nature genetics

ANALYSIS

https://doi.org/10.1038/s41588-020-0612-7

Identifying genetic variants underlying phenotypic variation in plants without complete genomes

Yoav Voichek and Detlef Weigel ✉

Structural variants and presence/absence polymorphisms are common in plant genomes, yet they are routinely overlooked in genome-wide association studies (GWAS). Here, we expand the type of genetic variants detected in GWAS to include major deletions, insertions and rearrangements. We first use raw sequencing data directly to derive short sequences, *k*-mers, that mark a broad range of polymorphisms independently of a reference genome. We then link *k*-mers associated with phenotypes to specific genomic regions. Using this approach, we reanalyzed 2,000 traits in *Arabidopsis thaliana*, tomato and maize populations. Associations identified with *k*-mers recapitulate those found with SNPs, but with stronger statistical support. Importantly, we discovered new associations with structural variants and with regions missing from reference genomes. Our results demonstrate the power of performing GWAS before linking sequence reads to specific genomic regions, which allows the detection of a wider range of genetic variants responsible for phenotypic variation.

GWAS support the systematic identification of candidate genomic loci responsible for phenotypic variation. A difficulty with plants is that their genomes are characterized by many structural variants (SVs), which can often cause phenotypic variation[1]. Although not usually analyzed, short sequencing reads can provide, in principle, information on many more variants in their source genomes than SNPs and short insertions/deletions (indels)[2]. Variants are typically discovered with short reads by mapping them to a reference genome, but common subsequences can also be directly compared among samples[3,4]. Such a direct approach is intuitively most powerful when there is no or only a poor reference genome assembly. Because short reads result from random shearing of genomic DNA, and because they contain sequencing errors, directly comparing short reads between two samples is not very effective. Instead, genetic variants in a population can be discovered by focusing on sequences of constant length *k* that are shorter than the original reads, termed *k*-mers. After *k*-mers have been extracted from all reads, *k*-mer sets from different samples can be compared against each other. Importantly, *k*-mers present in some samples, but missing from others, can identify a broad range of genetic variants. For example, two genomes differing in a SNP (Fig. 1a and Extended Data Figs. 1 and 2) will have *k*-mers unique to each genome, even if the SNP is found in a repeated region or a region not present in the reference genome. SVs such as large deletions, inversions and translocations, will also result in *k*-mer differences. Therefore, instead of defining genetic variants in a population relative to a reference genome, a *k*-mer presence/absence pattern in raw sequencing data can be directly associated with phenotypes to enlarge the tagged genetic variants in GWAS[5].

Reference-free GWAS based on *k*-mers have been used with bacteria, which have many dispensable genes[5–7]. They have also been applied to human genomes, which are much larger and have many more unique *k*-mers[3,8], but were restricted to case–control situations, and because of high computational load, not all *k*-mers were corrected for population structure. While *k*-mer-based approaches are likely to be especially appropriate for plants, the large genomes, highly structured populations and excessive genetic variation in plants[9–11] make the use of existing *k*-mer methods difficult. An attempt to use *k*-mer methods in plants was limited to a small subset of the genome and also accounted for population structure only for a small subset of *k*-mers[12].

Here, we present an efficient method for *k*-mer-based GWAS and compare it directly to the conventional SNP-based approach on more than 2,000 phenotypes from three species with different genome and population characteristics—*A. thaliana*, maize and tomato. In brief, we inverted the conventional approach of building a genome, using it to find population variants and finally associating variants with phenotypes. In contrast, we begin by associating sequencing reads with phenotypes and only then infer the genomic context of associated sequences. We posit that this change of order is especially effective in plants, for which defining the full population-level genetic variation based on reference genomes remains highly challenging.

Results

Comparison of SNP and *k*-mer genome-wide association on *A. thaliana* phenotypes. For an initial proof of concept, we examined a model trait in *A. thaliana*, flowering time. We used an existing dataset[13] to define the presence/absence patterns of 31-bp *k*-mers in over 1,000 inbred accessions. Of a total of 2.26 billion unique *k*-mers, 439 million appeared in at least five accessions (Extended Data Figs. 3a and 4). Using the presence or absence of a *k*-mer as two allelic contrasts, we performed genome-wide association analysis with a linear mixed model (LMM) to account for population structure (Extended Data Fig. 3b)[14] and compared it to analysis with SNPs and short indels (Fig. 1b).

To define a set of *k*-mers most likely to be associated with flowering time, we had to set a *P*-value threshold. Unfortunately, a single genetic variant is typically tagged by several *k*-mers, and the Bonferroni threshold would not accurately reflect the effective number of independent tests. To account for non-independence, we defined a threshold based on permutations of the phenotype[15]. This is computationally challenging, as the full genome-wide association analysis has to be run many times. We therefore implemented an LMM-based genome-wide association analysis specifically optimized for the *k*-mer application (Extended Data Fig. 3c)[16,17].

Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany. ✉e-mail: weigel@weigelworld.org

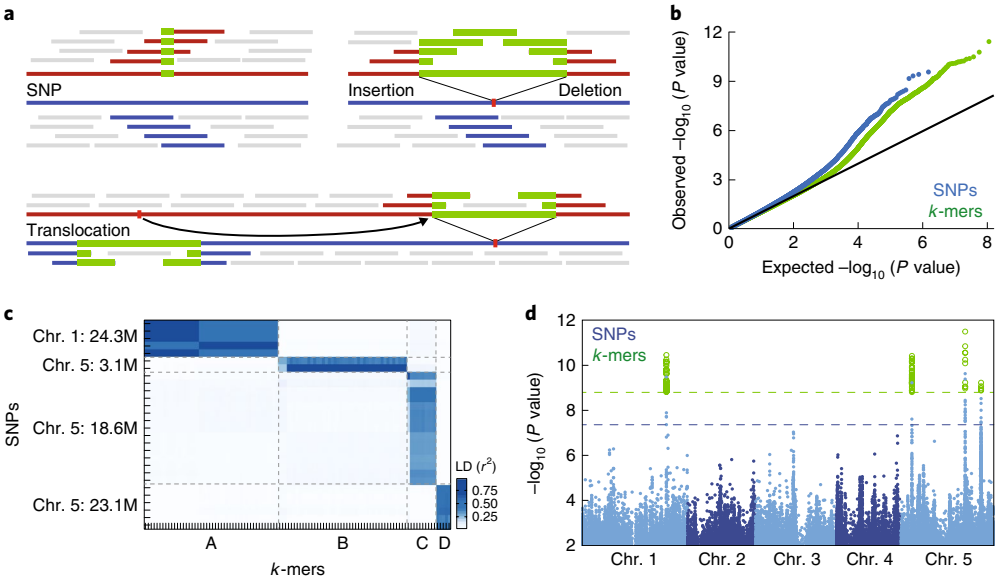


Fig. 1 | Flowering time associations in *A. thaliana*. **a**, *k*-mers and different genetic variants. The blue and red lines represent two individual genomes. The colored short bars mark *k*-mers unique to each genome, and the gray bars mark *k*-mers shared between genomes. **b**, *P*-value quantile–quantile plot of SNP and *k*-mer associations with flowering time at 10 °C. Deviation from $y = x$ indicates stronger than chance associations. **c**, LD between SNPs and *k*-mers passing *P*-value thresholds. Both methods identified four highly linked families of variants. Extended Data Fig. 5a,b shows SNP-to-SNP and *k*-mer-to-*k*-mer LD. **d**, *P* values of all SNPs and the subset of *k*-mers passing the *P*-value threshold as a function of their genomic position. The dashed lines mark the thresholds for SNPs (blue) and *k*-mers (green). Extended Data Fig. 5c,d shows separate plots for *k*-mers and SNPs.

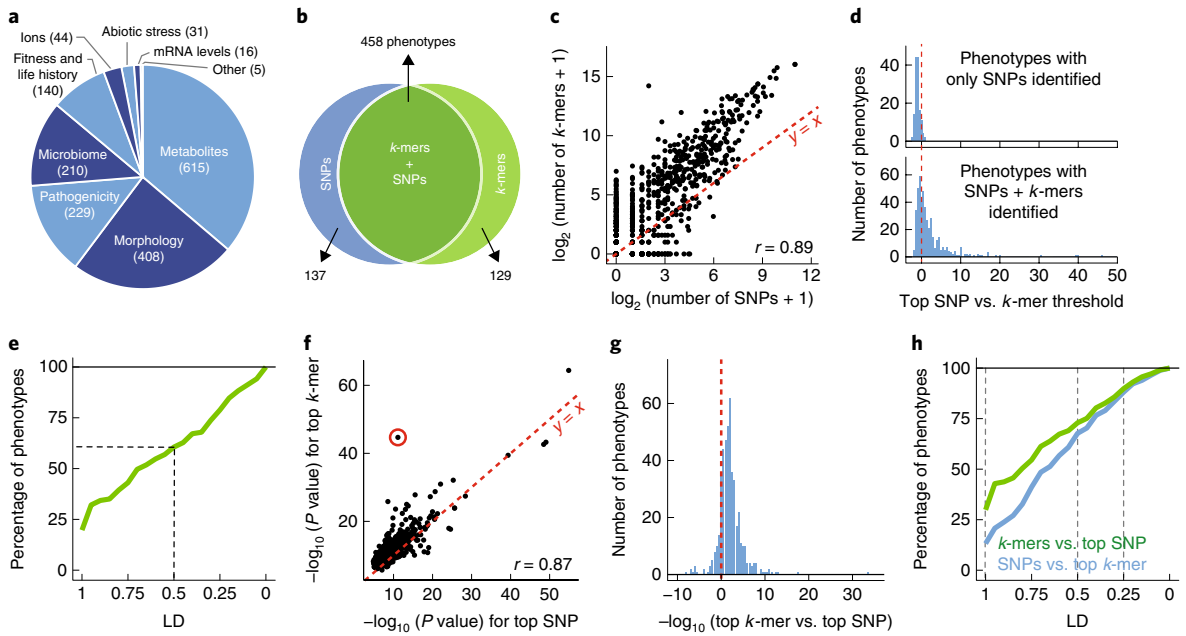


Fig. 2 | SNP- and *k*-mer-based GWAS on 1,582 *A. thaliana* phenotypes. **a**, Categories of collected phenotypes. **b**, Overlap between phenotypes with SNP and *k*-mer hits. **c**, Correlation of numbers of significant *k*-mers versus SNPs. **d**, Ratios (in $\log_{10}$) of the top SNP *P* value versus the *k*-mer threshold for 137 phenotypes with only significant SNPs (top) and for 458 phenotypes with significant SNPs and *k*-mers (bottom). **e**, The percentage of 137 phenotypes that had only significant SNP hits, for which a *k*-mer passing the SNP threshold could be found within different LD cutoffs. For a minimum LD = 0.5 (dashed lines), 61% of phenotypes had a linked *k*-mer that passed the SNP threshold. **f**, Correlation of *P* values of top *k*-mers with SNPs ($r = 0.87$). The red circle marks the strongest outlier (see Fig. 3a,b). **g**, Ratio between top *P* values (expressed as $-\log_{10}$) for the two methods for 458 phenotypes with *k*-mer and SNP hits. **h,** The percentage of all phenotypes for which a significant SNP could be found within different LD cutoffs of the top *k*-mer (blue) and vice versa (green). The dashed lines mark LD = 1, 0.5 and 0.25, with the percentage of phenotypes being 29%, 73% and 90% for the green curve, and 13%, 67% and 88% for the blue curve, respectively.

