population-specific pangenomes to become standard practice in synthetic inbreeding populations, outcrossing breeding populations, and natural populations research.

Methods

Plant materials and short read sequencing

The PeanutMAGIC population has been described in detail previously[13]. Briefly, eight founder parents: "Georgia-13M" (13M), "SunOleic 97R" (97R), "GT-C20" (C20), "Florida-07" (F07), "NC94022" (NC), "TifNV-High O/L" (TNV), "Tifrunner" (TR), and "GP-NC WS16" (WS16) were crossed in a simple funnel-like design. The critical 8-way cross consisted of 150 successful 4-way pairs, generating 950 unique F_1 offspring. The population was advanced to generate 3,187 $F_{2:7}$ RILs. A subset of 310 RILs were randomly selected for this study termed the MAGIC Core. These RILs were subject to CTAB DNA extraction followed by low-coverage (1×) Illumina sequencing.

The 'S' population has previously been described[11,46,49]. Briefly, this bioparental mapping population was derived from SunOleic 97R and NC94022 of 352 RILs. SunOleic 97R is a high-oleic fatty acid peanut cultivar. The two parents of the 'S' population are founders of the PeanutMAGIC. For this study, 144 RILs from the 'S' population were selected and sequenced at 5× with Illumina technology[46].

Founder assembly and pangenomes construction

The DNA of seven founders of the PeanutMAGIC was extracted using a high molecular-weight DNA extraction method and sequenced using PacBio Sequel II or Revio sequencing platforms at 25× coverage using the CCS mode to generate PacBio HiFi reads. The eighth founder Tifrunner (TR) (version 2 had previously been sequenced using PacBio and Illumina sequencing technologies[14]. PacBio reads were assembled using hifiasm (v0.19.6)[50] with the parameters '–u –l0'. The founders 13M, 97R, C20, and F07 were scaffolded with Hi-C data and polished using Juicer[51], Juicebox[52], and 3d-DNA[53] workflow. The founders NC, TNV, and WS16 were scaffolded using RagTag[25] with TR as the reference.

The resulting chromosome-level assemblies were aligned to TR as reference to assess quality and address erroneous assembly and orientation using minimap2[54] and dotPlotly for visualization. Assembly completeness was evaluated using BUSCO (v5.5.0)[26] with the "fabales_odb10" database.

A population-specific pangenome for PeanutMAGIC was constructed using the Mini-Graph Cactus Pangenome Pipeline (v2.6.7)[27,28] with TR as the reference genome. To generate a population-specific pangenome for the 'S' population, the 97R and NC assemblies were unified using the same pipeline, with NC as the reference. To assess the graphs vg (v1.56.0)[55] subcommands "stats" and "deconstruct"[56] were used to extract components and variant locations for each pangenome.

Syntenic block analysis

Synteny information was analyzed using the SynMap tool on the CoGe platform[57] using TR as reference. Only coding sequences were considered to identify unique gene parallels between subgenomes, minimizing oversampling of low-complexity regions.

Khufu pangenomic marker calling

KhufuPAN requires a graphical fragment assembly (GFA) file. The file should include the genomes forming the pangenome and a well-assembled reference genome (designated null reference), which is used to assign the coordinates of SNPs and structure variants.

Minigraph-Cactus (https://github.com/ComparativeGenomics/Toolkitcactus)[27] is recommended for creating the GFA file for the Khufu environment. The first step in the environment is KhufuPAN bootstrapping. The GFA is deconstructed to produce a parental VCF file. Then a series of filters is applied to remove odd alleles, low-quality variants, monomorphic variants, or those having polymorphism only with the null reference genome to generate a Filtered-Variant set. All files are packed into a single folder, which can be reused for multiple purposes.