We calculated the *P*-value thresholds for SNPs and *k*-mers, with a 5% chance of one false positive. The threshold for *k*-mers was higher than for SNPs (35-fold), but lower than the increase in test number (140-fold) due to the higher dependency between *k*-mers (Fig. 1a). Twenty-eight SNPs and 105 *k*-mers passed their corresponding thresholds. Using linkage disequilibrium (LD), we directly

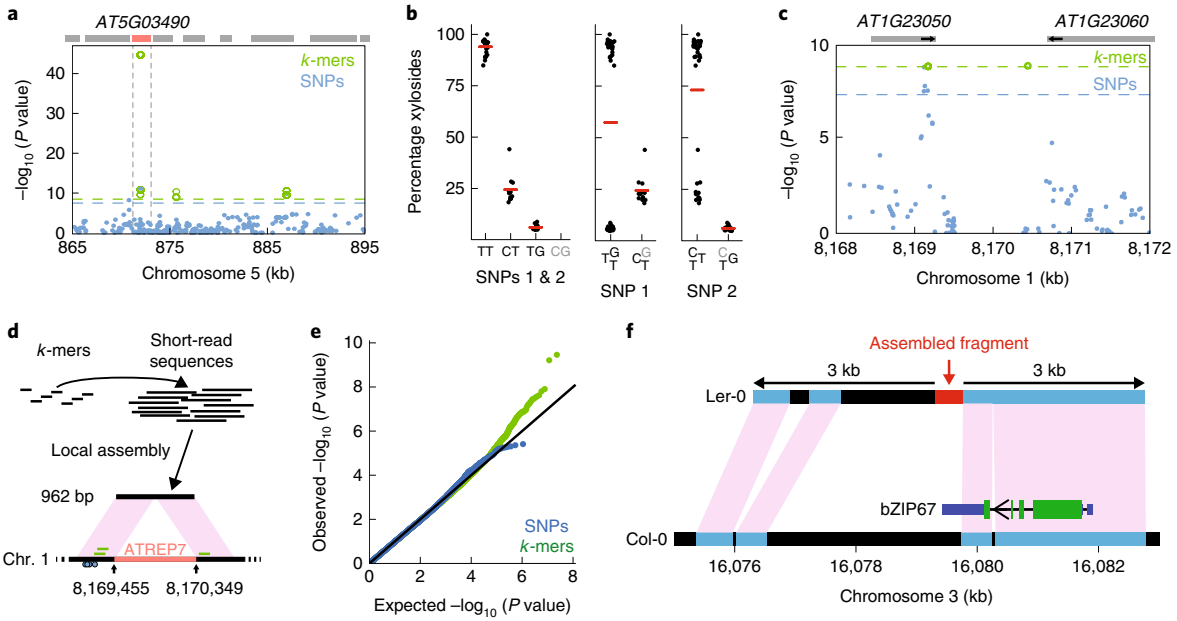


Fig. 3 | Specific cases of *k*-mer superiority. a, Associations with xyloside fraction in the indicated region. Boxes on top indicate genes, with *AT5G03490* marked in red. **b**, Xyloside percentage grouped by the states at two adjacent SNPs (872,003 and 872,007 bp), with averages marked by red bars. The left plot shows simultaneous grouping based on both SNPs, as is possible with *k*-mers. The middle and right plots show groupings based on each SNP alone. Haplotypes in this region will not capture this association (Extended Data Fig. 7b). **c**, Associations with seedling growth inhibition in the presence of flg22. Absence of SNPs in the central 1-kb region is likely due to the presence of a TE to which short reads cannot be unambiguously mapped. Gene orientations are indicated with short black arrows. **d**, Assembly of reads identified with the seven unmappable *k*-mers resulted in a 962-bp fragment that lacked the central 892-bp region from the reference genome, with similarity to the ATREP7 TE. The small blue circles on the bottom represent significant flanking SNPs, and the short green bars above represent the three significant mappable *k*-mers. **e**, *P*-value quantile–quantile plot of associations with germination time in darkness and low levels of nutrients. Only *k*-mers show stronger than expected associations. **f**, Assembled reads (red bar) containing significant *k*-mers from genome-wide association of germination time (**e**) match a region on chromosome 3 of Ler-0. Other regions that cannot be aligned between the reference genomes are indicated in black. The 3′ UTR of the gene encoding bZIP67 in Col-0 is indicated in dark blue; its exact extent in Ler-0 is unknown. Coding sequences are marked in green.

linked SNPs to *k*-mers without locating the *k*-mers in the genome. Four families of linked genetic variants were identified with both methods (Fig. 1c). As expected, the *k*-mers tagged the same genomic loci as the corresponding SNPs (Fig. 1d), and with similar results for 25-bp *k*-mers (Extended Data Fig. 5e). Therefore, *k*-mers identified the same genotype–flowering time associations as SNPs.

To increase the chances of discovering new associations, we evaluated 1,582 phenotypes from 104 *A. thaliana* studies (Supplementary Table 1 and Fig. 2a). There was substantial overlap between significant SNP and *k*-mer associations (Fig. 2b), and the numbers of *k*-mers and SNPs for each phenotype were highly correlated (Fig. 2c and Extended Data Fig. 6a). For 137 phenotypes, only a significant SNP could be identified, likely due to the more stringent thresholds for *k*-mers, as the most significant SNPs rarely passed the *k*-mer threshold in these cases (Fig. 2d). Moreover, a *k*-mer passing the SNP threshold was in high LD with the top SNP (Fig. 2e). Although the *k*-mer thresholds were more stringent than the SNP thresholds (Extended Data Fig. 6b), only *k*-mer associations were identified for 129 phenotypes. The *P* values of top SNPs and *k*-mers were highly correlated (Fig. 2f), with top *k*-mers having a lower *P* value in almost nine of ten cases (Fig. 2g). In addition, we found that, on average, associated *k*-mers were closer to top SNPs than the other way around: 29% of top SNPs were in complete LD with associated *k*-mers whereas 13% of top *k*-mers were in complete LD with associated SNPs; the corresponding proportions were 73% and 67% at LD ≥ 0.5 (Fig. 2h). This is consistent with *k*-mers often containing the top SNP, while SNPs in many cases were only linked to the causal variant identified by *k*-mers.

Case studies of *k*-mer superiority. In addition to simply improving the strength of associations (Extended Data Fig. 7a), we sought to identify cases where *k*-mers provided a conceptual improvement. We first looked at the fraction of dihydroxybenzoic acid (DHBA) xylosides among total DHBA glycosides (Fig. 2f)[18]. In this case, all significant *k*-mers mapped uniquely near *AT5G03490*, encoding a UDP-glycosyltransferase, already identified in the original study (Fig. 3a and Extended Data Fig. 7c). The stronger *k*-mer associations could be traced back to two nonsynonymous SNPs, 4 bp apart, in the gene's coding region. Due to their proximity, one *k*-mer held the state of both SNPs, and their combined information was more predictive of the phenotype than either SNP alone (Fig. 3b).

Our next case study involved seedling growth in the presence of a flg22 variant[19], for which we could map only three of the ten significant *k*-mers to the reference genome, in the proximity of significant SNPs in *AT1G23050* (Fig. 3c and Extended Data Fig. 7d). To identify the genomic source of the seven remaining *k*-mers, we assembled the short reads from which they originated. The resulting 962-bp fragment also included the three mappable *k*-mers (Fig. 3d), but did not contain a 892-bp helitron transposable element (TE)[20] present in the reference genome. While the *k*-mer method did not identify a new locus, it revealed an SV as the likely cause of differences in flg22 sensitivity.

Finally, we looked for phenotypes for which only significant *k*-mers were identified. One was germination in darkness under low nutrient supply[21], for which none of the 11 *k*-mers identified (Fig. 3e and Extended Data Fig. 7e,f) could be traced back to the reference genome. The reads containing these *k*-mers assembled into a 458-bp fragment that had a hit in the genome of Ler-0,

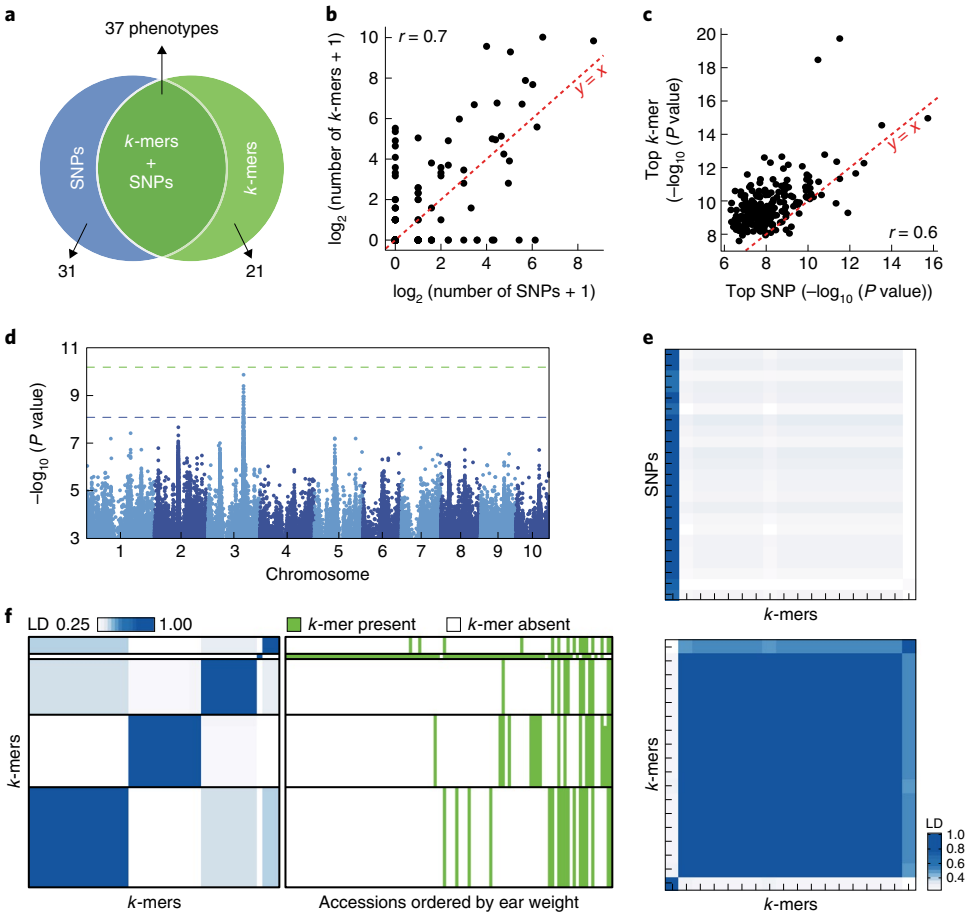


Fig. 4 | SNP- and *k*-mer-based GWAS in maize. a, Overlap between SNP and *k*-mer hits (Extended Data Fig. 8b,c). **b**, Correlation of the numbers of significant *k*-mers versus SNPs. See Extended Data Fig. 8e. **c**, Correlation of *P* values of top *k*-mers and SNPs. **d**, SNP associations with days to tassel (environment 06FL1). Dashed lines represent the SNP (blue) and *k*-mer (green) thresholds. **e**, LD between 23 significant SNPs and 18 *k*-mers (top) or between *k*-mers and *k*-mers (bottom) for days to tassel. The order of the *k*-mers is the same in both heat maps. **f**, LD between 45 *k*-mers associated with ear weight (environment 07A; left) and *k*-mer presence/absence patterns in different accessions ordered by ear weight (right).

a non-reference accession[22]. The flanking sequences were syntenic with the reference genome, with a 2-kb SV that included the assembled 458-bp fragment (Fig. 3f). This variant affected the 3′ UTR of the gene encoding the bZIP67 transcription factor. Accumulation of the bZIP67 protein but not *bZIP67* mRNA appears to mediate environmental regulation of germination[23]; an SV in the 3′ UTR is consistent with translational regulation of bZIP67. This case demonstrates the ability of our *k*-mer method to reveal associations with SVs not tagged by SNPs.

***k*-mer-based GWAS in maize.** To demonstrate the usefulness of our approach with larger, more complex genomes, we analyzed maize[24], which has a genome of ~2.5 Gb and extensive presence/absence variation of genes[10,25,26]. We applied our approach to 252 mostly morphological traits[27] in 150 inbred lines with short-read sequence coverage of at least 6× (ref. [28]). A total of 2.3 billion *k*-mers were present in at least five accessions (Extended Data Fig. 8a). Significant associations were identified for 89 traits by at least one of the methods and for 37 traits by both methods (Fig. 4a). As in *A. thaliana*, statistically significant variants as well as top associations were well correlated between the two methods (Fig. 4b,c and Extended Data Fig. 8b–d). Top *k*-mers had lower *P* values than top SNPs (Extended Data Fig. 8e), and the *k*-mer method detected associations not found by SNPs.

A major challenge for maize is the high fraction of short reads that do not map uniquely to the genome. Previously, additional information was used to find the genomic position of SNPs, including population LD and genetic map position[28]. We therefore compared SNPs and *k*-mers using LD, without locating *k*-mers in the genome. In several cases, a *k*-mer marked a common allele in the population with strong phenotypic effects, without the allele having been identified with SNPs. For example, for days to tassel, one clear SNP hit was also tagged by *k*-mers (Fig. 4d,e) but a second variant was only identified with *k*-mers. Another example is ear weight, for which no SNP was found (Extended Data Fig. 8f) but several unlinked *k*-mer-tagged variants were identified (Fig. 4f). Thus, new alleles with high predictive power for maize traits can be revealed using *k*-mers.

As with SNPs, the difficulty of unique short-read mapping also undermined our ability to identify the source of *k*-mers associated with specific traits. For example, we attempted to locate the genomic position of the *k*-mer corresponding to the SNP associated with days to tassel on chromosome 3 (Fig. 4d). Only about 1% of reads from which the *k*-mers originated could be mapped uniquely to the reference genome. However, when we assembled all originated reads into a 924-bp contig, we could place it to the same position as the identified SNPs. This fragment had two single-base-pair differences relative to the reference genome, and was not located near any gene. Thus, we could use the richness of combining reads from several accessions to more precisely locate variant origin.

***k*-mer-based GWAS in tomato.** At ~900 Mb, the tomato genome is smaller than that of maize, but it presents its own challenges, as there is a complex history of introgressions from wild relatives into

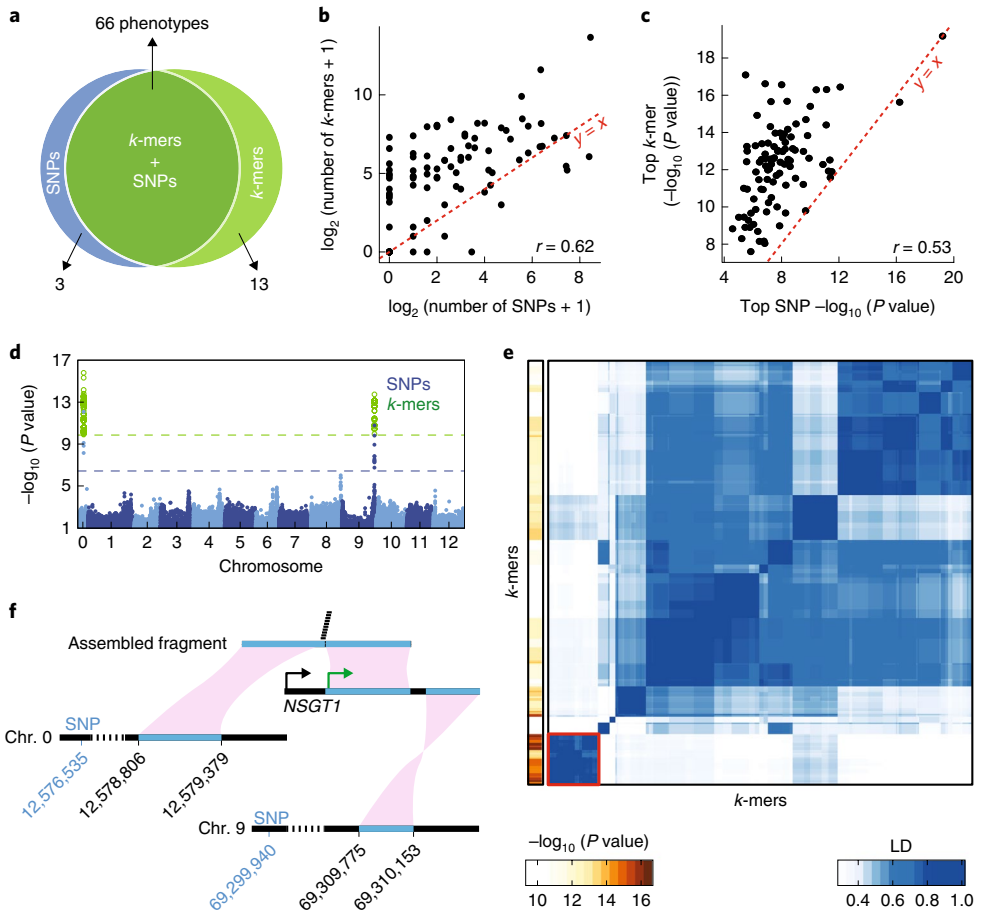


Fig. 5 | SNP- and *k*-mer-based GWAS in tomato. a, Overlap between SNP and *k*-mer hits (Extended Data Fig. 9b,c). **b**, Correlation of the numbers of significant *k*-mers versus SNPs. See Extended Data Fig. 9e. **c**, Correlation of *P* values of top *k*-mers and SNPs. **d**, SNP and *k*-mer associations with guaiacol concentration. Dashed lines represent the SNP (blue) and *k*-mer (green) thresholds. **e**, LD among 293 *k*-mers associated with guaiacol concentration (right) and *P* values of each *k*-mer (left). The red square (bottom left) indicates the 35 *k*-mers with the lowest *P* values and no mapping to the reference genome. **f**, Part of a fragment assembled from the 35 unmapped *k*-mers (**e**) mapped to chromosome 0 and another part mapped to the unanchored complete *NSGT1* gene from a non-reference accession. Only the 3′ end of *NSGT1* is found in the reference genome, on chromosome 9. The green and black arrows mark the start of the *NSGT1* ORF in the R104 'smoky' and 'non-smoky' lines[33]. The two significant SNPs closest to the two regions in the reference genome are indicated in blue.

domesticated tomatoes[29,30]. Starting with 981 million *k*-mers from 246 accessions (Extended Data Fig. 9a), we performed genome-wide association analysis on 96 metabolite measurements[31,32]. For many metabolites, an association was identified by both methods, but 3 had only SNP hits and 13 only *k*-mer hits (Fig. 5a). Similarly to the other species, the number of identified variants as well as top *P* values were correlated between methods (Fig. 5b,c). Top *k*-mer associations were also stronger than associations for top SNPs (Extended Data Fig. 9d), even more so than was seen in *A. thaliana* or maize.

In a case study, we examined the concentration of guaiacol, responsible for a strong off-flavor in tomato[31]. Associated SNPs were found on chromosome 9 and on 'chromosome 0' (Fig. 5d), which contains sequence scaffolds not assigned to the 12 nuclear chromosomes. Of the 293 guaiacol-associated *k*-mers, 180 could be mapped uniquely to the genome, all close to significant SNPs. Among the remaining *k*-mers, of particular interest was a group of 35 *k*-mers in high LD and with especially low *P* values (Fig. 5e). Assembly of the corresponding short reads resulted in a 1,172-bp fragment, of which the first 574 bp aligned near significant SNPs in chromosome 0 (Fig. 5f) and the remainder matched the non-reference *NSGT1* (non-smoky glycosyltransferase1) gene, which had been originally pinpointed as causal for variation in guaiacol[33]. The 35 significant *k*-mers covered the junction between these two mappable regions. Most of the *NSGT1* coding sequence is absent from the reference genome but present in other accessions. The significant SNPs identified on chromosomes 0 and 9 apparently represent the same region in other accessions, connected by the fragment we assembled (Fig. 5f). Thus, we identified an association outside of the reference genome and linked the SNPs on chromosome 0 to chromosome 9.

***k*-mer-based kinship estimates.** We have shown that the assembly of short fragments from *k*-mer-containing short reads reveals hits not only in the reference genome, but also in other published sequences. This opens up the possibility of applying our method to species without a high-quality reference genome, as contigs that include multiple genes can be relatively easily and cheaply generated[34]. The major question with such an approach is then how to correct for population structure in the genome-wide association step without kinship information from SNPs, determined by mapping to a reference genome. To learn kinship directly from *k*-mers, we estimated relatedness using *k*-mers, with presence or absence as the two alleles. We calculated relatedness matrices for *A. thaliana*,

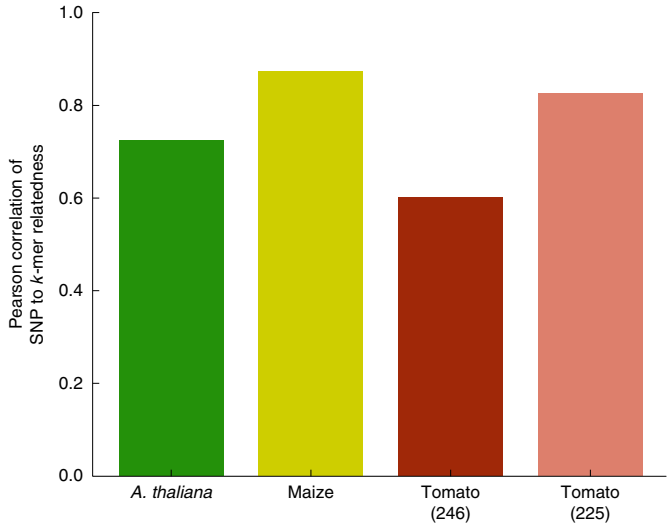


Fig. 6 | Kinship matrix estimates with *k*-mers. Relatedness between accessions was independently estimated based on SNPs and *k*-mers. The correlation between the two methods for tomato could be improved by removing 21 accessions (Extended Data Fig. 10).

maize and tomato and compared them to the SNP-based relatedness. In all three species, there was agreement between the two methods, although initial results were clearly better for *A. thaliana* and maize than for tomato (Fig. 6). The inferior performance in tomato was due to 21 accessions (Extended Data Fig. 10) that appeared to be more distantly related to the other accessions when kinship was estimated on the basis of *k*-mers rather than SNPs. This is likely because these accessions contain diverged genomic regions that perform poorly in SNP calling, resulting in inaccurate relatedness estimates. In conclusion, *k*-mers can be used to calculate relatedness between individuals, thus paving the way for GWAS in organisms without high-quality reference genomes.

Discussion

The complexity of plant genomes can make SNP-based identification of genotype–phenotype associations challenging. We have shown that *k*-mers can identify not only almost all associations found by SNPs and short indels but also SVs and variants in sequences not present in reference genomes. The expansion of variant types detected by the *k*-mer method complements SNP-based approaches and increases opportunities for finding and exploiting complex genetic variants driving phenotypic differences in plants, regardless of reference genome quality.

k-mers mark genetic polymorphisms in the population, but the types and genomic positions of these polymorphisms are initially not known. While one can also use *k*-mers for predictive models without knowing their genomic context, in many cases the genomic context of associated *k*-mers is of interest. The simplest solution is to align *k*-mers or the corresponding short reads to a reference genome[35]. Of interest are cases where *k*-mers cannot be placed on the reference genome. For these, one can first identify the originating short reads and assemble these into larger fragments, which is an effective path to uncover the genomic context of *k*-mers. The resulting fragment also captures phased haplotype information. Combining reads from multiple accessions can provide high local coverage around *k*-mers of interest, increasing the chances that sizeable fragments can be assembled and located.

A further improvement will be the use of *k*-mers to tag heterozygous variants. In our current implementation, which relies on presence/absence of *k*-mers, one of the homozygous states has to be clearly differentiated not only from the alternative homozygous state but also from the heterozygous state. This did not affect our comparisons between SNPs and *k*-mers in this study, as we only looked at inbred populations where homozygous, binary states are expected. Another improvement will be the use of *k*-mers to detect causal copy number variations. So far, we can only tag copy number variants if the junctions produce unique *k*-mers, but it would be desirable to also use *k*-mers inside copy number variants. Normalized *k*-mer counts would create a framework that could, at least in principle, detect almost any kind of genomic variation.

The comparison of *k*-mer- and SNP-based GWAS provides an interesting view on trade-offs in the characterization of genetic variability. The lower top *P* values obtained with *k*-mers where a SNP is the underlying variant suggest incomplete use of existing information in SNP calling. On the other hand, our analysis likely included some *k*-mers that represent only sequencing errors. While requiring *k*-mers to appear multiple times in a sequencing library and in multiple individuals removes most sequencing errors, this can also lead to some *k*-mers being labeled erroneously as absent. Finally, the increase in test load is an inevitable result of increasing the search space to tag more genetic variants.

k-mer-based approaches invert how GWAS are usually done. Instead of first locating sequence variations in the genome, we begin with sequence–phenotype associations and only then find the genomic context of associated sequences. Technological improvements in short- and long-read sequences, as well as methods to integrate them into a population-level genetic variation data structure, will expand the covered genetic variants[36,37]. While traditional GWAS methods will benefit from these improvements, so will *k*-mer-based approaches, which will be able to use tags spanning larger genomic distances. Therefore, we posit that, for GWAS, *k*-mer-based approaches are ideal because they minimize arbitrary choices when classifying alleles and they capture more, almost optimal, information from raw sequencing data.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0612-7.

Received: 31 October 2019; Accepted: 10 March 2020;
Published online: 13 April 2020

References

1. Saxena, R. K., Edwards, D. & Varshney, R. K. Structural variations in plant genomes. *Brief. Funct. Genomics* **13**, 296–307 (2014).
2. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226–232 (2012).
3. Salzberg, S. L., Pertea, M., Fahrner, J. A. & Sobreira, N. DIAMUND: direct comparison of genomes to detect mutations. *Hum. Mutat.* **35**, 283–288 (2014).
4. Zielezinski, A. et al. Benchmarking of alignment-free sequence comparison methods. *Genome Biol.* **20**, 144 (2019).
5. Lees, J. A. et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat. Commun.* **7**, 12797 (2016).
6. Sheppard, S. K. et al. Genome-wide association study identifies vitamin B_5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc. Natl Acad. Sci. USA* **110**, 11923–11927 (2013).
7. Lees, J. A. et al. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *eLife* **6**, e26255 (2017).
8. Rahman, A., Hallgrímsdóttir, I., Eisen, M. & Pachter, L. Association mapping from sequencing reads using *k*-mers. *eLife* **7**, e32920 (2018).
9. Gordon, S. P. et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184 (2017).
10. Sun, S. et al. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nat. Genet.* **50**, 1289–1295 (2018).