KhufuVAR is a KhufuPAN sub-tool used to call and filter GFA variants for markers. Raw FASTQ files, short or long-reads, are applied to Khufu-core (https://www.hudsonalpha.org/khufudata/), and then mapped to the GFA file using vg giraffe (https://github.com/vgteam/vg/wiki/Mapping-short-reads-with-Giraffe)[58]. Calls under the minimum depth cutoff are masked, and variants overlapping with the Filtered-Variant set are extracted. Variants are segregated differently based on the population structure. Therefore, another series of filters are applied, i.e., minor allele frequency (>0.01), polymorphism, and the percentage of missing data (>75% missing). To efficiently utilize computational resources, KhufuVAR splits the process into two parts. Within the first part, samples run in batches, and every batch is aligned with the Filtered-Variant set and calculates measurements for the population filters. The second part combines the batches and runs the population filters. The final output calls are exported in Khufu panmap format, which facilitates pangenome-based markers for genomic analysis (Supplementary Method 1). The generated panmap file can then be used to extract a RDS file for Cyclops to visualize markers throughout the population in chromosomal units (https://w-korani.shinyapps.io/cyclops_eye_ii/).

MAGIC Core pangenome-based genomic characterization

To examine marker segregation within the MAGIC Core, r^2 values were calculated in a 300-marker sliding window on TASSEL 5.0 software[59] and plotted according to the physical location. The 300-marker window size was selected based on the increased marker density relative to the sliding window used in Thompson et al.[13]. To maintain comparable physical window sizes, the marker count per window was scaled up according to the average fold increase in markers (3×), simplifying comparison between single and pangenome references. For the 'S' population, r^2 values were calculated in a 100-marker sliding window to ensure equal physical distance window sizes.

The neighbor-joining tree was constructed in R using the *ape* package[60]. The workflow involved generating a distance matrix using "dist.gene", converting the matrix into a tree with 'nj', and plotting the tree with "plot", following the approach outlined in Thompson et al.[13].

Oleic fatty acid contents and GWAS analysis

To collect phenotypic data for association studies, the seeds were planted in three replications in the field for 2 years along with the parental lines. The harvested seeds were analyzed chemically in USDA-ARS laboratory in Griffin, GA, for oil composition and content by analyzing five seeds per line using gas chromatography (GC)[11].

PeanutMAGIC specific pangenome markers were used to calculate a distance matrix for multidimensional scaling (MDS) to control population structure. A general linear model was applied for GWAS analysis in TASSEL 5.0 software[59], using oleic acid contents and genotype data from three sources: MAGIC Core pangenome-based markers, MAGIC Core single-reference markers, and the 'S' pangenome markers. Each genotype source was paired with its respective phenotypes and MDS to maintain consistent parameters across associations.

PCR genotyping

Perfect markers for *AhFAD2A* and *AhFAD2B* were previously developed and optimized for peanut breeding applications[5]. Fresh leaf tissue was collected, and DNA was extracted using a NaOH solution and a TE buffer dilution. Diluted DNA was then used for melting curve and KASP genotyping techniques[5].

Personalized pangenome genotyping

Sequencing reads for peanut MAGIC Core were used to generate personalized pangenomes for each individual following the methods detailed in the vg GitHub wiki (https://github.com/vgteam/vg/wiki/Haplotype-Sampling)[48]. First, kmc was used to count *k*-mers from the sequence reads[61]. These *k*-mer counts were then incorporated into haplotype sampling with vg giraffe[58]. The subsequent GBZ and minimizer files were processed with vg "pack" to compute read support for genotyping[55]. Finally, genotyping calls were made using vg "call"[62]. This workflow generated personalized pangenomes and genotypes that were used to extract variation information for *AhFAD2C* across MAGIC Core and the 'S' populations.

Statistical testing

All phenotypic statistical analyzes were conducted and visualized using R statistical software (v4.4.1)[63] with the *ggplot2* package[64]. ANOVA was performed with the base "aov" function. The Student's *t*-test was conducted using the "stat_compare_means(method = "t.test")" function from the *ggpubr* package[65]. The post-hoc Tukey's HSD was performed with the base 'TukeyHSD' function.

Long read RNA sequencing

Total RNA was extracted from 6 founder genotypes 13M, C20, NC, TNV, TR, and WS16 from seed tissue at the seed development stage Patee 7[66] using Qiagen RNeasy extraction kit (https://www.qiagen.com/us). The purified RNA was then prepared for sequencing on Pacbio Revio using Kinnex full-length RNA kit (https://www.pacb.com/wp-content/uploads/Procedure-checklist-Preparing-Kinnex-libraries-using-the-Kinnex-full-length-RNA-kit.pdf). To identify FAD2 genes two methods were used. Transcripts were aligned to TR using minimap2 and counted for annotated FAD2 genes[54]. The transcripts were also extracted using unique sequences found for each annotated FAD2 gene, to verify minimap2 alignments. The read counts were normalized by dividing the number of identified reads by the total number of reads in the sample and multiplying it by 1 million.