11. Minio, A., Massonnet, M., Figueroa-Balderas, R., Castro, A. & Cantu, D. Diploid genome assembly of the wine grape Carménère. *G3* **9**, 1331–1337 (2019).
12. Arora, S. et al. Resistance gene cloning from a wild crop relative by sequence capture and association genetics. *Nat. Biotechnol.* **37**, 139–143 (2019).
13. 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491 (2016).
14. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
15. Abney, M. Permutation testing in the presence of polygenic variation. *Genet. Epidemiol.* **39**, 249–258 (2015).
16. Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. & Aulchenko, Y. S. Rapid variance components-based method for whole-genome association analysis. *Nat. Genet.* **44**, 1166–1170 (2012).
17. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
18. Li, X. et al. Exploiting natural variation of secondary metabolism identifies a gene controlling the glycosylation diversity of dihydroxybenzoic acids in *Arabidopsis thaliana*. *Genetics* **198**, 1267–1276 (2014).
19. Vetter, M., Karasov, T. L. & Bergelson, J. Differentiation between MAMP-triggered defenses in *Arabidopsis thaliana*. *PLoS Genet.* **12**, e1006068 (2016).
20. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
21. Morrison, G. D. & Linder, C. R. Association mapping of germination traits in *Arabidopsis thaliana* under light and nutrient treatments: searching for G×E effects. *G3 (Bethesda)* **4**, 1465–1478 (2014).
22. Zapata, L. et al. Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. *Proc. Natl Acad. Sci. USA* **113**, E4052–E4060 (2016).
23. Bryant, F. M., Hughes, D., Hassani-Pak, K. & Eastmond, P. J. Basic LEUCINE ZIPPER TRANSCRIPTION FACTOR67 transactivates *DELAY OF GERMINATION1* to establish primary seed dormancy in *Arabidopsis*. *Plant Cell* **31**, 1276–1288 (2019).
24. Schnable, P. S. et al. The B73 maize genome: complexity, diversity and dynamics. *Science* **326**, 1112–1115 (2009).
25. Gore, M. A. et al. A first-generation haplotype map of maize. *Science* **326**, 1115–1117 (2009).
26. Springer, N. M. et al. The maize W22 genome provides a foundation for functional genomics and transposon biology. *Nat. Genet.* **50**, 1282–1288 (2018).
27. Zhao, W. et al. Panzea: a database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Res.* **34**, D752–D757 (2006).
28. Bukowski, R. et al. Construction of the third-generation *Zea mays* haplotype map. *Gigascience* **7**, 1–12 (2018).
29. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012).
30. Lin, T. et al. Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* **46**, 1220–1226 (2014).
31. Tieman, D. et al. A chemical genetic roadmap to improved tomato flavor. *Science* **355**, 391–394 (2017).
32. Zhu, G. et al. Rewiring of the fruit metabolome in tomato breeding. *Cell* **172**, 249–261 (2018).
33. Tikunov, Y. M. et al. Non-smoky glycosyltransferase1 prevents the release of smoky aroma from tomato fruit. *Plant Cell* **25**, 3067–3078 (2013).
34. Sohn, J.-I. & Nam, J.-W. The present and future of de novo whole-genome assembly. *Brief. Bioinform.* **19**, 23–40 (2018).
35. Pascoe, B. et al. Enhanced biofilm formation and multi-host transmission evolve from divergent genetic backgrounds in *Campylobacter jejuni*. *Environ. Microbiol.* **17**, 4779–4789 (2015).
36. Schneeberger, K. et al. Simultaneous alignment of short reads against multiple genomes. *Genome Biol.* **10**, R98 (2009).
37. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **27**, 665–676 (2017).

Methods

Curation of an *A. thaliana* phenotype compendium. Studies containing phenotypic data on *A. thaliana* accessions were located in PubMed using a set of general search terms. For most studies, relevant data were obtained from the supplementary information; otherwise, requests were sent to the corresponding authors. Data already uploaded to the AraPheno dataset[38] were downloaded. Phenotypic data in PDF format were extracted using Tabula software. Different sets of naming for accessions were converted to accession indices. In the case where an index for an accession could not be located, we omitted the corresponding data point. In the case where an accession could potentially be assigned to different indices, we first checked whether it was part of the 1001 Genomes (1001G) project; if so, we used the 1001G index, but otherwise one of the possible indices was randomly assigned. Phenotypes of metabolite measurements from two studies[39,40] were filtered to a reduced set by the following procedure: taking the first phenotype and sequentially retaining phenotypes if the correlation with all previously taken phenotypes was lower than 0.7. Data from the second study[40] were further filtered for phenotypes with a title. The assignment of categories for each phenotype was done manually (Supplementary Table 1). All processed phenotypic data can be found at https://zenodo.org/record/3701176#.XmX9u5NKhhE.

Whole-genome sequencing data and variant calls of *A. thaliana*. Whole-genome short reads for 1,135 *A. thaliana* accessions were downloaded from the NCBI Sequence Read Archive (SRA; accession SRP056687). Accessions with fewer than 10^8 unique *k*-mers (a proxy for low effective coverage) were removed, resulting in a set of 1,008 accessions. The 1001G VCF file with SNPs and short indels was downloaded (http://1001genomes.org/data/GMI-MPI/releases/v3.1) and condensed into these accessions using vcftools v0.1.15 (ref. [41]). We required a minor allele count (MAC) of five individuals, resulting in 5,649,128 genetic variants. The VCF file was then converted to a PLINK binary file using PLINK v1.9 (ref. [42]). The TAIR10 reference genome was used for short-read and *k*-mer alignments. The coordinates for genes in figures were taken from Araport11[43].

Whole-genome sequencing data and variant calls of maize. Whole-genome short reads of maize accessions corresponded to the '282' set of the maize HapMap3.2.1 project[28]. Sequencing libraries '×2' and '×4' were downloaded from NCBI SRA (accession PRJNA389800) and combined. Coverage per accession was calculated as the number of reads multiplied by read length and divided by the genome size. Data for 150 accessions with coverage of > 6× were used. Phenotypic data for the 252 traits measured for these accessions were downloaded from Panzea[27].

Two phenotypes were constant over more than 90% of the 150 accessions, and these were removed from further analysis (NumberofTilleringPlants_env_07A, TilleringIndex-BorderPlant_env_07A). The HapMap3.2.1 VCF files (c*_282_corrected_onHmp321.vcf.gz) of SNPs and indels were downloaded from Cyverse. Variant files were filtered using vcftools v0.1.15 to the relevant 150 accessions. Variants were further filtered for a MAC of ≥ 5, resulting in a final set of 35,522,659 variants. The B73 reference genome, version AGPv3 (ref. [44]), which was used to create the VCF file, was downloaded from MaizeGDB and used for alignments[44].

Whole-genome sequencing data and variant calls of tomato. Whole-genome short reads were downloaded for 246 accessions with coverage of > 6× from NCBI SRA and the European Nucleotide Archive (ENA) portal of the European Bioinformatics Institute (accession numbers SRP045767, PRJEB5235 and PRJNA353161). A table with coverage per accession was shared by the authors[31]. Metabolite measurements (adjusted values) were taken from Tieman et al.[31], as well as a subset from Zhu et al.[32]. These were filtered to a reduced set according to the procedure described previously for *A. thaliana*. Metabolites were ordered as reported originally[32]. Only the repeat with the most data points, and requiring at least 40, was retained. The VCF file with SNPs and short indels[31] was obtained from the authors and filtered for the relevant 246 accessions. Variants were further filtered for a MAC of ≥ 5, resulting in a final set of 2,076,690 variants. The reference genome SL2.5 (ref. [29]; https://www.ncbi.nlm.nih.gov/assembly/GCF_000188115.3/) used to create the VCF file was used for alignments.

Calculation and comparison of kinship matrices. Kinship matrices of relatedness between accessions were calculated as in EMMA[46] using default parameters. The algorithm was recoded in C++ to directly read PLINK binary files. For *k*-mer-based relatedness, the same algorithm was used, coding presence/absence as two alleles. For comparison of *k*-mer- and SNP-based relatedness, we correlated (Pearson) the values for all $\frac{n}{2}$ pairs for *n* accessions. For tomato, 3,492 pairs had a relatedness of more than 0.15 lower for *k*-mers than for SNPs. Approximately 3,300 (94.4%) of these pairs were between a set of 21 accessions and all other 225 accessions. We calculated the correlation twice: for all pairs and for pairs of these 225 accessions.

Genome-wide association on SNPs and short indels or on the full *k*-mers table. Genome-wide association on the full set of SNPs and short indels was conducted using LMMs with the kinship matrix in GEMMA (v0.96)[14]. Minor allele frequency (MAF) and MAC were set to 5% and 5, respectively, with a maximum of 50% missing values (-miss 0.5). To run genome-wide association on the full set of *k*-mers (for example, in Fig. 1b), *k*-mers were first filtered for those having only unique presence/absence patterns on the relevant set of accessions, a MAF of at least 5% and a MAC of at least 5. Presence/absence patterns were then condensed to only the relevant accessions and output directly as a PLINK binary file. GEMMA was then run using the same parameters as for the SNP genome-wide association described above.

Phenotype covariance matrix estimation and phenotype permutation. EMMA (emma.REMLE function) was used to calculate the variance components, which were used to calculate the phenotypic covariance matrix[46]. We then calculated 100 permutations of the phenotype using the mvnpermute R package[15]. The *n*% (for example, *n* = 5 gives 5%) family-wise error-rate threshold was defined by taking the *n*th top *P* value from the 100 top *P* values of running genome-wide association analysis on each permutation. In all cases, unless indicated otherwise, the 5% threshold was used.

Scoring *P* values from genome-wide association analysis for similarity to uniform distribution and filtering phenotypes. For each SNP-based genome-wide association analysis, we scored for a general bias in *P*-value distribution, similarly to Atwell et al.[47]. All SNP *P* values were collected; the 99% higher *P* values were tested against the uniform distribution using a Kolmogorov–Smirnov test; and the test statistic was used to filter phenotypes for which the distribution deviated significantly from the uniform distribution. A threshold of 0.05 was applied, filtering 89, 0 and 295 phenotypes for *A. thaliana*, maize and tomato, respectively.

***k*-mer genome-wide association analysis.** Genome-wide association of *k*-mers was done in two steps, with the aim of selecting the *k*-mers with the most significant *P* values. The first step was based on the approach used in Bolt-lmm-inf or GRAMMAR-Gamma[16,17]. For phenotypes *y*, genotypes *g* and a covariance matrix **Ω**, the *k*-mer score is:

$$T^2_{\text{score}} = \frac{1}{\gamma}\frac{(\tilde{g}^T\Omega^{-1}\tilde{y})^2}{\tilde{g}^T\tilde{g}}$$

where $\tilde{g} = g - E(g)$ and $\tilde{y} = y - E(y)$. The first step was used only to filter a fixed number of top *k*-mers; thus, we could use any score monotonous with T^2_{score} and specifically $\frac{(\tilde{g}^T\Omega^{-1}\tilde{y})^2}{\tilde{g}^T\tilde{g}}$, which is independent of γ (see the Supplementary Note on calculation optimization and Supplementary Table 3). In the second step, the best *k*-mers were run using GEMMA to calculate the likelihood ratio test *P* values[14].

The number of *k*-mers filtered in the first step was set to 10,000 for *A. thaliana* and 100,000 for maize and tomato. Both steps associated *k*-mers while accounting for population structure and, while the first step used an approximation, the second used an exact model. Therefore, real top *k*-mers may be lost, as they would not pass the first filtering step. To control for this, we first defined the 5% family-wise error-rate threshold based on the phenotype permutations and then identified all of the *k*-mers that passed the threshold. Next, we used the following criteria to minimize the chance of losing *k*-mers: we checked whether all the identified *k*-mers were in the top N/2 *k*-mers from the ordering of the first step ($N = 10{,}000$ or 100,000, depending on species). For example, in maize, all *k*-mers passing the threshold in the second step should be in the top 50,000 *k*-mers from the first step. The probability that this will happen randomly is 2^{-m}, where *m* is the number of identified *k*-mers, which is unlikely in most phenotypes. In 8.5% of phenotypes from *A. thaliana*, the criteria were not fulfilled and for these phenotypes we reran both steps with 100× more *k*-mers filtered (1,000,000) in the first step. For six phenotypes, the criteria still did not hold and these were not used in further analysis. In tomato, 33% of phenotypes did not fulfill these criteria. In these cases, we reran the first step with 100× more *k*-mers filtered (10,000,000), and 17 phenotypes still did not pass the threshold and were omitted from further analysis. The permutations were not rerun, and the threshold defined using 100,000 *k*-mers was used, as the top *k*-mer used to define the threshold tended to be high in the list. For maize, all phenotypes passed the criteria without rerunning.

SNP-based GWAS on phenotype permutations. To calculate thresholds for SNP-based GWAS, we used the two-step approach as used for *k*-mers. The permuted phenotypes were run in two steps, as we were only interested in the top *P* value to define thresholds. We filtered 10,000 variants in the first step, which were then run using GEMMA to get the exact scores[14]. The non-permuted phenotypes were run using GEMMA on all the variants.

Calculation of linkage disequilibrium. Two variants, *x* and *y*, can either be a *k*-mer or a SNP. For a *k*-mer, variants were coded as 0 or 1 if absent or present, respectively. For SNPs, one variant was coded as 0 and the other as 1. If one of the variants had a missing or heterozygous value in a position, this position was not used in the analysis. LD was calculated using the r^2 measure[48]. The LD value was calculated using the formula:

$$r^2 = \frac{(p(x = 1 \wedge y = 1) - p(x = 1) \times p(y = 1))^2}{p(x = 1) \times p(y = 1) \times p(x = 0) \times p(y = 0)}$$

Comparing Col-0 and Ler genome assemblies with *k*-mers. The lists of 31-bp *k*-mers that are part of the Col-0 TAIR10 and the Ler genomes[22] were created using KMC v3 (ref. [45]). The lists from both genomes were filtered for *k*-mers appearing in a single genome and those appearing only once in a genome. The positions of the filtered *k*-mers were identified by checking each position in the genome against the filtered lists. In Extended Data Fig. 1, *k*-mers from these lists are plotted around four variants, as defined previously[22]. The statistics presented in Extended Data Fig. 2 are for all variants reported in the supplementary tables of Zapata and colleagues[22], under the titles 'Lindel_Allelic', 'Lindel_NonAllelic', 'IntraChromTransloc', 'InterChromTransloc' and 'InversionSites'.