Whole genome long read low coverage sequencing

High molecular weight DNA was extracted from 94 MAGIC Core lines using the unofficial Pacbio high-throughput gDNA workflow for plants (https://www.pacb.com/support/documentation/). The subsequent DNA was then multiplexed and formed libraries using Pacbio Hifi plex prep kit for Revio sequencing. The reads were then separated into their respective genotypes to generate raw long reads. The long reads were mapped to the PeanutMAGIC pangenome using the personalized pangenome genotyping approach and verified using minimap2[54].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All raw sequencing data and assembly results of genomes generated in this study have been deposited in the National Cener for Biotechnology Information under the accession number PRJNA1212195 [https://www.ncbi.nlm.nih.gov/bioproject/1212195]. Source data are provided with this paper.

References

1. Jones, J. B. et al. A randomized trial on the effects of flavorings on the health benefits of daily peanut consumption. *Am. J. Clin. Nutr.* **99**, 490–496 (2014).
2. Kris-Etherton, P. M., Hu, F. B., Ros, E. & Sabaté, J. The role of tree nuts and peanuts in the prevention of coronary heart disease: multiple potential mechanisms. *J. Nutr.* **138**, 1746–1751 (2008).
3. Alper, C. M. & Mattes, R. D. Peanut consumption improves indices of cardiovascular disease risk in healthy adults. *J. Am. Coll. Nutr.* **22**, 133–141 (2003).
4. Davis, J. P. et al. Peanut oil stability and physical properties across a range of industrially relevant oleic acid/linoleic acid ratios. *Peanut Sci.* **43**, 1–11 (2016).
5. Chu, Y. et al. Marker-assisted selection to pyramid nematode resistance and the high oleic trait in peanut. *Plant Genome* **4**, 110–117 (2011).
6. Jung, S., Powell, G., Moore, K. & Abbott, A. The high oleate trait in the cultivated peanut [*Arachis hypogaea* L.]. II. molecular basis and genetics of the trait. *Mol. Gen. Genet.* **263**, 806–811 (2000).
7. López, Y. et al. Isolation and characterization of the Δ12-fatty acid desaturase in peanut (*Arachis hypogaea* L.) and search for polymorphisms for the high oleate trait in spanish market-type lines. *Theor. Appl. Genet.* **101**, 1131–1138 (2000).
8. Chu, Y., Ramos, L., Holbrook, C. C. & Ozias-Akins, P. Frequency of a loss-of-function mutation in oleoyl-PC desaturase (ahFAD2A) in the mini-core of the U.S. peanut germplasm collection. *Crop Sci.* **47**, 2372–2378 (2007).
9. Chu, Y., Holbrook, C. C. & Ozias-Akins, P. Two alleles of *ahFAD2B* control the high oleic acid trait in cultivated peanut. *Crop Sci.* **49**, 2029–2036 (2009).
10. Branch, W. D., Brown, N. & Perrera, M. A. Inheritance of mid-oleic fatty acid ratio seed trait in peanut. *Peanut Sci.* **49**, 26–29 (2022).
11. Pandey, M. K. et al. Identification of QTLs associated with oil content and mapping FAD2 genes and their relative contribution to oil quality in peanut (*Arachis hypogaea* L.). *BMC Genet.* **15**, 133 (2014).
12. Scott, M. F. et al. Multi-parent populations in crops: a toolbox integrating genomics and genetic mapping with breeding. *Heredity* **125**, 396–416 (2020).
13. Thompson, E. et al. Genetic and genomic characterization of a multiparent advanced generation intercross (MAGIC) population of peanut (*Arachis hypogaea* L.). *Crop Sci.* **65**, 1 (2024).
14. Bertioli, D. J. et al. The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nat. Genet.* **51**, 877–884 (2019).
15. Bertioli, D. J. et al. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet.* **48**, 438–446 (2016).
16. Danilevicz, M. et al. Plant pangenomics: approaches, applications and advancements. *Curr. Opin. Plant Biol.* **54**, 18–25 (2020).
17. Guo, L. et al. Super pangenome of *Vitis* empowers identification of downy mildew resistance genes for grapevine improvement. *Nat. Genet.* **57**, 741–753 (2025).
18. Branch, W. D. Registration of 'Georgia-13M' peanut. *J. Plant Regist.* **8**, 253–256 (2014).