Calculating linkage disequilibrium of the closest SNP/*k*-mer. In Supplementary Fig. 1, to calculate LD between all *k*-mers and SNPs in the *A. thaliana* 1001G project, the 1001G-imputed SNPs matrix was used[49] (provided by Ü. Seren) to avoid dealing with missing values in the original VCF file. The imputed matrix was condensed to the 1,008 accessions used in the *k*-mers table, and only SNPs with MAF ≥ 0.05 were considered. *k*-mers were also filtered for MAF ≥ 0.05. We were left with 898,869 SNPs and 163,644,699 *k*-mers; therefore, the complexity of calculating all LD was $(898{,}869)\times(163{,}644{,}699)\times(1{,}008)\geq 10^{17}$. This calculation was done using the SSE4 command set by representing the variant per individual as one bit and combining 64 individuals in one CPU word. Only the maximal LD of each SNP to all *k*-mers and of each *k*-mer to all SNPs were saved.

Linkage disequilibrium cumulative graph. In Fig. 2e,h, for a set of phenotypes and for every $l=0, 0.05, \ldots, 1$, we calculated the percentage of phenotypes for which there was a *k*-mer or a SNP in the predefined group, which had $LD\geq l$ with the top SNP or top *k*-mer, respectively. The predefined groups were (1) all of the *k*-mers that passed the SNP-defined threshold in Fig. 2e or (2) all of the SNPs or *k*-mers that passed their own defined thresholds in Fig. 2h. The percentage was then plotted as a function of *l*.

Retrieving source reads of a specific *k*-mer and assembling them. For a *k*-mer identified as being associated with a phenotype, we first looked in the *k*-mers table and identified all accessions included in the association analysis and having this *k*-mer present. For each of these accessions, we went over all sequencing reads and filtered out all paired-end reads that contained the *k*-mer. To assemble paired reads, SPAdes v3.11.1 was used with the '--careful' parameter[50].

Alignment of reads or *k*-mers to the genome. Paired-end reads were aligned to the genome using bowtie2 v2.2.3 with the '-very-sensitive-local' parameter. *k*-mers were aligned to the genome using bowtie v1.2.2 with the '-best-all-strata' parameter[51].

Analysis of flowering time in 10C. In Fig. 1 and Extended Data Fig. 5, to find the location of the 105 identified *k*-mers in the genome, *k*-mers were first mapped to the *A. thaliana* genome. Of the 105 *k*-mers, 84 had a unique mapping, 1 was mapped to multiple locations and 20 could not be mapped. For the 21 *k*-mers with no unique mapping, we located the sequencing reads that they originated from and mapped these to the *A. thaliana* genome. For each *k*-mer, we looked only at the reads with the top mapping scores. For the single *k*-mer that had multiple possible alignments, the originating reads did not have a consensus mapping location in the genome. For every *k*-mer from the 20 non-mapped *k*-mers, all top reads per *k*-mer (in some cases except one) mapped to a specific region spanning a few hundred base pairs. The middle of this region was defined as the *k*-mer position for the Manhattan plot in Fig. 1d. To find the locations of all *k*-mers presented in Extended Data Fig. 5d, we used only uniquely mapped *k*-mers.

To find the location of the 93 associated *k*-mers of length 25 bp (Extended Data Fig. 5e), we followed the same procedure: 87 *k*-mers had unique mapping, 1 mapped multiple times and 5 could not be mapped. For the 5 non-mappable *k*-mers and the *k*-mer with non-unique mapping, we located the short reads from which they originated and aligned them to the genome. For each of the 5 *k*-mers, all reads with the top mapping score were mapped to a specific region of a few hundred base pairs, and we took the middle of the region as the location in the Manhattan plot. For the *k*-mer with multiple mappings, 15 of 17 reads mapped to the same region and we used this location. All *k*-mers mapped to the 4 locations in the genome for which SNPs were identified, except one—AAGCTACTTGGTTGATAATACTAAT; the reads from which this *k*-mer originated mapped to the same region on chromosome 5 at position 3,191,745–3,192,193 and we used the middle of this region.

Analysis of the xyloside fraction. In Fig. 3a,b and Extended Data Fig. 7b,c, all *k*-mers passing the threshold were mapped uniquely to chromosome 5 in the region 871,976–886,983. Of the 123 identified *k*-mers, 27 had the same minimal *P* value ($-\log_{10} P$ value = 44.7). These *k*-mers mapped to chromosome 5 in positions 871,976–872,002, all covering the region 872,002–872,007. For the 60 accessions used in this analysis, all reads from 1001G were mapped to the reference genome. The mapping in region 872,002–872,007 of chromosome 5 was examined manually by Integrative Genomics Viewer (IGV) in all accessions[52]. The SNPs at 872,003 and 872,007 were called manually without knowledge of the phenotype value.

The hierarchical clustering in Extended Data Fig. 7b was done according to all SNPs on chromosome 5 from position 870,000–874,000. The distance between every two accessions was defined as the average number of SNPs with different values, taking into account only SNPs with no missing values.

Analysis of growth inhibition in the presence of flg22. In Fig. 3c,d, the phenotype in the original study was labeled 'flgPsHRp' (ref. [19]). For each of the seven *k*-mers that could not be mapped uniquely to the genome, the originated reads from all accessions were retrieved and assembled. All seven cases resulted in the same assembled fragment (SEQ1; Supplementary Table 2). Using NCBI BLAST, we mapped this fragment to chromosome 1: position 40–265 was mapped to 8,169,229–8,169,455 and position 262–604 was mapped to 8,170,348–8,170,687. For every accession from the 106 that were used in the genome-wide association analysis, we tried to locally assemble this region to see whether the junction between chromosome 1 positions 8,169,455–8,170,348 could be identified. We used all 31-bp *k*-mers from the above assembled fragment as bait and located all the reads for each accession separately. For 11 of the 13 accessions that had all ten identified *k*-mers, we got a fragment from the assembly process. In all 11 cases, the exact same junction was identified. For one of the four accessions that had only part of the ten identified *k*-mers, we got a fragment from the assembler that had the same junction. For 43 of the 89 accessions that had none of the identified *k*-mers, the assembly process resulted in a fragment, but in none of these cases could the above junction be identified.

Analysis of germination in darkness and low nutrients. In Fig. 3e,f, the phenotype in the original study was labeled 'k_light_0_nutrient_0' (ref. [21]). The 11 identified *k*-mers had two possible presence/absence patterns, separating them into two groups of four and seven *k*-mers. The short-read sequences containing the four or seven *k*-mers were collected separately and assembled, resulting in the same 458-bp fragment (SEQ2; Supplementary Table 2). This fragment was used as a query in NCBI BLAST search, resulting in alignment to Ler-0 chromosome 3 (LR215054.1) position 15,969,670–15,970,128. The region from 15,969,670–3000bp to 15,970,128 + 3000 bp in LR215054.1 was retrieved and used as the query for a NCBI BLAST search. The fragment mapped to Col-0 reference genome chromosome 3 (CP002686.1). Region 1–604 mapped to 16,075,369–16,075,968, region 930–1445 mapped to 16,076,025–16,076,532, region 3,446–3,946 mapped to 16,079,744–16,080,244 and region 3,958–6,459 mapped to 16,080,301–16,082,781.

Analysis of root branching zone. In Supplementary Fig. 2, the phenotype in the original study was labeled 'Mean(R)_C', that is, branching zone in no treatment[53]. No SNPs and one *k*-mer (AGCTACTTTGCCACCCACTGCTACTAACTCG) passed their corresponding 5% thresholds. The *k*-mer mapped to the chloroplast genome at position 40,297, with one mismatch. No SNPs and another *k*-mer (CCGGCGATTACTAGAGATTCCGGCTTCATGC) passed the 10% family-wise error-rate threshold. This *k*-mer mapped non-uniquely to two places in the chloroplast genome: 102,285 and 136,332.

Analysis of lesion by *Botrytis cinerea* UKRazz. In extended Data Fig. 7a, the lesion by *Botrytis cinerea* UKRazz phenotype was labeled as 'Lesion_redgrn_m_theta_UKRazz'[39]. In the genome-wide association analysis, 19 *k*-mers and no SNPs were identified. All *k*-mers had the same presence/absence pattern. The short-read sequences from which the *k*-mers originated were mapped to chromosome 3 around position 72,000, and contained a 1-bp deletion of a T nucleotide in position 72,017. Whole-genome sequencing reads were mapped to the genome for the 61 accessions with phenotypes used in these analyses. We manually observed the alignment around position 72,017 of chromosome 3, without knowing whether the accession had the identified *k*-mers. For 20 accessions, we observed the 1-bp deletion in position 72,017, and all 19 accessions containing the *k*-mers were part of these 20.

Analysis of days to tassel and ear weight in maize. In Fig. 4, ear weight phenotype was labeled 'EarWeight_env_07A' in the original dataset[27]. Days to tassel were measured in growing degree days (GDD) and labeled as 'GDDDaystoTassel_env_06FL1' in the original dataset. In a comparison of LD between *k*-mers and SNPs in days to tassel (Fig. 4e, top), two SNPs were filtered out as having more than 10% heterozygosity and one as having exactly 50% missing values. In days to tassel, the *k*-mer that was similar to the identified SNPs was AGAAGATATCTTATGAACTCCTCACCAGTAA. The 171 paired-end reads from which this *k*-mer originated mapped to the genome as follows: 2 (1.17%) aligned concordantly zero times, 2 (1.17%) aligned concordantly exactly once and 167 (97.66%) aligned concordantly more than once. The assembly of these reads produced two fragments: the first of length 273 bp with coverage of 1.23 and the second of length 924 bp with coverage of 27.41 (SEQ3; Supplementary Table 2). We aligned this second fragment to the genome using Minimap2 with default parameters[54]. Minimap2 reported only one hit to chromosome 3 (NC_024461.1) in position 159,141,222–159,142,137.

Analysis of guaiacol concentration in tomato. For Fig. 5d–f, guaiacol concentration was labeled '$\log_3$_guaiacol' in the original study. From the 293 *k*-mers

passing the threshold, 184 could be mapped uniquely to the genome: 135 to chromosome 0 at position 12,573,795–12,576,534, 45 to chromosome 9 at position 69,301,436–69,305,717, 3 to chromosome 6 at position 8,476,136–8,476,138 and 1 to chromosome 4 at position 53,222,324. The four *k*-mers that mapped to chromosomes 4 and 6 were checked manually by locating their reads and aligning them to the genome. In all cases, no reads could be aligned to the genome (>99.5%). For the 35 *k*-mers not mapping to the genome and in high LD (Fig. 5e), all reads containing at least one of the *k*-mers were retrieved and assembled (SEQ4; Supplementary Table 2). An NCBI BLAST search of this fragment resulted in positions 1–574 mapped to position 12,578,806–12,579,379 on chromosome 0 of the tomato genome (CP023756.1) and positions 580–1,169 mapped to positions 289–878 in *NSGT1* (KC696865.1). The R104 smoky accession *NSGT1* ORF starts at position 307, as reported previously[33]. An NCBI BLAST search of *NSGT1* (KC696865.1) identified mapping to chromosome 9 of the tomato genome (CP023765.1) from 975–1,353 to positions 69,310,153–69,309,775.

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

Data availability

A list of all phenotypes and top SNPs or *k*-mers passing their corresponding thresholds can be found at https://zenodo.org/record/3701176#.XmX9u5NKhhE. The authors declare that all other data supporting the findings of this study are available within the Supplementary Information files.

Code availability

Code is available at https://github.com/voichek/kmersGWAS.

References

38. Seren, Ü. et al. AraPheno: a public database for *Arabidopsis thaliana* phenotypes. *Nucleic Acids Res.* **45**, D1054–D1059 (2017).
39. Fordyce, R. F. et al. Digital imaging combined with genome-wide association mapping links loci to plant–pathogen interaction traits. *Plant Physiol.* **178**, 1406–1422 (2018).
40. Chan, E. K. F., Rowe, H. C., Hansen, B. G. & Kliebenstein, D. J. The complex genetic architecture of the metabolome. *PLoS Genet.* **6**, e1001198 (2010).
41. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
42. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
43. Cheng, C.-Y. et al. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* **89**, 789–804 (2017).
44. Portwood, J. L. 2nd et al. MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Res.* **47**, D1146–D1154 (2019).
45. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating *k*-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
46. Kang, H. M. et al. Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
47. Atwell, S. et al. Genome-wide association study of 107 phenotypes in Arabidopsis *thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
48. Devlin, B. & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322 (1995).
49. Togninalli, M. et al. The AraGWAS Catalog: a curated and standardized *Arabidopsis thaliana* GWAS catalog. *Nucleic Acids Res.* **46**, D1150–D1156 (2018).
50. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
52. Robinson, J. T. et al. Integrative Genomics Viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
53. Ristova, D., Giovannetti, M., Metesch, K. & Busch, W. Natural genetic variation shapes root system responses to phytohormones in *Arabidopsis*. *Plant J.* **96**, 468–481 (2018).
54. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

Acknowledgements

We thank the many colleagues who have shared *A. thaliana* phenotypic information with us. We thank in particular G. Zhu and S. Huang for help with tomato genotypic and phenotypic information and C. Romay, R. Bukowski and E. Buckler for help with maize genotypes and phenotypes. We thank K. Swarts, F. Rabanal and I. Soifer for fruitful discussions. This work was supported by the DFG ERA-CAPS 1001 Genomes Plus and the Max Planck Society.