19. Gorbet, D. W. & Knauft, D. A. Registration of 'SunOleic 97R' peanut. *Crop Sci.* **40**, 1190–1191 (2000).
20. Liang, X. Q., Holbrook, C. C., Lynch, R. E. & Guo, B. Z. β-1,3-glucanase activity in peanut seed (*Arachis hypogaea*) is induced by inoculation with *Aspergillus flavus* and copurifies with a conglutin-like protein. *Phytopathology* **95**, 506–511 (2005).
21. Gorbet, D. W. & Tillman, B. L. Registration of 'Florida-07' peanut. *J. Plant Regist.* **3**, 14–18 (2009).
22. Culbreath, A. K. et al. High levels of field resistance to tomato spotted wilt virus in peanut breeding lines derived from *hypogaea* and *hirsuta* botanical varieties. *Peanut Sci.* **32**, 20–24 (2005).
23. Holbrook, C. C. et al. Registration of 'TifNV-High O/L' peanut. *J. Plant Regist.* **11**, 228–230 (2017).
24. Tallury, S. P. et al. Registration of two multiple disease-resistant peanut germplasm lines derived from *Arachis cardenasii* Krapov. & W.C. Gregory GKP 10017. *J. Plant Regist.* **8**, 86–89 (2014).
25. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
26. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
27. Hickey, G. et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).
28. Armstrong, J. et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).
29. Clevenger, J. P. & Ozias-Akins, P. SWEEP: a tool for filtering high-quality SNPs in polyploid crops. *G3: Genes Genomes Genet.* **5**, 1797–1803 (2015).
30. Clevenger, J. P., Korani, W., Ozias-Akins, P. & Jackson, S. Haplotype-based genotyping in polyploids. *Front. Plant Sci.* **9**, 564 (2018).
31. Pandey, M. K. et al. Genetic dissection of novel QTLs for resistance to leaf spots and tomato spotted wilt virus in peanut (*Arachis hypogaea* L.). *Front. Plant Sci.* **8**, 25 (2017).
32. Korani, W. et al. De novo QTL-seq identifies loci linked to blanchability in peanut (*Arachis hypogaea*) and refines previously identified QTL with low coverage sequence. *Agronomy* **11**, 11 (2021).
33. Liu, Z. et al. Grapevine pangenome facilitates trait genetics and genomic breeding. *Nat. Genet.* **56**, 2804–2814 (2024).
34. Li, X. et al. Large-scale gene expression alterations introduced by structural variation drive morphotype diversification in Brassica oleracea. *Nat. Genet.* **56**, 517–529 (2024).
35. Khan, A. W. et al. *Cicer* super-pangenome provides insights into species evolution and agronomic trait loci for crop improvement in chickpea. *Nat. Genet.* **56**, 1225–1234 (2024).
36. Zhang, Y. et al. Telomere-to-telomere *Citrullus* super-pangenome provides direction for watermelon breeding. *Nat. Genet.* **56**, 1750–1761 (2024).
37. Zhuang, W. et al. The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat. Genet.* **51**, 865–876 (2019).
38. Chen, X. et al. Sequencing of cultivated peanut, *Arachis hypogaea*, yields insights into genome evolution and oil improvement. *Mol. Plant* **12**, 920–934 (2019).
39. Newman, C. S. et al. Initiation of genomics-assisted breeding in virginia-type peanuts through the generation of a de novo reference genome and informative markers. *Front. Plant Sci.* **13**, 1073542 (2023).
40. Clevenger, J. et al. Gene expression profiling describes the genetic regulation of *Meloidogyne arenaria* resistance in *Arachis hypogaea* and reveals a candidate gene for resistance. *Sci. Rep.* **7**, 1317 (2017).
41. Chu, Y. et al. Identification of rare recombinants leads to tightly linked markers for nematode resistance in peanut. *Peanut Sci.* **43**, 88–93 (2016).
42. Simpson, C. E. Pathways for introgression of pest resistance into *Arachis hypogaea* L.1. *Peanut Sci.* **18**, 22–26 (1991).
43. Zhang, H., Yu, Y., Wang, M., Dang, P. & Chen, C. Effect of genotype-by-environment interaction on oil and oleic fatty acid contents of cultivated peanuts. *Horticulturae* **9**, 1272 (2023).
44. Wang, M. L. et al. Genotype, environment, and their interaction effects on peanut seed protein, oil, and fatty acid content variability. *Agron. J.* **116**, 1440–1454 (2024).

45. Clevenger, J., Chu, Y., Scheffler, B. & Ozias-Akins, P. A developmental transcriptome map for allotetraploid *Arachis hypogaea*. *Front. Plant Sci.* **7**, 1446 (2016).
46. Agarwal, G. et al. A recombination bin-map identified a major QTL for resistance to tomato spotted wilt virus in peanut (*Arachis hypogaea*). *Sci. Rep.* **9**, 18246 (2019).
47. Khera, P. et al. Mapping quantitative trait loci of resistance to tomato spotted wilt virus and leaf spots in a recombinant inbred line population of peanut (*Arachis hypogaea* L.) from SunOleic 97R and NC94022. *PLoS ONE* **11**, e0158452 (2016).
48. Sirén, J. et al. Personalized pangenome references. *Nat. Methods* **21**, 2017–2023 (2024).
49. Qin, H. et al. An integrated genetic linkage map of cultivated peanut (*Arachis hypogaea* L.) constructed from two RIL populations. *Theor. Appl. Genet.* **124**, 653–664 (2012).
50. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
51. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
52. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
53. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
54. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
55. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
56. Liao, W. W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
57. Lyons, E. et al. Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.* **148**, 1772–1781 (2008).
58. Sirén, J. et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, 6574 (2021).
59. Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
60. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **34**, 526–528 (2018).
61. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating *k*-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
62. Hickey, G. et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).
63. R. Core Team. R. Core Team (2014). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. http://www.R-project.org/. (2014).
64. Wickham, H. Ggplot2: create elegant data visualizations using the grammar of graphics. *Wiley Interdiscip. Rev. Comput. Stat.* **3**, 1 (2011).
65. Kassambara, A. ggpubr: 'ggplot2' based publication ready plots. R package version 0.2. https://CRAN.R-project.org/package=ggpubr (2020).
66. Pattee, H. E., Johns, E. B., Singleton, J. A. & Sanders, T. H. Composition changes of peanut fruit parts during maturation. *Peanut Sci.* **1**, 57–62 (1974).

Acknowledgements

We would like to acknowledge Billy Wilson and Hui Wang for maintaining and progressing the PeanutMAGIC population. Stephanie Botton and Sydney Webb for KASP marker work. Akshaya Biswal and Carlos Cardon for RNA extraction and purification. Kendell Lee, Zack Myers, Paige Phelps, Tajinder Singh, and Yangbing Wang for their work with long read DNA and RNA sequencing. David Bertioli and Mark Mackiewicz for comments and suggestions for the manuscript. This research was partially supported by the US Department of Agriculture Agricultural Research Service (USDA-ARS), the Peanut Research Foundation, the Georgia Peanut Commission, and National Peanut Board (B.G.). Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The USDA is an equal opportunity employer and provider.

Author contributions

B.G., J.P.C., and R.K.V. conceived and designed the study. E.T., W.K., D.W., and V.G. performed data analyses. E.T., W.K., and J.P.C. performed genome assembly. J.P.C. and E.T. contributed to Hi-C alignment. E.T. constructed the pangenomes. W.K. performed KhufuPAN genotype calling. B.T. and M.W. performed oil chemical analysis. E.T. generated the original figures, supplementary data, and wrote the first draft of the manuscript with input from D.W., V.G., B.G., J.P.C., R.K.V., C.C.H., P.O-A., and A.K.C. All the authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-67371-7.

Correspondence and requests for materials should be addressed to Rajeev K. Varshney, Baozhu Guo or Josh P. Clevenger.

Peer review information *Nature Communications* thanks Jeffrey Dunne, Liang Guo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025