Author contributions

Y.V. and D.W. designed the study and wrote the paper. Y.V. conducted the analysis.

Competing interests

The authors declare no competing interests.

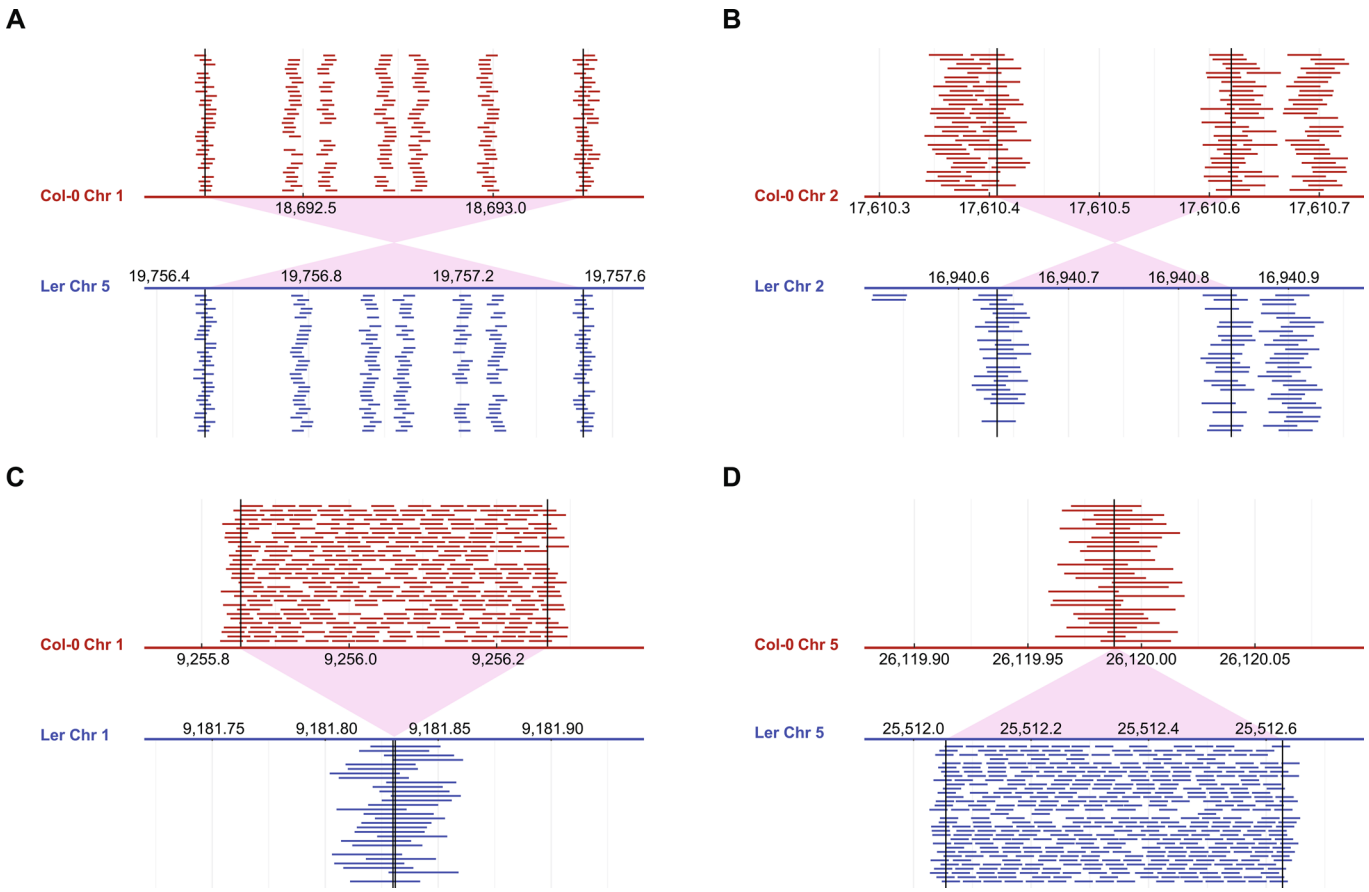
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-020-0612-7.

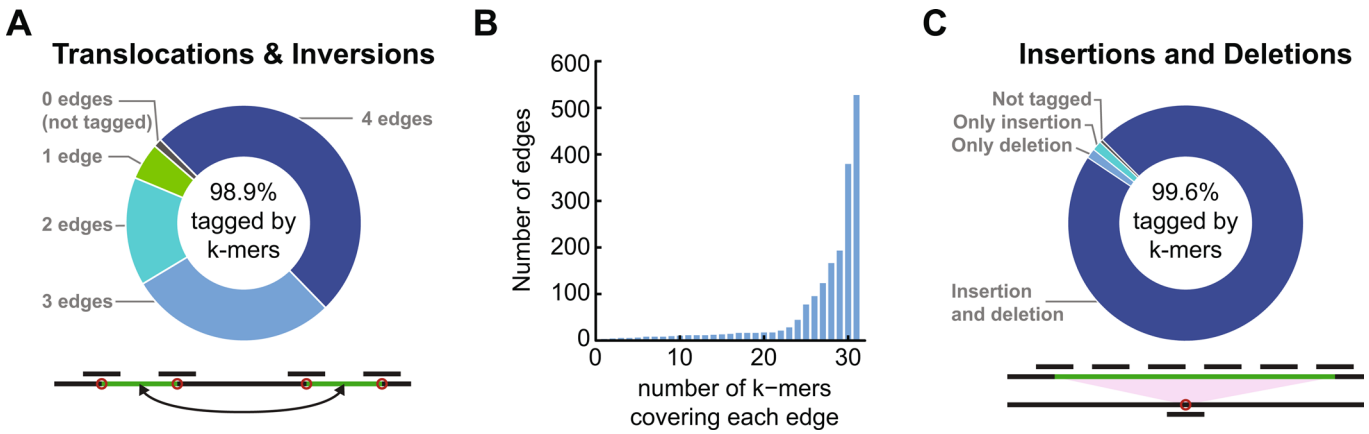
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0612-7.

Correspondence and requests for materials should be addressed to D.W.

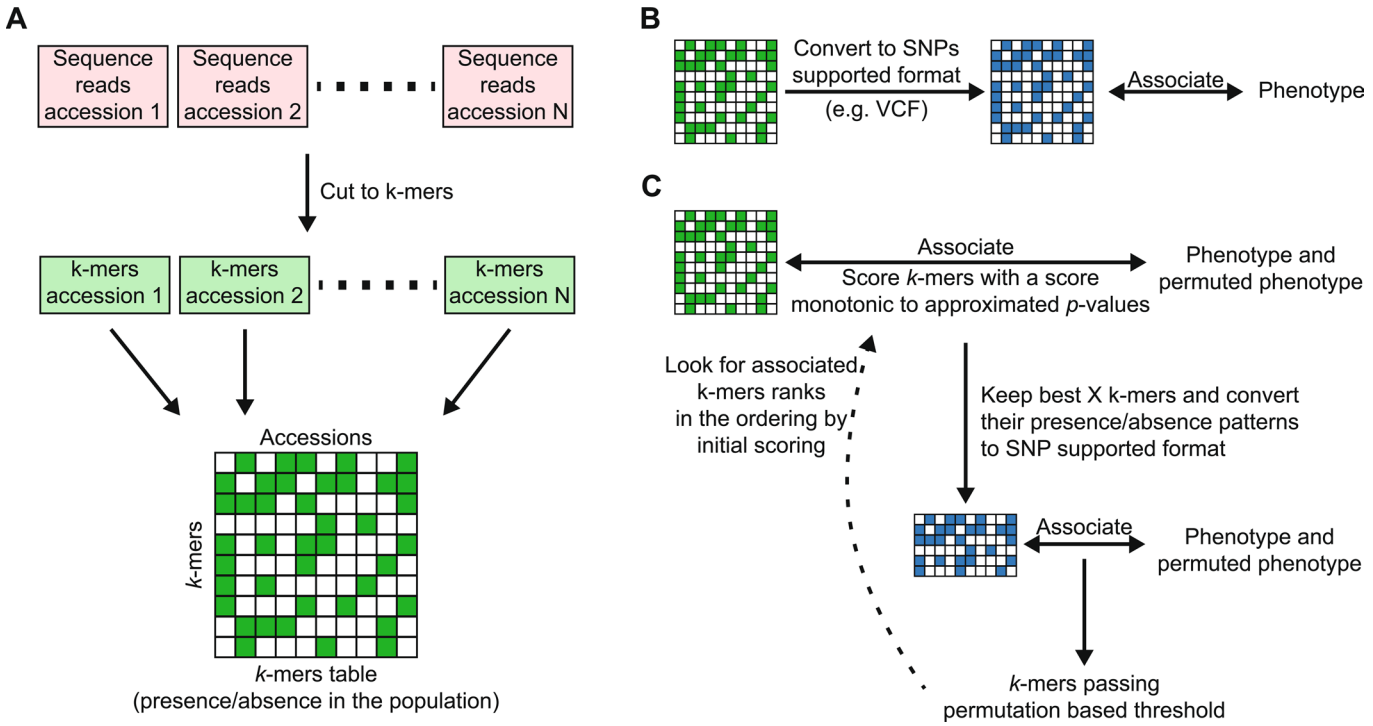
Reprints and permissions information is available at www.nature.com/reprints.



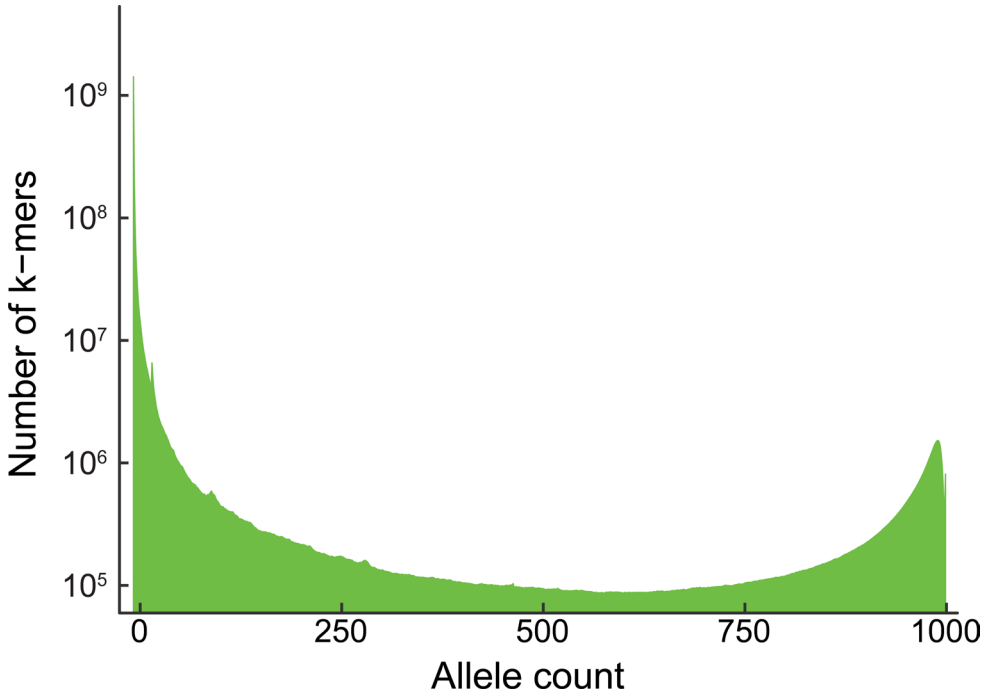
Extended Data Fig. 1 | Examples of well characterized structural variant tagged by *k*-mers. Examples of how *k*-mers tag well characterized structural variants[22] between the Col-0 reference genome and the Ler fully assembled genome. The two genomes were used to count 31 bp *k*-mers, and all *k*-mers unique to one genome and appearing only once in it were plotted in the indicated regions. The **a** translocation, **b** inversion and **c-d** insertion/deletion positions are indicated by vertical lines and red shades. The *k*-mers unique to Col-0/Ler are plotted in the upper/lower panels in red/blue, respectively. The five positions tagged by *k*-mers inside the translocation presented in **a** are either SNPs or 1 bp indels.



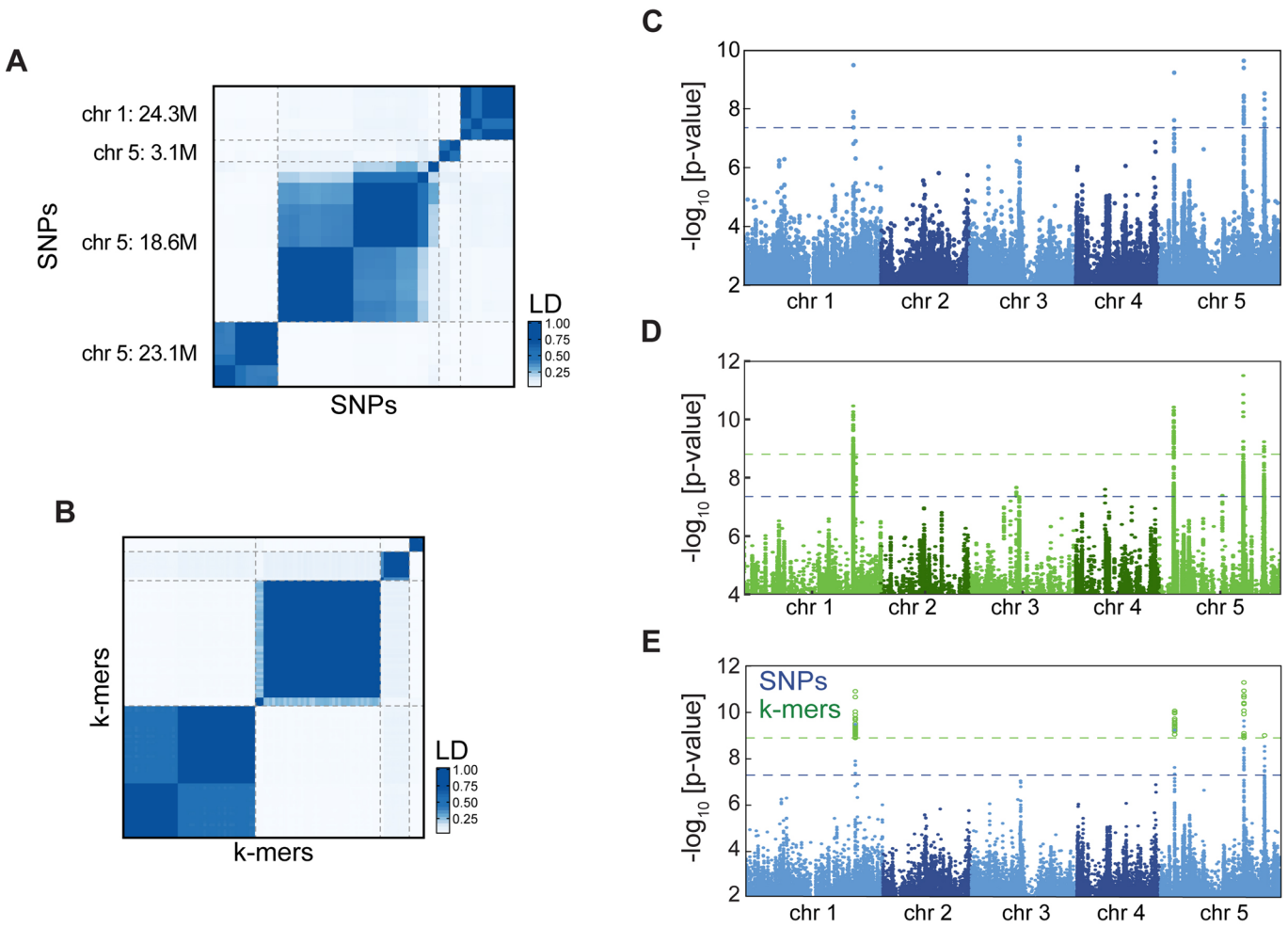
Extended Data Fig. 2 | Genome-wide evaluation of *k*-mer potential to detect SVs in well-characterized genomes. a, For every translocation or inversion, previously identified[22] between the Col-0 reference genome or the Ler genome we evaluate if it is tagged by 31 bp *k*-mers. Each translocation or inversion will affect 4 edges between the translocated fragment and the neighbouring genomic regions (bottom panel). For every previously identified translocation or inversion, the number of edges (0-4) which are tagged by *k*-mers unique to one genome were counted. Only 1.1% of these SVs were not tagged by any *k*-mer unique to one genome (upper panel). **b**, For every edge tagged by *k*-mers, described in A, we plot the number of *k*-mers unique to one genome which tagged it. The histogram is enriched with edges covered by the maximal number of *k*-mers, 31. **c**, Evaluating the potential to tag by *k*-mers long insertions/deletions between the well characterized genomes of Col-0 and Ler[22]. While in the genome with the apparent deletion only the junction between the two fragments will be tagged by unique *k*-mers, in the genome with the apparent insertion, the entire insert will be tagged (bottom panel). Only 0.4% of the previously characterized long insertions/deletions are not tagged by unique *k*-mers.



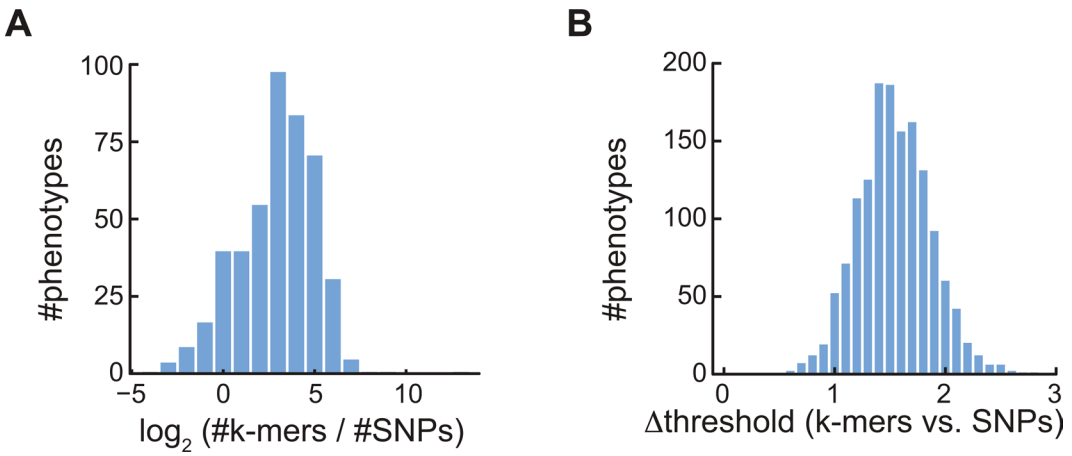
Extended Data Fig. 3 | Pipeline for *k*-mer-based GWAS. a, Creating the *k*-mer presence/absence table: Each accession's genomic DNA sequencing reads are cut into *k*-mers[45], filtering *k*-mers appearing less than twice/thrice in a sequencing library. *k*-mers are further filtered to retain only those present in at least 5 accessions, and ones that are found in both forward and reverse-complement form in at least 20% of accessions they appeared in. All *k*-mer lists are combined into a *k*-mer presence/absence table. **b**, Genome-wide associations on the full *k*-mers table using SNP-based software: the *k*-mers table is converted into PLINK binary format, which is used as input for SNP-based association mapping software[14,42]. **c**, GWA optimized for the *k*-mers: *k*-mers presence/absence patterns are first associated with the phenotype and its permutations using a LMM to account for population structure[16,17]. This first step is done by calculating an approximated score of the exact model. Best *k*-mers from this first step (for example 100,000 *k*-mers) are passed to the second step, In which an exact *p*-value is calculated[14] for both the phenotype and its permutations. A permutation-based threshold is calculated, and all *k*-mers passing this threshold are checked for their rank in the scoring from the first step. If not all *k*-mers hits are in the top 50% of the initial scoring, then the entire process is rerun from the beginning, passing more *k*-mers from the first to the second step. This last test is built to confirm that the approximation of the first step will not remove true associated *k*-mers.



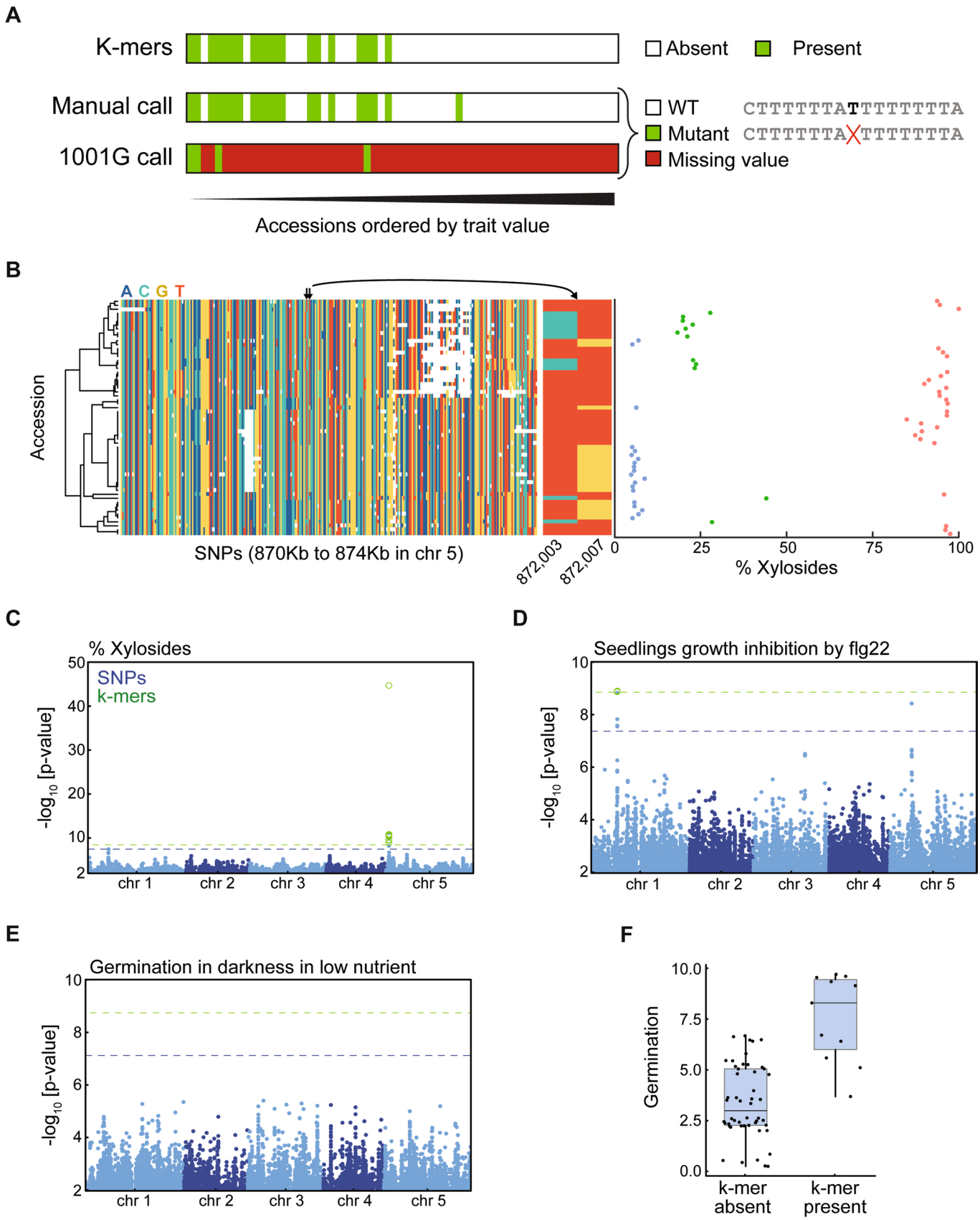
Extended Data Fig. 4 | Allele counts for *A. thaliana* 1001G *k*-mers. Histogram of *k*-mer allele counts: For every N=1..1008, the number of *k*-mers appeared in exactly N accessions is plotted.



Extended Data Fig. 5 | Flowering time-genotype associations in *A. thaliana* identified with *k*-mers. a, LD between SNPs associated with flowering time. Dashed lines represent the four variant types, as in Fig. 1c. **b**, LD between *k*-mers associated with flowering time, Dashed lines represent the four variant types, as in Fig. 1c. **c**, Same as Fig. 1d with only SNPs. **d**, Same as Fig. 1d with only *k*-mers presented, showing also *k*-mers lower than the threshold. **e**, Manhattan plot of SNPs and *k*-mer associations with flowering time in 10 °C as in Fig. 1d for *k*-mers of length 25 bp.

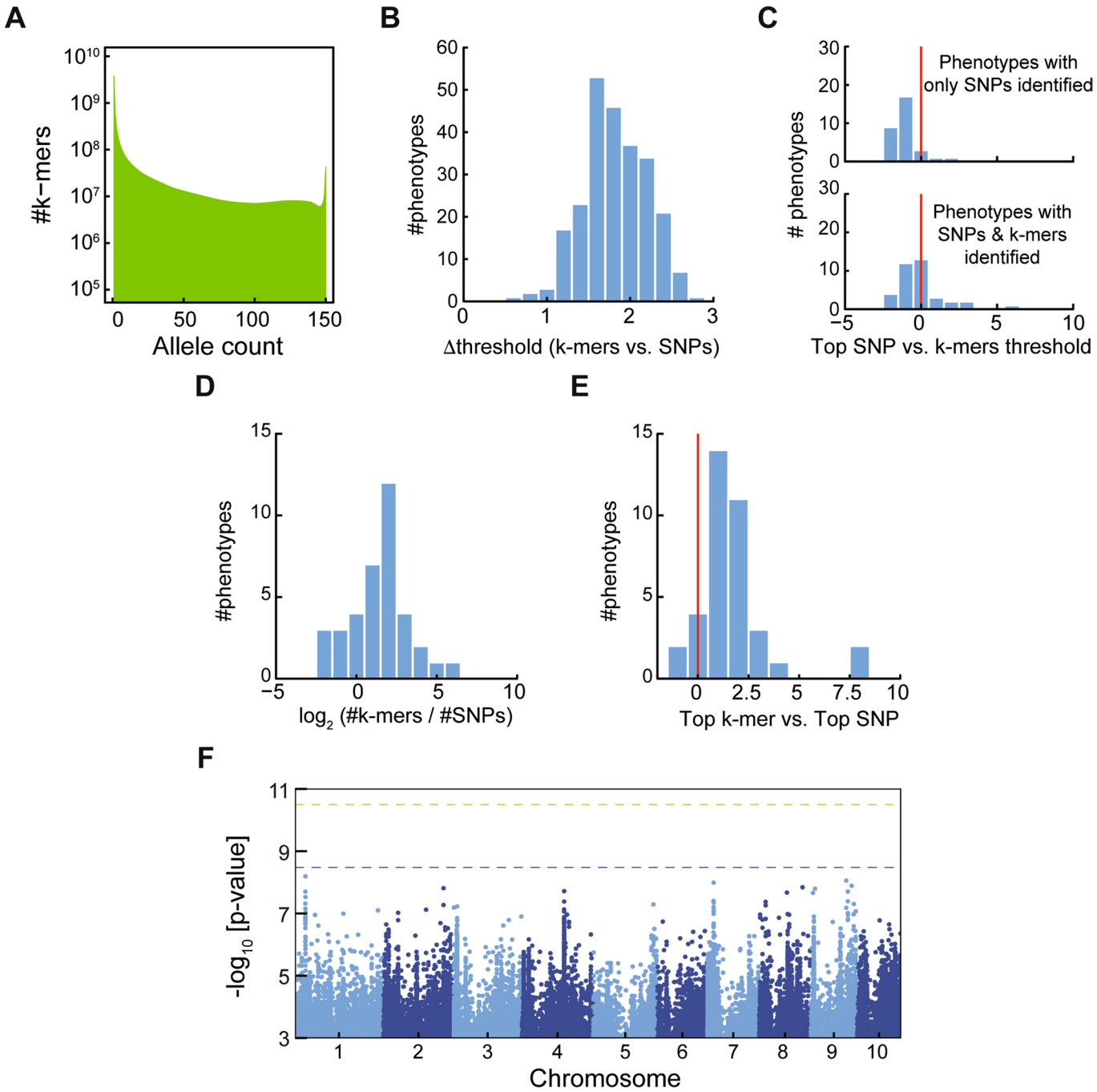


Extended Data Fig. 6 | Comparison of SNP- and *k*-mer-GWAS on phenotypes from 104 studies on *A. thaliana* accessions. a, Histogram of the number of identified *k*-mers vs. identified SNPs (in $\log_2$) for *A. thaliana* phenotypes. Only the 458 phenotypes with both variant types identified were used. **b**, Histogram of thresholds difference of *k*-mers vs. SNPs of all *A. thaliana* phenotypes. Thresholds were $-\log_{10}$ transformed.

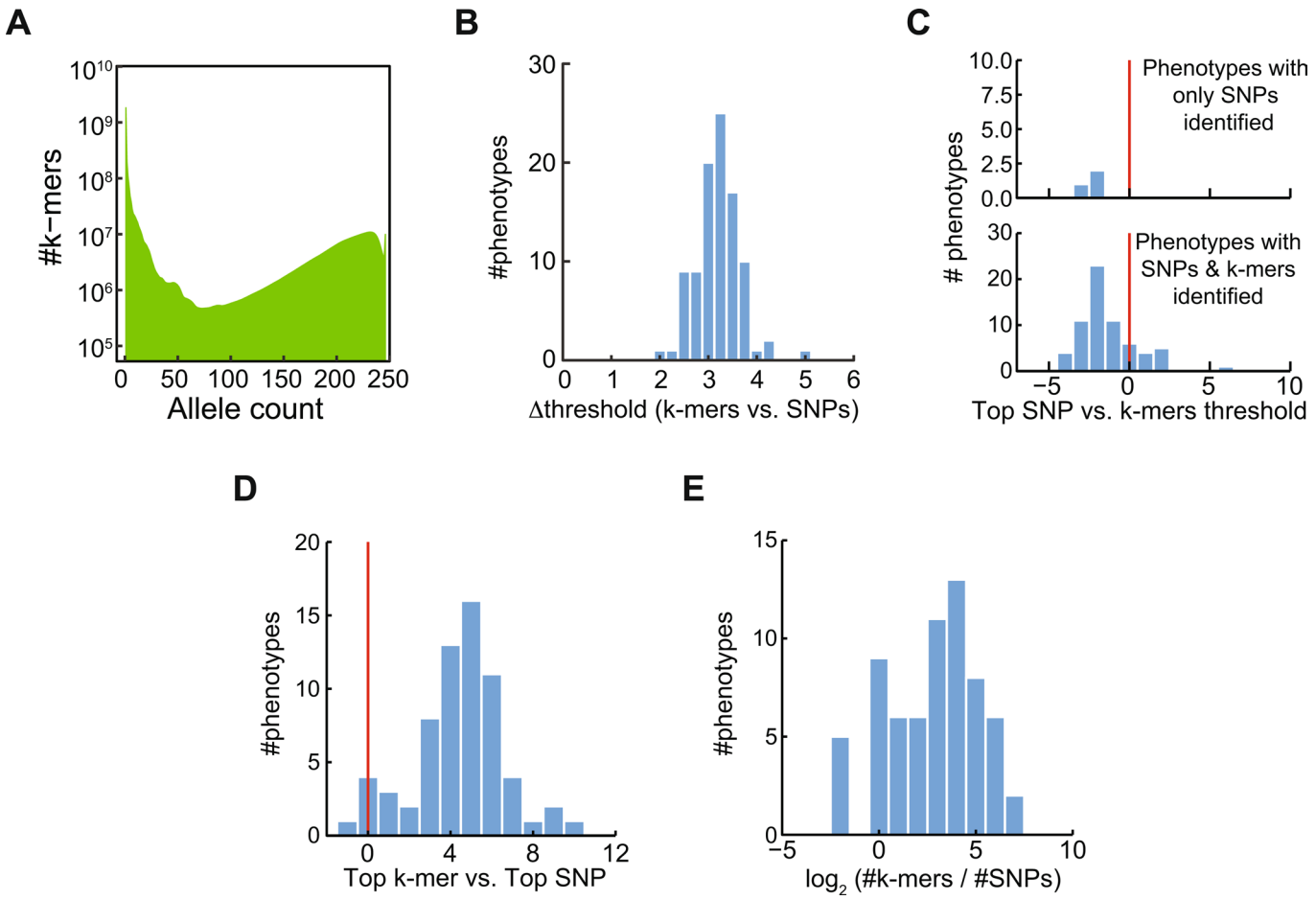


Extended Data Fig. 7 | See next page for caption.

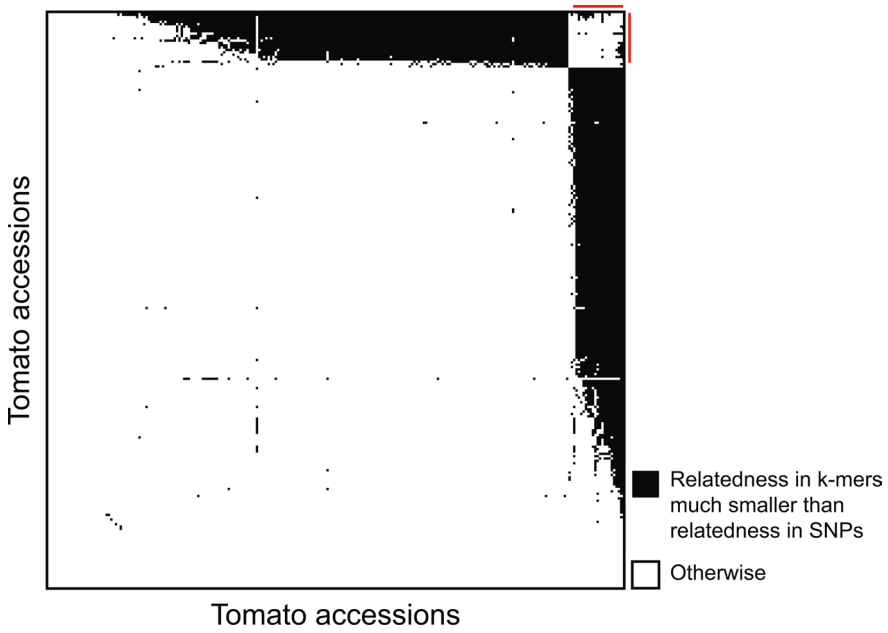
Extended Data Fig. 7 | Specific case studies in which *k*-mers are superior to SNPs. a, Results from GWAS on measurements of lesions by *Botrytis cinerea* UKRazz strain[39]. An example of *k*-mers having better hold on a short variant: 19 *k*-mers and no SNPs were identified, all *k*-mers in complete LD (top row). Sequence reads containing the *k*-mers mapped to chromosome 3, with a single T nucleotide deletion out of an eight T's stretch, in position 72,017. Manual (middle) and the 1001G project (bottom) calls are shown. In the 1001G, 57 of 61 accessions contain missing values. **b**, Haplotypes around SNPs associated with xylosides concentrations are not correlated with this trait. All SNPs in positions 870,000 to 874,000 in chromosome 5 were hierarchically clustered (left panel, white mark missing values). The two identified SNPs are marked by arrows and a close-up of their state is shown (middle panel). Phenotypic values colored according to the two SNPs: TG blue, TT red, and CT green (right panel). **c-e**, Manhattan plot for: **c**, xyloside percentage, **d**, seedling growth inhibition by a flg22 variant, **e**, germination in darkness in low nutrient conditions. **f**, Germination phenotype plotted for accessions with top associated *k*-mer present or absent. Boxes cover 25%- 75% percentiles, medians marked by horizontal lines, and whiskers cover the full range of values.



Extended Data Fig. 8 | Comparison of SNP- and *k*-mer- based GWAS in maize. a, Histogram of *k*-mer allele counts for maize accessions. **b**, Histogram of difference between threshold values of SNPs and *k*-mers for maize phenotypes. **c**, Histogram of the top SNP *P*-value divided by the *k*-mers defined threshold, in (-log10), for maize phenotypes. Plotted for phenotypes with only identified SNPs (upper panel) or for phenotypes with both SNPs and *k*-mers identified (lower panel). **d**, Histogram of the number of identified *k*-mers vs. identified SNPs for maize phenotypes. **e**, Histogram of the difference between top (-log10) *p*-values in the two methods for maize phenotypes identified by both methods. Plotted as in Fig. 2g. **f**, Manhattan plot of associations with ear weight (environment 07A). Associated *k*-mers could not be located in the reference genome, and are thus not presented.



Extended Data Fig. 9 | Comparison of SNP- and *k*-mer-based GWAS in tomato. a, Histogram of *k*-mers allele counts for tomato accessions. **b**, Histogram of difference between threshold values of SNPs and *k*-mers for tomato phenotypes. **c**, Histogram of the top SNP *P*-value divided by the *k*-mers defined threshold, in -log10, for tomato phenotypes. Plotted for phenotypes with only identified SNPs (upper panel) or for phenotypes with both SNPs and *k*-mers identified (lower panel). **d**, Histogram of the difference between top (-log10) *p*-values in the two methods for tomato phenotypes. **e**, Histogram of the number of identified *k*-mers vs. identified SNPs for tomato phenotypes.



Extended Data Fig. 10 | Kinship matrix calculation based on *k*-mers for tomato accessions. Identification of pairs of tomato accessions for which relatedness as measured with *k*-mers is much lower than relatedness as measured with SNPs. For every pair among the 246 accessions, a black square is plotted if the difference in relatedness between SNPs and *k*-mers is larger than 0.15. Accessions are ordered by the number of black square in their row/column. Red lines mark the 21 accessions with most black squares, that is, those for which the *k*-mer/SNP difference in relatedness is larger than 0.15 for the most pairs.

nature research

Corresponding author(s): Detlef Weigel

Last updated by author(s): Mar 9, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No new data was generated in this manuscript.
Data analysis	The code utilized was uploaded to GitHub: https://github.com/voichek/kmersGWAS. We also used the following software: vcftools v0.1.15, PLINK v1.9, KMC v3, GEMMA v0.96, and mvnpermute R package.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

No new data was generated in this study. The collection of previous published data sets can be found in https://zenodo.org/record/3701176#.XmX9u5NKhhE

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No new data was generated in this study. Previously published data sets were used as is.
Data exclusions	No data were excluded.
Replication	Not applicable
Randomization	Not applicable
Blinding	Not applicable

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging