European ash pangenome reveals widespread structural variation and a diverse genetic basis for low ash dieback susceptibility

Authors: Daniel P. Wood[1], Mohammad Vatanparast[1], Dario Galanti[1], Katherine Wheeler[1,2], Emma Curran[3], Levi Yant[3,4], Richard Whittet[5], Richard A. Nichols[6], Richard J. A. Buggs[1,6], Laura J. Kelly[1]

1. Royal Botanic Gardens, Kew, Richmond, UK
2. Centre for Novel Agricultural Products, University of York, York, UK
3. School of Life Sciences, University of Nottingham, Nottingham, UK
4. Linnean Centre for Plant Biology, Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Sciences, Uppsala, Sweden
5. Forest Research, Northern Research Station, Roslin, Midlothian, UK
6. School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK

Email: d.wood@kew.org; l.kelly@kew.org

Abstract

European Ash (*Fraxinus excelsior*) is a keystone forest tree species, whose populations are being decimated by Ash Dieback disease (ADB). Uncovering the genetic basis of low susceptibility to this devastating disease relies on a comprehensive understanding of genetic variation present in *F. excelsior.* A linear reference genome from a single individual cannot contain the total sequence variability within a species, including its genic regions; a pangenome more fully captures total sequence content. In this study, we developed a *F. excelsior* pangenome reference using a new chromosomal-level

phased linear reference genome, and *de novo* assemblies and long-read data from a geographically diverse set of fifty *F. excelsior* samples. We identified 362,965 structural variants (SVs), including 174Mb of sequence absent from the linear reference genome (22% of the linear reference size). We demonstrated that failing to explicitly link SVs with gene sequences can lead to substantial overestimation of dispensable genes (those that vary in their presence/absence between individuals) due to variability in the annotation process. Controlling for this, we identified 3,065 dispensable genes including 133 with gene ontologies associated with low disease susceptibility. We used the pangenome to analyse existing genomic data from over 1,200 individuals to identify loci associated with reduced susceptibility to ADB. This analysis revealed that different genetic backgrounds harbour distinct sets of loci associated with low susceptibility, highlighting the complexity of this trait but also the potential benefits of combining variants from different genetic backgrounds to boost efforts to ensure the future survival of *F. excelsior*.

Introduction

High quality reference genomes constructed from single individuals have been used to great effect to assess genetic diversity within species, allowing mapping of short reads from many individuals and discovery of genetic variants[1]. However, this approach provides limited insights into larger structural variants (SVs) between individuals and can lead to some miscalling of single nucleotide polymorphisms (SNPs)[2]. Pangenomes – reference genome data structures incorporating highly contiguous sequences from a representative set of individuals – have revealed the total sequence contained in SVs

can exceed the average genome size of a species[3–6]. These SVs can encompass genes, with some studies identifying the majority of genes within a species as dispensable (i.e. not present in all individuals[5,7–9]). Pangenomes thus often contain genes that would be entirely missed if using a single reference genome approach[10,11]. Indeed, novel disease resistance genes are frequently identified in pangenomes[12–14] including in trees[15–17], suggesting pangenome approaches may be particularly valuable when seeking to detect loci contributing to pathogen resistance. Analysis of existing data using pangenome references has uncovered the basis of missing heritability in some systems[18] and has enhanced our understanding of the genetic basis of a variety of traits, often adding complementary information to that uncovered by analyses using a linear reference genome[5,19–21]. Thus, to better understand genetic variability within a species and its impact on traits, pangenome references incorporating diverse samples are required[10,22].

Fraxinus excelsior (European ash), a keystone tree species within Europe, is under severe threat from ash dieback disease (ADB) caused by the invasive non-native fungus *Hymenoscyphus fraxineus*[23,24]. Populations of ash are undergoing substantial and rapid mortality[25–27], with high associated management costs and biodiversity loss[28–30], attracting substantial public concern[31–34]. A small fraction of trees remain relatively healthy under ADB inoculum pressure and quantitative resistance to the disease has a narrow-sense heritability of approximately 0.4[35–39]. Understanding the genetic basis of this trait is important for ascertaining the feasibility of producing new individuals with lower susceptibility[40–44]. Stocks et al.[42] identified SNPs associated with resistance in

mainly UK populations by mapping short reads to a single reference genome; this identified 192 SNPs significantly associated with low ADB susceptibility[38,40,45–47]. Studies in other European populations did not find SNPs strongly associated with low ADB susceptibility[43,44], with one indicating many SNPs from Stocks et al.[42] were at very low frequencies in mainland European populations[48], raising the possibility of diverse mechanisms of resistance between European populations[42–44,47–49]. A pangenome reference for the species that captures variation from different genetic backgrounds would represent an invaluable resource to enable improved understanding of this critically important trait. A recently published "superpangenome" of the *Fraxinus* genus identifies interspecific SVs associated with salt tolerance, suggesting a role for these variants in adaptation to abiotic stressors over millions of years[50]. However, this does not capture intraspecific SVs in *F. excelsior* that may be involved in low ADB susceptibility.

Here we construct a chromosome level, phased genome assembly for the same individual as the first reference assembly for European ash[51] and a graph pangenome using long-read sequencing from a diverse sample of 50 individuals across Europe. This revealed 174Mb of additional sequence captured in structural variants, an increase of 22% from the new linear reference genome size from a single individual. By explicitly associating putative dispensable genes with SVs, we identified 3,065 high confidence dispensable genes in *F. excelsior*, including 133 with gene ontology terms associated with disease resistance. Reanalysis of poolseq data from over 1,200 individuals[42] revealed that the majority of alleles associated with low ADB susceptibility are not

widespread across most of the seed sources within the study, highlighting the potential for improved ADB resistance via crossing between populations within the UK.

Results

A high quality linear reference genome for European ash

A phased linear reference genome was produced for a single *Fraxinus excelsior* individual, with total sizes of 793 and 787 Mb, N50s of 32.9 and 32.2 Mb, and L90s of 22 for the haplotype 1 (henceforth referred to as “BATG-1.0”) and haplotype 2 assemblies respectively (Supplementary Table 1). *Fraxinus excelsior* has 23 pairs of chromosomes[52] – the top 23 scaffolds of BATG-1.0 and haplotype 2 contain 93.7% and 94.5% of the total sequence in each assembly, with the 24th scaffold for each of the haplotypes being 2.1% and 1.5% of the size of the 23rd scaffold. Telomeres were detected at both ends of 19 of the top 23 scaffolds in BATG-1.0 and 17 of the top 23 in haplotype 2**.** Together these results indicate the assemblies are pseudochromosomal. Unlike other recently published pseudochromosmal assemblies of *F. excelsior*[50,53], this assembly is not scaffolded using the *F. pennsylvanica* genome assembly[53,54]. We also assembled and annotated the *F. excelsior* chloroplast genome; this was 155.6kb in length with 130 annotated genes.

98.2% and 98.3% of eudicot BUSCO genes were completely assembled in the BATG-1.0 and haplotype 2 genomes; 21% of BUSCOs were duplicated, a similar proportion as in other *F. excelsior* assemblies (Supplementary Table 1) and consistent with the fact that there are several whole genome duplication events in the Oleaceae, the family to

114 which *F. excelsior* belongs[50,55]. 98.3% of PacBio reads mapped to one or both of the
115 haploid assemblies. The total assembly sizes of 787 and 793Mb are smaller than the
116 1C genome size of 877Mb estimated by flow cytometry in Sollars et al. (2017) but are
117 similar to the size of unpublished pseudochromosomal assemblies of *F. excelsior*
118 (FRAX_001_PL; daFraExce3.hap1.1) that do not use *F. pennsylvanica* for scaffolding (see
119 Supplementary Table 1).

120 65.6% of BATG-1.0 consists of repeats, consisting mostly of LTR elements (33.0%) and
121 unclassified repeats (19.9%). This is higher than previously estimated for the same
122 individual[51], but is based on a more complete assembly that likely captures more
123 repetitive elements, and is similar to the repeat content of the assembly of another *F.*
124 *excelsior* individual (69.12%) and other *Fraxinus* assemblies[50]. This repeat content is
125 however substantially higher than *F. pennsylvanica* (48.4%[54]). Olive, also a member of
126 the Oleaceae, has a lower repeat content than *F. excelsior* (56.2%[56]), although tomato
127 (a fellow asterid with a similar 1C genome size) is more similar (65%-72%)[57,58].

128 Annotation of BATG-1.0 was performed using BRAKER3. Ten runs of BRAKER3 were
129 performed with short-read RNA-seq data and protein ortholog information, and 10 runs
130 with long-read RNA-seq data and protein orthologs. These identified 32,207 genes
131 annotated in every run, with a further 1,519 genes annotated in every run of the short-
132 read annotation only and 1,392 genes annotated in every run of the long-read
133 annotation only. A further 3,158 genes were inconsistently annotated, in either the
134 short- or long-read runs. To reduce the chances of erroneously identifying genes as
135 dispensable in the pangenome, we initially retained all of these genes, giving a total of
136 38,276. This is a similar number to that reported in the BATG-0.5 assembly (38,949[51])

137 and the *F. pennsylvanica* assembly (35,470[54]) but slightly lower than the 43,392 genes
138 reported in Franco Ortega et al.[53] and substantially lower than the *Fraxinus* assemblies
139 used in the *Fraxinus* superpangenome (48,389-53,538)[50] including their *F. excelsior*
140 assembly (50,327). Using OrthoFinder to identify orthologs from BATG-1.0, BATG-0.5,
141 *F. pennsylvanica*, *Olea europaea*, *Solanum lycopersicum*, *Erythranthe guttata*,
142 *Sesamum indicum* and *Salvia miltiorrhiza* (Supplementary Table 2), 5.5% of BATG-1.0
143 genes lacked an ortholog in another assembly, compared with 8.5% in BATG-0.5.
144 94.4% of eudicot BUSCO proteins were found in the total BATG-1.0 protein set,
145 suggesting the annotation captures the vast majority of genes in the assembly.

146 Structural variant calling and pangenome construction

147 Low susceptibility to ADB is likely a highly polygenic trait[42,49], so we prioritised sampling
148 a larger number of individuals to construct the pangenome instead of focusing on
149 constructing fewer highly contiguous assemblies, even if this restricts the types of SV
150 that can be analysed. Oxford Nanopore Technologies (ONT) data was generated for 49
151 individuals from a variety of European provenances as well as the reference individual
152 (Figure 1A, Supplementary Table 3). PCA of SNPs called from this data revealed little
153 geographic structure (Figure 1B), as was also found to be the case when a similar set of
154 provenances were analysed in earlier work[51].

155 Heterozygous structural variants called between the two high quality haploid genomes
156 of the reference individual were used to assess the accuracy of calling SVs using lower
157 coverage ONT data from the same individual with a variety of methods. These included
158 mapping the ONT reads to the reference haploid genome and calling SVs using

159 Sniffles2[59] and cuteSV[60], de novo assembling the ONT reads with Flye[61], Shasta[62] or
160 nextDenovo[63] and calling SVs from contigs mapped against the reference, and the
161 combination of read mapping and assembly mapping approaches (see Methods). This
162 analysis indicated that the read mapping approaches had higher recall of SVs detected
163 using the high quality haploid genomes, while assembly mapping approaches had
164 higher precision, and a combination of read mapping and mapping of the assembly
165 produced by Shasta gave the highest recall (84.4%) and F1 score (77.9%), with a
166 precision of 72.2% (Extended Data Figure 1). This third approach was therefore used to
167 call SVs for all 50 samples.

168 Merging SVs across all individuals identified a total of 737,995 SVs (Figure 2A).
169 Merging samples sequentially resulted in fewer new SVs discovered in the fiftieth
170 sample compared with the first, but the number newly detected did not plateau,
171 indicating that more samples would be required to exhaustively sample SVs in *F.*
172 *excelsior* (Figure 2B). SVs present in three or more individuals were retained for further
173 analysis (Figure 2C). This resulted in 362,965 SVs, of which 170,202 were insertions,
174 189,724 were deletions, 295 were inversions and 2,744 were tandem duplications
175 (Figure 2C). The total number of SVs represented 396Mb of variable sequence across
176 the *F. excelsior* individuals, including 174Mb of insertion sequences not present in the
177 linear reference genome (Figure 2B). 99.9% of SVs were on the 23
178 pseudochromosomal scaffolds and 69% overlapped with an annotated repeat in the
179 linear reference genome.

Pangenome annotation: explicitly linking dispensable genes with SVs

To identify dispensable genes with an explicit link to SVs in the pangenome, a genome sequence was generated for each individual by using the SV genotype calls to modify the BATG-1.0 reference fasta (i.e. deleting deletions, inserting insertion sequence, etc). Genes were *de novo* annotated for each individual genome sequence and consolidated across individuals using OrthoFinder to identify a set of unique gene models. Of the 47,449 genes, 31,117 (65.6%) were present in all individuals including the reference (Figure 3). Of the 9,449 genes absent from the BATG-1.0 genome (i.e. putatively undiscovered dispensable genes), 5,615 did not overlap with an SV in any individual. For those that did overlap with an SV, only 1,240 did so with an F1 score (a measure of how consistently a gene presence or absence was associated with an SV, see Methods) of 0.8 or above (Figure 3, Extended Data Figure 2). As in each case the non-SV sequence is identical to the reference sequence, these putatively dispensable genes may be the result of stochasticity in the annotation process, or the result of SVs elsewhere in the genome altering the mapping of RNA-seq data from the reference individual and thereby reducing evidence to support the gene annotation. Similarly, of the 6,883 genes present in the reference assembly but absent from one or more of the other assemblies, 3,159 did not overlap with an SV in any individual. For those that did overlap with an SV, only 1,825 did so with an F1 score of 0.8 or above (Figure 3, Extended Data Figure 2). Putatively dispensable genes with an F1 score of less than 0.8 were classified as invariant if present in the reference genome, or were otherwise

202 removed from the analysis. This brought the total number of genes in the pangenome to
203 39,420, of which 3,065 were dispensable (7.8% of the total). This is a substantial
204 reduction in dispensable genes compared with the estimated 34.4% of genes without
205 filtering for those consistently overlapping with SVs (Figure 3). Putatively dispensable
206 genes without a consistent SV association (i.e. those excluded from our final
207 annotation) could be affected by the limited RNA-seq data (available for the reference
208 individual and its mother tree only) or the particular annotation pipeline used. The final
209 pangenome gene set included 99.2% complete eudicot BUSCOs – more than the
210 98.2% captured in the BATG-1.0 sequence alone.

211 Functional annotation indicated that high confidence dispensable genes overall had
212 fewer functional annotations than invariant genes (Figure 4A). Putatively dispensable
213 genes that were filtered out also had a lower number of functional annotations (Figure
214 4A). This may be a result of these gene subsets being less conserved across species
215 and/or less well annotated. High confidence variable genes were enriched for ten GO
216 terms including “RNA-templated DNA biosynthetic process” and “transposition, RNA-
217 mediated”, as well as “floral meristem growth” (Supplementary Table 4). Putatively
218 dispensable genes that were filtered out were enriched for 28 terms including “defense
219 response” and “regulation of auxin polar transport” that were not significant in the
220 dispensable genes (Supplementary Table 4). A direct comparison of the high
221 confidence dispensable genes with the filtered genes provided little evidence of GO
222 terms only enriched in the dispensable genes, or the filtered genes, differing
223 meaningfully in frequency between these two sets (Figure 4B).The 3,065 dispensable
224 genes include those with GO terms indicating a role in response to stress (324),

225 developmental process (310) and defense response (133), with 29 genes linked to a
226 function in “defense response to fungus” specifically (Supplementary Table 4). Genes
227 annotated for functions in leaf development (43) and leaf senescence (16) were also
228 included within the dispensable set – variation in Spring leaf flushing and Autumn leaf
229 senescence has been linked to low ADB susceptibility in some studies[38,45,64].

230 Improved short read mapping, SNP calling and SV calling using 231 the pangenome

232 We next tested whether using our pangenome reference as opposed to a linear
233 reference sequence improves read mapping for short-read sequenced individuals.
234 While overall mapping to our pangenome produced little effect on the total number of
235 short reads mapped in proper pairs (Extended Data Figure 3), we observed general
236 improvements based on other metrics: this includes increases in mean mapping quality
237 and alignment score (1.4% and 1.2% on average, respectively Extended Data Figure 3),
238 a reduction in the total number of substitutions (3.0% on average, Extended Data Figure
239 3), and a large reduction in soft clipped sequence (33.5% on average, Extended Data
240 Figure 3). Overall, this indicates the pangenome improves the accuracy of short-read
241 read mapping.

242 Examination of allelic depths of homozygous alternate, reference and heterozygous
243 sites showed SVs and SNPs exhibit similar distributions of reference and alternate
244 alleles (Extended Data Figure 4). MAF distributions and deviation from HWE for the two
245 variant types were also similar (Extended Data Figure 5A, B). Patterns of linkage decay
246 with distance were broadly similar in SNP-SNP, SV-SV and SNP-SV comparisons, with

247 slightly faster linkage decay between SVs, and between SVs and SNPs, compared with
248 between SNPs (Extended Data Figure 5C). Together, these suggest SNP and SV calls
249 behave in a similar way for individually sequenced samples.

250 To investigate the accuracy of allele frequency estimations from short-read poolseq
251 data, we constructed an *in silico* pool with an equivalent number of individuals and
252 depth to the average *in vitro* pool from Stocks et al.[42]. Allele frequency estimates for
253 SNPs were estimated using popoolation2[65]. Similar software does not exist for
254 estimating SV allele frequencies from pangenomes. Whilst vg[66] does not explicitly
255 support poolseq allele frequency estimation, we called SVs for the pools using vg and
256 used the depths of reference vs. alternate alleles to estimate SV frequency in the pools.
257 Correlation between the alternate allele frequencies from the individual based calling
258 and estimates from the *in silico* pool was high for SNPs (slope = 0.94, R^2 = 0.93). For
259 SVs, whilst the overall correlation was high (slope = 0.96, R^2 = 0.85), there were a
260 minority of SVs where the pool based estimates differed substantially from the individual
261 based calling (Extended Data Figure 6A, B). Poolseq data from Stocks et al.[42] was
262 mapped to the pangenome, and SNP and SV frequencies analysed. PCA recovered
263 similar population structures from SV and SNP frequencies estimated using these pools
264 (Extended Data Figure 6C, D), as has been found in other studies[66,67]. Together, these
265 results suggest that allele frequencies of most SVs can be captured accurately by
266 mapping poolseq data to a pangenome.

267 # GWAS with SVs and SNPs

268 We reanalysed poolseq data from Stocks et al. [42] using the pangenome. In that study,
269 for 13 seed sources, a pool of healthy trees and a pool of unhealthy trees (42 trees per
270 pool on average) from field trials heavily affected by ADB were sequenced (see
271 Methods for details). We performed a GWAS for low ADB susceptibility using SV allele
272 frequencies estimated from this data, with a Cochran-Mantel Haenszel (CMH) test[68,69],
273 a commonly used statistic for identifying associations from poolseq data[42,65,70–74]. This
274 identified 531 significant SVs (p-value cutoff of 1×10^{-13}; Extended Data Figure 7A).
275 However, many of these sites showed inconsistent allele frequency estimates from the
276 *in silico* pools vs. individual genotyping as assessed in the above section (261 had a
277 discrepancy in frequency estimation > 0.2; see Extended Data Figure 7B vs. Extended
278 Data Figure 6B). Whilst some of the 531 significant SVs may be truly correlated with low
279 ADB susceptibility, many are likely false positives resulting from errors in the estimation
280 of allele frequencies. These results indicate SV allele frequency estimates from poolseq
281 data are insufficiently accurate for a reliable GWAS.

282 Our estimated allele frequencies for SNPs called from the Stocks et al.[42] poolseq data
283 were used to perform a GWAS for low susceptibility to ADB using the CMH test, and to
284 identify regions with the top 1% of F_{st} values per chromosome between multiple pairs of
285 healthy and unhealthy pools (Figure 5A). The CMH test identified 92 SNPs passing a
286 significance threshold of 1×10^{-13}. These had large allele frequency shifts in some pools
287 (Supplementary Figure 1) but generally did not shift consistently between healthy and
288 unhealthy pools for the majority of pairs. Only 15.2% of significant SNPs showed allele

289 frequency shifts in the same direction in more than 80% of pairs (Figure 5B). This was
290 partially due to many of these SNPs only being present in subsets of pools - only 28.2%
291 of significant SNPs were segregating in more than 80% of pairs (see also
292 Supplementary Note 1, Supplementary Figure 1). A quasibinomial generalized linear
293 moldel (qGLM) test, able to identify consistent allele frequency shifts across
294 heterogenous pairs of pools[70], did not identify any significant sites (Supplementary Note
295 1, Supplementary Figure 2). Similarly, the maximum number of pairs of pools sharing an
296 F_{st} outlier window was 8, from a total of 15 pairs of pools. For F_{st} outlier windows shared
297 between 5 and 8 pairs of pools, diverse combinations of provenances were
298 represented. For example, the 80 outlier windows shared by 6 pairs of pools were made
299 up of 34 different combinations of provenances, with the most frequently occurring
300 combination found 12 times (Supplementary Table 5**)**. Together, these analyses
301 suggest that the genetic basis of low ADB susceptibility is highly variable across the
302 sampled provenances.

303 A rare insertion is strongly associated with low ADB susceptibility 304 in some provenances

305 In the SNP GWAS, a single 1,767 bp region on chromosome 15 of BATG-1.0 was
306 identified in both the CMH and F_{st} outlier analyses as associated with ash dieback
307 damage. This region contained 55 of the 92 significant SNPs from the CMH test, with up
308 to 7 of the 15 pairs of pools showing F_{st} outliers in this window (Figure 5C). 1.3kb of this
309 region was annotated as repetitive, including an 888bp fragment annotated as
310 LTR/Gypsy (Figure 5C). No gene was annotated in this region or within 5kb of it. Five

311 pairs of pools showed clear increases in depth across this region for one of the pools
312 (e.g. Figure 5D**;** see Extended Data Figure 8 for all pairs), consistent with the presence
313 of a high frequency insertion sequence closely resembling this region but not
314 necessarily in this location. Searching this region against the sequences of insertions
315 detected by the long-read sequencing – including those present in fewer than 3
316 individuals, and therefore not included in the pangenome – did not identify an insertion
317 sequence corresponding to the region of increased depth. This indicates that this SV
318 was not captured during construction of the pangenome. Of the five pairs of pools
319 showing an increased depth in this region, four (FTT-SO, NSZ-302, NSZ-303 and NSZ-
320 405) showed an increase in the healthy pool, whereas one (NSZ-107) showed an
321 increase in the unhealthy pool (Extended Data Figure 8). It should be noted that NSZ-
322 107 had two biological replicates, each representing healthy or unhealthy individuals
323 from two different sets of trial sites[42]; only one of the two replicates showed a difference
324 in depth in this region between the two pools.

325 Discussion

326 Here we present a new high quality linear reference genome for the same individual as
327 used for the first ash reference genome[51] (BATG-0.5), and a pangenome for European
328 ash from a geographically diverse sample of individuals, including some with low
329 susceptibility to ash dieback (ADB). The new linear reference assembly is 793Mb long,
330 with 94% in 23 scaffolds. This represents a substantial improvement compared to
331 BATG-0.5, with an L90 of 22 in BATG-1.0 compared to 12,210 in BATG-0.5. The
332 pangenome of 50 individuals captures 362,965 SVs, representing 396Mb of variable

333 sequence, including 174Mb that is absent from the linear reference genome (equivalent
334 to approximately 22% of its length). This is broadly similar to other recent pangenome
335 studies, although there is substantial variability across these[3,15,75–88] (Extended Data
336 Figure 9A, 9B; Supplementary Table 6). Our analyses likely underestimate the total SV
337 diversity in *F. excelsior*, as each additional individual contributed novel SVs to the
338 collection. A comparison of SVs called between the chromosome-level haploid
339 assemblies, and ONT data for the same individual, indicates our SV calling approach
340 has a precision of 72.2% (although a recall of 84.4%) compared with using
341 chromosome-level phased assemblies, suggesting some SVs included in the
342 pangenome may be false positives. However, it should be noted that i) as heterozygous
343 sites only have half the read support of homozygous alternate sites – it is likely the latter
344 can be detected with higher recall and precision, and ii) the fact that we only included
345 SVs called in at least three samples in the pangenome potentially reduces the number
346 of false positives. Our results highlight the substantial diversity of SVs in this species,
347 consistent with high SNP genetic diversity[51], but also underline that more extensive
348 sampling is required to fully capture commonly segregating SVs.
349
350 We identified 6.6% of genes in the pangenome as dispensable; this is substantially
351 lower than most other pangenomes, which typically report 20-60% of genes or gene
352 families as being variable[3,15,75–88] (Extended Data Figure 9C, Supplementary Table 6).
353 As noted above, our pangenome only contained variants called in at least three
354 individuals, so potentially misses the rarest dispensable genes. However, a substantial
355 proportion of the difference in gene number is explained by our filtering of genes not

356 reliably associated with an SV. Without this filter, a more typical 33% of genes in the
357 pangenome are classified as dispensable (Extended Data Figure 9C, Supplementary
358 Table 6). We believe this filter is justified as variability in the annotation process has
359 previously been found to underlie putatively dispensable gene sets[85,88]. Indeed,
360 previous pangenome studies have identified “core” genes as being more highly
361 expressed[7,77,80,83,89,90]. This relationship could be due to these genes performing critical
362 biological functions, but also could be the result of more highly expressed genes having
363 more RNA-seq data evidence and being more accurately annotated as a result.
364 Functional groups typically found as enriched among dispensable genes in other
365 studies, such as defense and auxin response[3,7,77,85,87,91] were enriched in our initial set
366 of variable genes, but not following our filtering process. It should however be stressed
367 that there was no clear functional distinction between the final set of dispensable genes
368 and those removed during filtering, and that RNA-seq data from the specific samples
369 used to construct the pangenome may have resulted in more accurate annotation.
370 Nevertheless, our results emphasise the importance of confirming that putatively
371 dispensable genes are driven by underlying sequence differences, to distinguish them
372 from possible non-biological annotation variability. The genes we identified as
373 dispensable with high confidence include 133 with defense-related functions,
374 emphasising the importance of pangenomes in understanding the genetic basis of
375 disease resistance traits.
376
377 Using a pangenome reference allows improved short-read mapping for *F. excelsior*, as
378 has been found for other species[11], and we were able to successfully call individual

379 SNPs and SVs from individually sequenced samples. SV genotyping from short-reads is
380 generally more challenging than SNPs[92–94], and we are not aware of previous studies
381 calling allele frequencies of poolseq data mapped to a pangenome graph. Our results
382 indicate poolseq data can be used to estimate allele frequencies accurately for most
383 SVs, but the approach generates more discrepancies with individual based genotyping
384 than is the case for SNPs. SVs showing these discrepancies were particularly likely to
385 be identified as significant in a GWAS using a CMH test, resulting in many likely false
386 positives. The development of dedicated software for estimating SV allele frequencies
387 from poolseq data, equivalent to those available for SNPs, is required in order to make
388 full use poolseq data in the pangenome era. Nevertheless, some of the SVs identified
389 as significant from the CMH test are potentially involved in low ADB susceptibility –
390 future studies using individually genotyped data could be used to confirm this.
391
392 Analysis of SNP allele frequency estimates from the Stocks et al.[42] poolseq data did not
393 find allele frequency shifts that were consistent across the majority of provenances
394 sampled, which were mostly of UK origin. SNPs identified as significantly associated
395 with low ADB susceptibility by a CMH test mostly changed frequency between healthy
396 and unhealthy pools for a small number of provenances. This pattern can sometimes be
397 explained by the fact that the site was not polymorphic in some of the pairs. For those
398 provenances where allele shifts were detected, these were not always in a consistent
399 direction; such patterns could reflect epistasis, or interactions with test site effects.
400 Similarly, windows of high F_{st} between healthy and unhealthy pools, when shared
401 across provenances, were often shared by diverse combinations of provenances. These

402 results suggest that different populations possess different alleles associated with low
403 ADB susceptibility. If these are additive, crossing individuals from different provenances
404 may lead to improved resistance to this disease, although possible epistasis or site-
405 specific effects could confound this. Variation in allele presence or effects in different
406 populations may contribute to the difficulty in identifying individual SNPs significantly
407 associated with low ADB susceptibility in diverse sets of populations[43,44], and is
408 consistent with a lack of variability of many of the SNPs associated with low ADB
409 susceptibility in Stocks et al. (2019) in other European populations[48].
410
411 The potential importance of SVs in mediating low ADB susceptibility is highlighted by
412 the fact that the region most strongly associated with this trait in our SNP analyses likely
413 represents an insertion sequence. This variant is segregating in a minority of
414 provenances, but seemingly has a strong (if not entirely consistent) association with low
415 susceptibility to ADB. Despite a wide sampling, with particular emphasis on UK
416 provenances, this candidate SV was not detected by the individual SV calling in this
417 study. This is consistent with the number of SVs captured in the pangenome failing to
418 plateau when adding the fiftieth sample, despite a large total number of SVs and
419 substantial extra sequence being added (Figure 2). Nevertheless, our newly developed
420 pangenome encompasses a diverse set of SVs from across the geographic range of
421 European ash, including a large quantity of previously uncharacterised sequence and
422 dispensable genes, making it an essential resource for full investigation of the role of
423 both SNPs and SVs in key traits of this important tree species.

Methods

Reference assembly

To construct a phased, chromosome-level genome, we sampled tree 2451S, which was produced in 2002 by the FRAXIGEN project via the selfing of a tree in UK native seed zone (NSZ) 403 and archived at the Earth Trust. We sampled a potted graft of 2451S growing at Royal Botanic Gardens, Kew. The graft was created by Forest Research in 2017 using a rootstock from a *F. excelsior* individual of unknown provenance. Leaf tissue from this graft was flash frozen in liquid nitrogen in May 2022 and sent to Cantata Bio, CA, USA (formerly Dovetail Genomics) on dry ice, where DNA extractions, library preparation, sequencing and genome assembly was performed as outlined below.

DNA was extracted by Cantata Bio using an in-house CTAB protocol and quantified using Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). The SMRTbell Express Template Prep Kit 2.0 (PacBio, Menlo Park, CA, USA) using the manufacturer recommended protocol was used to construct a PacBio SMRTbell library (~20kb) for PacBio Sequel. The library was bound to polymerase using the Sequel II Binding Kit 2.0 (PacBio) and loaded onto PacBio Sequel II). Sequencing was performed on PacBio Sequel II 8M SMRT cells, resulting in 66Gb data (approximately 75X coverage based on an estimated 1C genome size of 877Mb[51]. For the Omni-C libraries, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with DNAse I, chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter containing ends. After proximity

445 ligation, crosslinks were reversed and the DNA purified. Purified DNA was treated to
446 remove biotin that was not internal to ligated fragments. Sequencing libraries were
447 generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-
448 containing fragments were isolated using streptavidin beads before PCR enrichment of
449 each library. The library was sequenced on an Illumina HiSeqX platform to produce
450 approximately 30X sequence coverage in 150bp paired-end reads. The combined
451 Illumina and PacBio data was used to produce an initial assembly with hifiasm v0.15.4-
452 r347[95] using default parameters. The *de novo* assembly and Dovetail OmniC library
453 reads were used as input data for HiRise[96] to produce a draft assembly. OmniC library
454 reads were then aligned to the draft input assembly using BWA[97] and reads with MQ <
455 50 removed. The mapping information was analyzed using HiRise to produce a
456 likelihood model for genomic distance between read pairs, and the model was used to
457 identify and break putative misjoins, and to score and make joins.

458 After receiving the assemblies and raw data from Cantata Bio, the haploid genome
459 assemblies were assessed using QUAST v5.2[98] and BUSCO v5.5.4 using the
460 eudicots_odb10 database[99]. To assess potential contaminants, scaffolds from the most
461 contiguous of the two haploid genomes (henceforth referred to as “BATG-1.0”, with the
462 less contiguous haploid genome referred to as “haplotype 2”) were searched against the
463 NCBI nt database (downloaded 21/02/23) using blast+ v2.11.0[100] with settings -
464 max_target_seqs 10 -max_hsps 1 -evalue 1e-25. To assess coverage, the PacBio data
465 was mapped to the contigs using minimap2 v2.5[101]. The combined data was analysed
466 using blobtools v4.0.7[102]. Three scaffolds were identified as potential contaminants (with
467 lengths between 14,626 and 35,091kb), representing 0.009% of the total assembly size.

468 The PacBio data was used to *de novo* assemble a chloroplast genome for *F. excelsior*
469 using Oatk v1.0[103] using HMMER v.3.1b2 (hmmer.org) with parameter settings -c 30 -k
470 1001 and a chloroplast gene profile database provided by OatK (embryophyta_pltd).
471 Annotation was performed using GeSeq v2.03[104] using default settings.

472 Long-read RNA extraction and sequencing

473 Long-read RNA seq data from leaf and cambium tissue was generated to annotate the
474 new genome assembly from the reference individual. Leaf material was collected from
475 the same grafted tree as sampled for the reference assembly (see above) in May 2022
476 and flash frozen in liquid nitrogen. For cambium tissue, twigs were sampled in January
477 2023 and the layer of tissue immediately below the bark was isolated using a fresh
478 razor blade before flash freezing in liquid nitrogen. Tissue was stored in a -80 freezer
479 until RNA extraction.

480 For long-read RNA extraction, a modified version of the Qiagen RNeasy Plant Mini Kit
481 standard protocol was used (Supplementary Note 2). In brief, 25-30mg tissue was used
482 with the standard kit protocol (including the optional DNase step) and eluted in 50µl
483 RNase-free water. 450µl buffer RLT was added to the RNA, and the RNeasy kit protocol
484 was repeated (without the DNase step). RNA was quantified using a Nanodrop 2000
485 and a Promega QuantiFluor® RNA System Kit using a Quantus fluorometer. Samples
486 were shipped to the Deepseq facility at the University of Nottingham and fragment
487 length determined with an Agilent Tapestation using an RNA Screentape kit.

488 cDNA libraries were prepared using an Oxford Nanopore PCR-cDNA Sequencing Kit
489 (SQK-PCS109), loaded onto R9.4.1 flowcells and sequenced on a Gridion X5 Mk1 –
490 this was performed once for the cambium library and twice for the leaf library. Runs
491 were performed with high-accuracy basecalling (running on MinKNOW software version
492 22.12.5). All data reported relates to reads that passed a base calling quality filter
493 (PHRED score >9).

494 Long-read pangenome sampling

495 In sampling individuals for the ash pangenome, we aimed to i) sample a broad range of
496 European populations, with ii) a particular focus on UK diversity by sampling individuals
497 representing different NSZ in the UK[105] (see
498 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachmen
499 t_data/file/797909/Regions_of_provenance_and_seed_zone_map.pdf). With this
500 approach we aimed to capture a global set of variants from across the species range,
501 whilst increasing our chances of capturing SVs segregating in the mainly UK
502 populations for which short-read data is available from Stocks et al.[42].

503 In total, 50 samples were used for long-read sequencing. To maximise the chances of
504 identifying SVs segregating in the poolseq data from Stocks et al.[42] one individual from
505 each of the 13 NSZ used in the Forest Research (FR) mass screening field trial at North
506 Burlingham, Norfolk, UK (Site 16)[42] was sequenced, with the exception of NSZ-204
507 where two were sequenced, as individually sequenced data for this provenance were
508 available from Stocks et al.[42] The North Burlingham site was planted in 2013 with two-
509 year-old saplings that were grown for non-research purposes in commercial nurseries

510 from seed sources from 15 different provenances. The site contained four complete
511 replicate plots for each seed source, with 5,000 trees per ha (a spacing of 1 m× 2 m). At
512 the time of sampling, the site had high *H. fraxineus* inoculum pressure from naturally
513 established infections, and ADB damage was severe throughout. In 2022 (when trees
514 were eleven years old) individuals were sampled that appeared relatively healthy within
515 each plot, and scored between 3 and 5 on the damage scale of Pliura et al.[36] (we
516 excluded the score of 0, which indicates trees that died soon after planting) to increase
517 the chances of capturing alleles for low ADB susceptibility. Three wild trees of unknown
518 age from seed zones in the UK not covered by the FR trial populations (NSZ-305, NSZ-
519 406, NSZ-402) were also included to further capture UK diversity. To increase the
520 likelihood of including more of the range-wide diversity of structural variation in *F.*
521 *excelsior*, one individual from each of 29 mostly mainland European provenances was
522 sampled from the Realising Ash's Potential (RAP) trial (established in 2005), as well as
523 two from provenances in the FRAXIGEN trial (established in 2004), both at Paradise
524 Wood, Earth Trust, Oxfordshire. Finally, leaf material from the individual used for the
525 linear reference genome was also included. Sample information is outlined in
526 Supplementary Table 3, including approximate coordinates of origin where available. A
527 map based on these coordinates is shown in Figure 1A , overlaid with a shapefile of the
528 *F. excelsior* native range from Caudullo et al.[106].

529 Sampling for long-read DNA sequencing was performed in May 2022. For samples
530 collected in the field, intact stems with leaves were excised from individual trees using
531 secateurs or a pole pruner, wrapped in damp tissue paper and sealed in airtight plastic
532 bags before returning to RBG Kew for storage at 4°C. Leaf tissue was removed from the

533 stems and frozen within 48 hours of collection by immersing in dry ice, and then stored
534 at -80°C prior to DNA extraction.

535 Long-read DNA extraction and sequencing

536 For DNA extraction, a modified CTAB protocol was used[53] (see Supplementary Note 3
537 for full protocol). In brief, 300-400mg tissue was disrupted in liquid nitrogen and added
538 to 8ml Carlson buffer with 20µl beta-mercaptoethanol and 40µl Qiagen Proteinase K.
539 This was incubated at 65ºC for 1 hour, before adding 16µl Qiagen Rnase A and
540 incubated for a further 15 minutes. 1 volume of SEVAG solution was added, the tube
541 inverted, and centrifuged for 10mins at 4ºC before the aqueous phase was removed.
542 This step was then repeated. 0.7 volumes of isopropanol was added and the tube
543 incubated at 4ºC for 15mins. The sample was centrifuged at 4ºC for 30mins to
544 precipitate the DNA and resuspended in Qiagen buffer AE before being redissolved in
545 5ml Qiagen Buffer G2 for 15 minutes. A Qiagen Genomic Tip 100/G Kit was used to
546 purify the DNA; the “Isolation of Genomic DNA from Blood, Cultured Cells, Tissue,
547 Yeast or Bacteria using Genomic-tips” protocol from the kit handbook was used, with
548 the modification of using 10ml rather than 7.5ml buffer QC. DNA was eluted in 100µl
549 buffer 1x TE and left to resuspend at room temperature and then at 4ºC for up to 3 and
550 4 days respectively. DNA was quantified using a Quantus Flurometer with QuantiFlour
551 dsDNA kit, and a Nanodrop 2000. Integrity was measured using an Agilent Tapestation
552 with a Genomic DNA kit, and the DNA shipped to the University of York on dry ice to
553 prevent fragmentation during transit, where library preparation was performed as
554 described below.

555 Short fragment elimination was performed using a Circulomics SRE Standard Kit.
556 Library preparation was performed using the Oxford Nanopore Technologies (ONT)
557 SQKNBD114.24 (Protocol version NBE_9169_v114_revD_15Sep2022-promethion.pdf)
558 library preparation kit, with the following modifications: i) Reaction volumes were
559 proportionately scaled up to 60µl total for end repair and A-tailing, and incubation times
560 increased to 30 minutes each at 20°C and 65°C, ii) Reaction volumes were doubled for
561 the barcode adapter ligation step, which was performed at room temperature for 1 hour.
562 Barcode-ligated samples were bead cleaned separately before pooling at appropriate
563 volumes prior to adapter ligation. Native barcodes were ligated and reads demultiplexed
564 by barcode following the run. Libraries were loaded onto PromethION flowcells, runs
565 commenced on a Gridion with live super accuracy base calling (running on MinKNOW
566 software version 22.10.5) with barcode demultiplexing and adapter trimming. All data
567 reported relates to reads that passed a base calling quality filter (PHRED score >10).

568 Long-read data analysis

569 Simplex data for the ONT individuals was trimmed, firstly using Trimmomatic v0.39[107]
570 with settings LEADING:7 TRAILING:7, followed by Nanofilt v2.8.0[108] with minimum
571 average read quality of 7 and length 1000. Quality was inspected and summary
572 statistics generated using Nanoplot v1.4.10[109], FastQC v0.11.9[110] and multiQC v1.14[111].

573 Reads were mapped to BATG-1.0 using minimap2 v2.5[101], with PCR duplicates and
574 reads with mapping quality (MAPQ) <20 removed using samtools v1.9[97].

575 To inspect population structure of these samples, SNPs were called individually for all
576 samples other than the reference genome individual using Clair3 v.1.0.5[112] with model
577 r941_prom_sup_g5014. Genotypes with DP < 8 or > 60 were set to missing and SNPs
578 merged using bcftools v1.16[113]. Sites with >10% missing data or MAF < 0.05 were
579 excluded. Plink2 v2.00a3.3LM[114] was used to identify unlinked SNPs (using setting –
580 indep-pairwise 50 10 0.1) and calculate a PCA using default settings.

581 Read-based SV calling was performed using the long-read data generated for all 50
582 samples using Sniffles2 v2.0.7[59] and cuteSV v2.0.3[60], two tools that rank highly across
583 several comparisons of read-based SV calling software[115–122]. Both tools were run with
584 a minimum read support of 5, and cuteSV with the additional parameters –
585 max_cluster_bias_INS 100 –max_cluster_bias_DEL 100 –diff_ratio_merging_INS 0.3 –
586 diff_ratio_merging_DEL 0.3. Breakends (BND) or sites which were genotyped as
587 homozygous reference were removed.

588 To complement the read-based SV calling, *de novo* assemblies were produced from the
589 ONT read data. *De novo* assembly from trimmed reads was performed using shasta
590 v0.11.1[62] with the Nanopore-May2022 config file, Flye v2.9.2-b1786[61] with default
591 settings, and nextDenovo v.2.5.0[63] with an estimated genome size of 850Mb. Assembly
592 contiguity was assessed using QUAST. SVs were called using these assemblies by
593 mapping them to BATG-1.0 using minimap2 in asm10 mode. SVs were called with
594 svim-asm v.1.02[123] using alignments with a minimum MAPQ of 20 and minimum SV
595 size of 50. Haplotype 2 of the linear reference genome was also mapped to BATG-1.0,
596 and SVs called using the above method. SV calls between the two pseudochromosomal
597 haploid genomes were compared to SV calls from the ONT data from the same

598 individual using the various approaches, with SV sets merged using SURVIVOR
599 v1.0.7[124] using settings 200 2 1 1 1 50 for the intersection of sets and 200 1 1 1 1 50 for
600 the union. Assuming the SV calls between the two haploid linear reference genomes
601 are a “truth” set, precision, recall and F1 were calculated for the various SV calling
602 methods from the lower coverage ONT data. Based on these comparisons (see
603 Results), SVs for each individual were included if they were called in at least one of i)
604 cuteSV and Sniffles2 ii) svim-asm from mapping the shasta assembly to BATG-1.0.

605 These SVs were merged across all 50 individuals using SURVIVOR with settings 200 2
606 1 1 1 50. SVs that overlapped with Ns in the reference were removed. SVs that were
607 called in at least 3/50 individuals were identified using bcftools v1.16 and used for
608 downstream analysis. The overlap between these SVs and the repeat annotation was
609 assessed using the bedtools v2.28.0[125] window function with parameter -w 1.
610 Consensus sequences for insertions and deletions were searched against the ncbi nt
611 database using blast+ with parameters -max_hsps 1 -evalute $1e^{-25}$, with the resulting
612 data loaded into blobtools. Based on analysis with blobtools, four sequences were
613 identified as putative contaminants and the associated SVs removed from the vcf. The
614 resulting vcf, as well as the BATG-1.0 fasta was used to construct a pangenome using
615 the vg v1.64.0[66] construct command with option -v -S -a.

616 Linear reference and pangenome annotation

617 Repeats were identified using Repeatmodeler v2.0.1[126] in LTRStruct mode. The outputs
618 of this were combined with the custom repeat library used previously to mask the
619 BATG-0.5 assembly[51], and used to mask the genome with Repeatmasker v4.1.2[127] with

620 the parameters -wublast -nolow. Telomere sequences were identified using Teloscope
621 v0.0.9[128] with default settings using the canonical plant telomeric repeat CCCTAAA. To
622 annotate the linear reference genome, short-read RNA-seq data from leaf, root,
623 cambium and flower tissue that had been used to annotate the BATG-0.5 assembly[51],
624 was downloaded from ENA and mapped to BATG-1.0 using STAR v2.7.10b with default
625 parameters[129]. The long-read RNA-seq data generated in this study was trimmed firstly
626 using Trimmomatic v0.39[107] with settings LEADING:7 TRAILING:7, followed by Nanofilt
627 v2.8.0[108] with minimum average read quality of 7 and minimum length 100. The trimmed
628 data was then mapped to the BATG-1.0 reference genome using minimap2 v2.5[101] .
629 Variability in the gene annotation process could lead to false positives during the
630 identification of variable genes[16,130], so we repeated the annotation process multiple
631 times and combined the results. BRAKER3[131] was run 10 times on the softmasked
632 genome using the bam files from the short read data, and protein ortholog data from
633 Viridiplantae from orthodb v11[132]. A version of BRAKER3 capable of using long-reads
634 (singularity build braker3_lr.sif docker://teambraker/braker3:devel) was also run 10
635 times using the bam files from the long read data and the same protein ortholog set as
636 used with the short read data. To combine the outputs of the runs, the amino acid
637 sequences of the longest transcript per gene (minimum length 30 amino acids) for each
638 run were clustered into orthogroups using Orthofinder v2.5.5[133]. Within each orthogroup,
639 groups of sequences with non-overlapping positions within the BATG-1.0 assembly
640 (likely representing recent duplicates) were separated using the bedtools intersect
641 utility. For each of these groups, the gene model with the longest transcript from the set
642 of 20 BRAKER3 runs was chosen. The features for the selected gene models were

643 combined into a single bed file. The amino acid sequences from the final set of
644 transcripts was used as the input for Orthofinder, along with protein sequences for
645 BATG-0.5[51], *Fraxinus pennsylvanica*[54], *Olea europaea* subsp. *europaea*[134], *Solanum*
646 *lycopersicum*[135], *Erythranthe guttata*[136], *Sesamum indicum*[137] and *Salvia miltiorrhiza*[138]
647 (Supplementary Table 2).

648 To annotate the pangenome, sequences and locations of structural variants were used
649 to modify the sequence of the BATG-1.0 fasta, which was then annotated. As most SVs
650 were defined from read mapping to BATG-1.0 rather than from the individual *de novo*
651 assemblies, this approach maximises the opportunity for novel genes/gene deletions to
652 be explicitly linked to underlying SVs. The steps were as follows (see Extended Data
653 Figure 10 for an explanatory diagram): for each individual, a non-overlapping set of SVs
654 detected in that individual was extracted from the combined vcf of SVs used to build the
655 pangenome. Where SVs overlapped, insertions were retained in priority to other types
656 of SVs, and longer SVs were retained in priority to shorter ones. Based on these sets of
657 SVs, sequence was then inserted, deleted, or inverted in the reference fasta (the SV
658 sequence in the combined vcf of SVs used to build the pangenome was used, so the
659 same sequence was used to represent a given SV in different individuals) to produce a
660 transformed fasta for each individual in turn. The transformed fasta was then annotated
661 for each individual. Repeats were annotated and softmasked, and long and short read
662 RNA-seq data from the reference individual were mapped to each transformed fasta
663 using the same methods as for the reference sequence. A single run of the short-read
664 BRAKER3 protocol, and a single run of the long-read BRAKER3 protocol was
665 performed as detailed above. For these two runs for each individual, the Orthofinder

666 approach as outlined previously was used to consolidate the outputs of BRAKER3 into
667 a single set of proteins. Orthofinder was then run again for each set of proteins across
668 the 50 individuals, as well as the proteins from the consensus annotation of BATG-1.0.
669 Non-overlapping groups of sequences were separated using bedtools as before. For
670 each gene, bedtools intersect was used to identify SVs that overlapped the gene
671 sequence. Novel genes not present in the BATG-1.0 genome should overlap with an
672 SV, as should deletions of genes present in the reference genome. For each novel gene
673 that overlapped a SV at least once (i.e. in at least one sample, including the reference
674 individual sequenced with long reads), an F1 score was calculated to identify the
675 strength of the association between the new gene and the SV: N_Gene_SV / (
676 N_Gene_SV + 0.5*N_Gene_NSV + 0.5*N_NGene_SV) where N_Gene_SV is the
677 number of samples with the gene and the SV, N_Gene_NSV is the number of samples
678 with the gene but without the SV, and N_NGene_SV is the number of samples without
679 the gene but with the SV. For novel genes overlapping more than one SV, the SV with
680 the highest F1 score was retained as potentially causal for the novel gene. An
681 equivalent calculation was performed for genes present in the BATG-1.0 genome but
682 absent in at least one sample, where an SV overlapped with the annotation coordinates
683 for the gene in the reference. Genes with an F1 score of 0.8 or above were retained in
684 the final pangenome annotation as being dispensable; for those with an F1 score of <
685 0.8, those absent in the BATG-1.0 genome were removed, and those present in the
686 BATG-1.0 genome were classified as invariant. Genes did not overlap an SV in any
687 sample were classified as invariant if they were annotated in BATG-1.0 and were

688 absent in some other samples, and discarded if they were absent in BATG-1.0 but
689 present in some other samples.

690 The complete set of potential pangenome genes (i.e. before removing those with an F1
691 score <0.8) was functionally annotated. Protein sequences were searched against the
692 NCBI SR database (downloaded 20/08/24), and UniProt-SwissProt database
693 (downloaded 24/07/24) using DIAMOND v2.1.9.163[139] in blastp mode. Interproscan
694 v5.69.101.0[140] with default settings was used to search for protein domains. GO terms
695 were assigned using eggNOG-mapper v2.1.12[141] by using DIAMOND blastp to search
696 the eggNOG database. GO enrichment was performed using topGO v2.52.0[142] using
697 annot = annFun.gene2GO, nodeSize = 10 and the exact Fisher test with Bonferonni
698 correction. OrthoFinder was run with the complete set of potential pangenome protein
699 sequences using the same species and parameters described above.

700 SNP and SV calling from short read data, from individuals and 701 pools

702 To test the impact of using the pangenome on read mapping accuracy, short-read data
703 from 42 individuals separately whole genome sequenced in Stocks et al.[42] (this number
704 selected to reflect the average number of individuals in a pool in that study, see
705 generation of *in silico* pools below) was mapped using vg giraffe[66] to either i) the full
706 pangenome or ii) the BATG-1.0 genome converted into pangenome format using vg
707 construct. The resulting gam files were converted to bam files in the BATG-1.0
708 coordinate system using vg surject. Mapping statistics were calculated using bcftools
709 v1.19.

710 For SV calling from these individuals, vg pack was used with parameter -Q 5, and SVs
711 called using vg call with the vcf used to construct the pangenome as a reference. For
712 SNP calling, bams were produced using vg surject, with PCR duplicates and reads with
713 MAPQ < 20 removed using samtools. SNPs were then called using bcftools v1.19 call.
714 SNPs and SVs were merged across individuals, and calls with DP < 5 or >50 were set
715 to missing, then sites with >10% missing individuals or a MAF of <0.05 were removed.
716 To produce similar numbers of SNPs and SVs for linkage analysis, 1% of SNPs were
717 randomly sampled using vcflib 1.0.3[143] function vcfrandomsample. SNP and SV data
718 were combined using bcftools v1.19 concat. Pairwise linkage between markers was
719 calculated using plink v1.9.-170906[144], with parameters –ld-window-kb 1000 and –ld-
720 window 99999. Linkage decay with distance between SVs alone, between SNPs and
721 SVs, and between SNPs alone were analysed separately.

722 To identify the accuracy of SNP and SV calling from poolseq data using the
723 pangenome, read files from the 42 individually sequenced samples (which each
724 contained approximately the same number of reads, equating to approximately 22X 1C
725 genome coverage) were combined, and 10% of read pairs randomly subsampled with
726 seqkitt v2.8.0[145]. This corresponds to approximately 2.2X coverage per individual, the
727 average per-individual coverage of the poolseq data in Stocks et al.[42]. SNPs and SVs
728 were called from this *in silico* pool using the approaches above, but with sites being set
729 to missing if they had a DP < 50 or DP > 200. Allele frequencies for SNPs/SVs were
730 calculated as the number of reads supporting the alternate allele divided by the total
731 number of reads. For the merged individually called SNPs/SVs, the correlation between

732 the alternate allele frequency in the individually called samples and the estimated allele
733 frequency for the *in silico* pool was tested with a linear model in R v4.3.1[146].

734 GWAS for low susceptibility to ash dieback from poolseq data

735 A poolseq dataset from Stocks et al.[42] was used to perform a GWAS for low
736 susceptibility to ADB. In brief, this data was generated as follows: leaf samples were
737 taken from four FR mass screening trial sites in England (including the North
738 Burlingham site where individuals were sampled for constructing the pangenome), from
739 six year old ash trees from a total of 12 different seed sources (10 from UK NSZ, one
740 from Ireland [IRL-DON] and one from Germany [DEU]), as well as a group from a
741 breeding seedling orchard planted by the Future Trees Trust (FTT-SO). For each of the
742 13 sources, a pool of healthy trees and a pool of unhealthy trees (42 trees per pool on
743 average) were sequenced with the Illumina HiSeq X platform in 150bp paired-end
744 reads. Two seed sources had a second pair of pools, produced as biological replicates,
745 for a total of 30 pools (excluding one technical replicate). See Stocks et al.[42] for full
746 details. The trimmed, paired short read data from ENA (accession ERA1732352,
747 Submitted files) was used.

748 Here, these reads were mapped to the pangenome and read depths for the reference or
749 alternate paths were extracted from the vcf for each pool to produce estimates of SV
750 allele frequencies, as described in the previous section. The SV allele frequencies were
751 then merged across individuals. For SNPs, the output of vg giraffe was converted to
752 linear reference coordinates using vg surject, with PCR duplicates and reads with
753 MAPQ < 20 removed using samtools v1.9[147]. SNPs were called from the resulting bam

754 files using samtools mpileup with -d 8000 and the mpileup2sync function from
755 Popoolation2[65] with –min-qual 20 to produce a .sync file. Sites were removed if there
756 was a third most frequent allele with a total depth higher than 15. The frequencies of
757 major alleles for i) SVs and ii) a random 1% of SNPs were used to perform PCA using
758 the R prcomp function.

759 To identify SV and SNP allele frequency changes associated with ADB damage,
760 Cochran-Mantel Hahn (CMH) tests were performed using the cmh-test.pl function from
761 Popoolation2, with parameters min_count 15, min_coverage 40 and max_coverage
762 200. For sites with a significant result ($p < 1 \times 10^{-13}$), the Woolf test was used from the
763 vcd R package[148] to test for significant deviation from homogeneity across groups using
764 the alternate/reference read counts for each pool. An alternative statistical test for
765 GWAS, a quasibinomial generalized linear model (qGLM), was run using the
766 poolFreqDiffTest_QGLM.py script from the PoolFreqDiff repository[70], using parameters -
767 mincnt 15 -minc 40 -maxc 200 -n 45. For SNPs identified as significant from the CMH
768 test, nearby genes were identified by using the bedtools window utility with a window
769 size of 5kb.

770 From the .sync files generated for each chromosome, F_{st} was calculated using the fst-
771 sliding.pl function of popoolation2 for the pair(s) of healthy vs. unhealthy pools for each
772 provenance, with a window size of 500bp and step size 100bp, with –pool-size set as
773 40. Sliding windows with a window size of 50bp with step size 10bp gave similar results,
774 so are not reported here. The top 1% of values per chromosome (“outliers”) were
775 reported for each pair of pools, and windows reported which contained an outlier for
776 more than five pairs of pools.

Acknowledgements

We would like to thank i) Genomics Laboratory, Bioscience Technology Facility, University of York, ii) Deep Seq: Next Generation Sequencing Facility, University of Nottingham, iii) Cantata Bio LLC, California for their contribution to sequencing and genome assembly, iv) Earth Trust and other landowners for facilitating access for sampling. We would also like to thank Mike

Code availability

Code needed to reproduce the analysis will be made available at https://github.com/danielwood1992/Ash_pangenome.

Figures

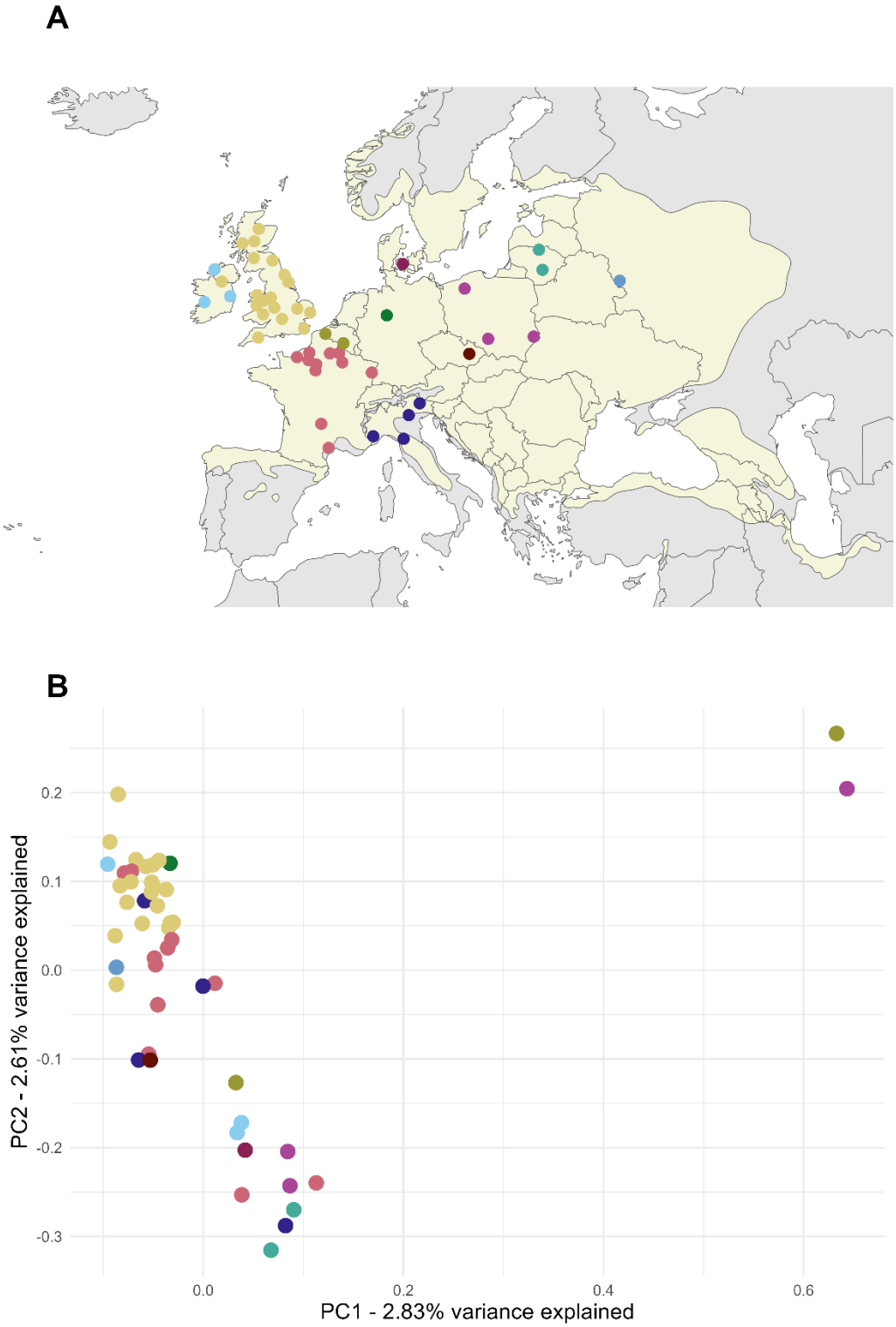

A
B
0.2
0.1
0.0
-0.1
-0.2
-0.3
PC2 - 2.61% variance explained
0.0
0.2
0.4
0.6
PC1 - 2.83% variance explained

Figure 1 - Geographic origins and genetic structure of samples

A) Map showing location of origin for individuals used to construct the pangenome; see Methods and Supplementary Table 1 for full details. Points represent individuals sampled or approximate sample locations for seedlots from which individuals were grown. Light yellow areas represent the native range of *F. excelsior* using shapefiles from Caudullo et al.[106]. B) Principal components analysis of single nucleotide polymorphism (SNP) calls from long-read sequencing data from individuals sampled. First two principal components are plotted, with the percentage of variation explained by each principal component shown on the axis labels. In Figures A and B, individuals are coloured by country of origin.

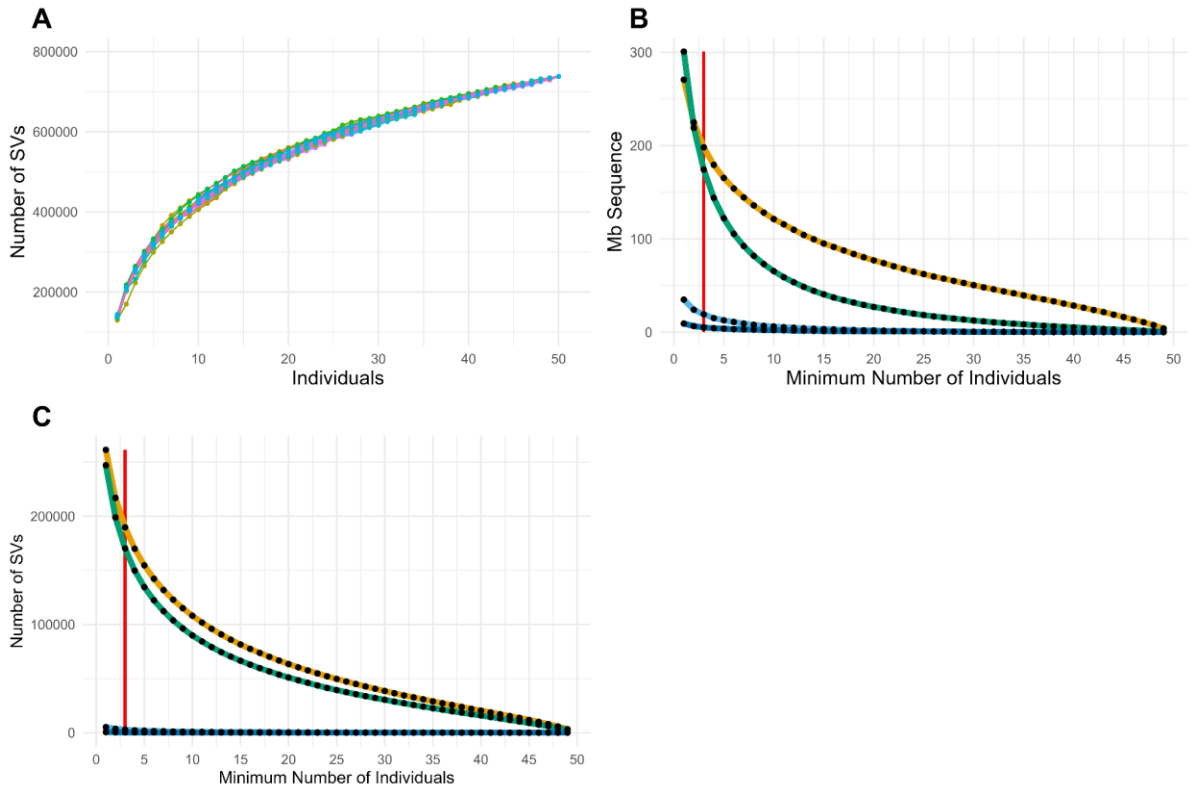


Figure 2 - Structural variant properties

A) Saturation curves of sequentially merging SVs from different individuals. Each sequential series of SVs was generated by combining individuals in a random order, one at a time, with the total number of SVs in the group being displayed; separate series of random combinations are designated by colour, with points in the same series joined by a line. The X axis represents the number of individuals merged, the Y axis the total number of SVs in those individuals. B), C): The total quantity of sequence (B) and total number of SVs (C) of different SV classes, at different minimum numbers of individuals called from long-read data. Colour-coding indicates different types of SVs: insertions (green), deletions (orange), tandem duplications (light blue) and inversions (dark blue). A vertical red line at n = 3 indicates the cutoff for SVs being included in the pangenome.

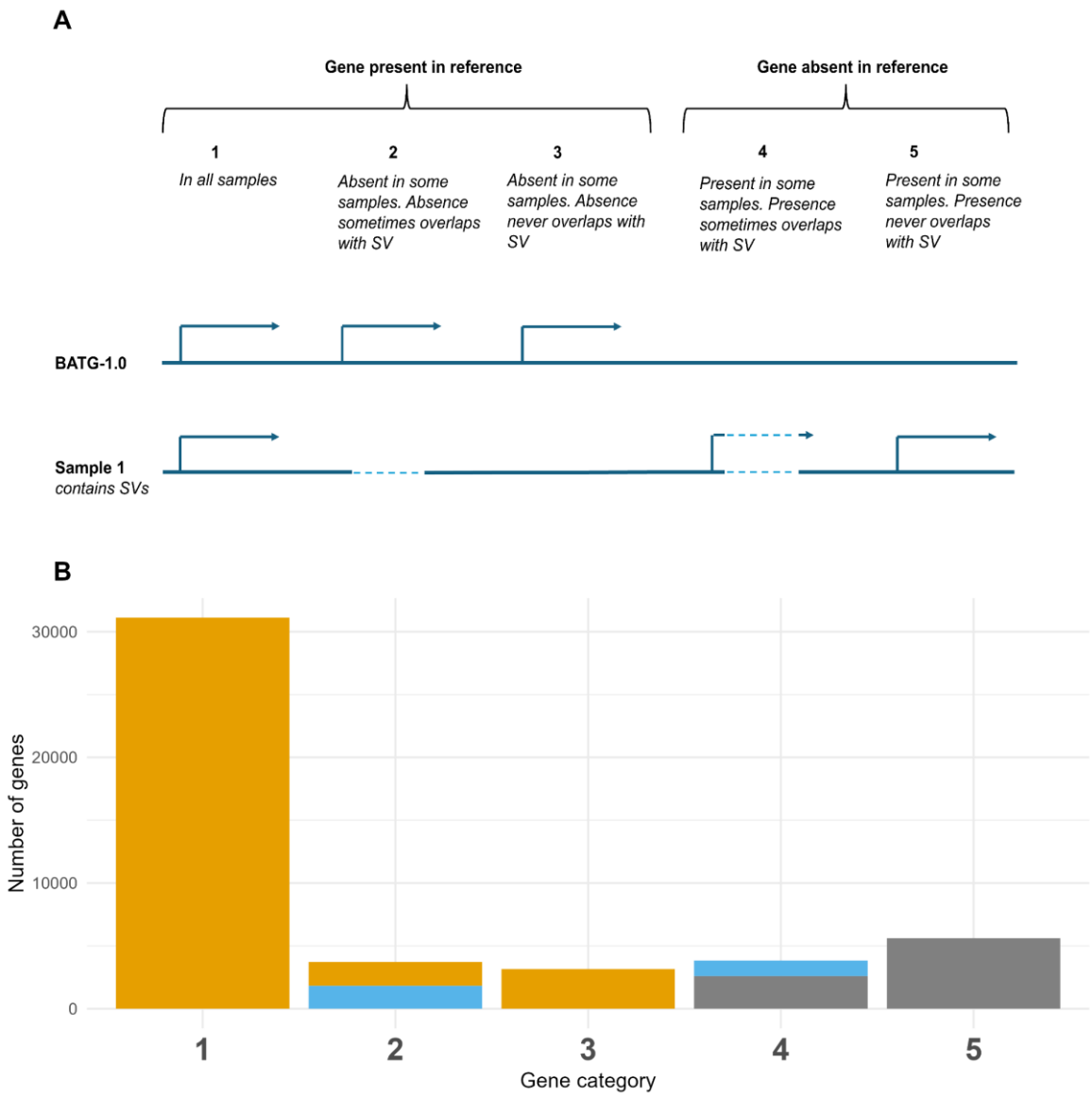
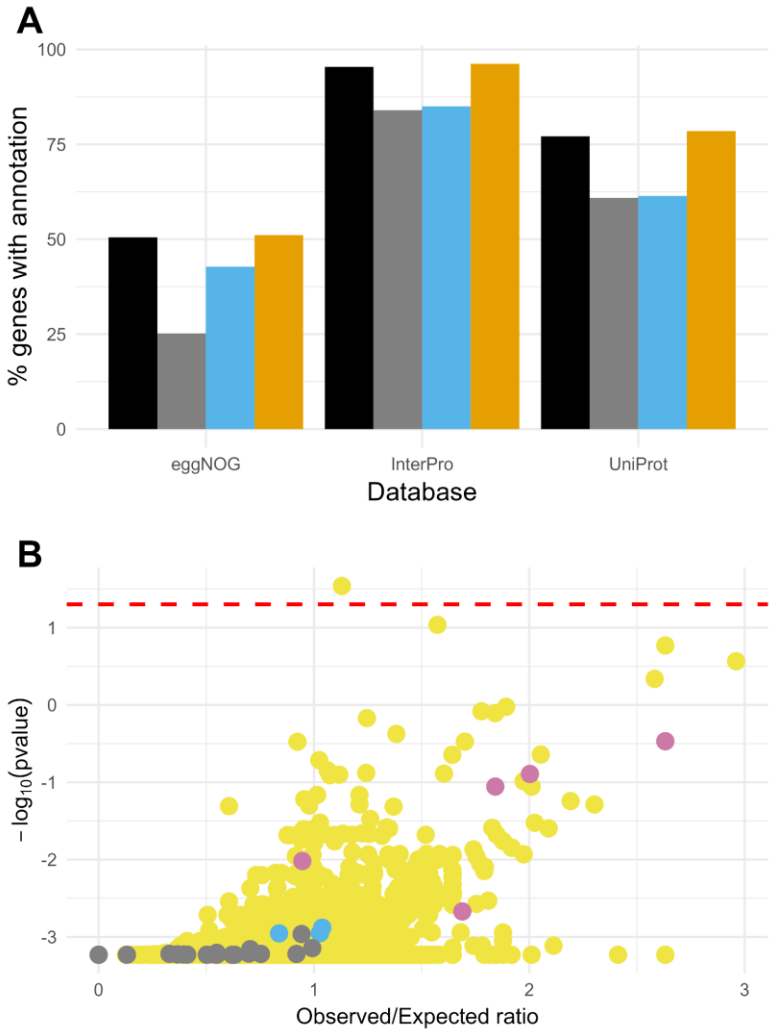


Figure 3 - Consistency of association of putatively dispensable genes with structural variants

A) Illustration of classification of genes in the pangenome based on their presence/absence, and how this corresponds with structural variants (SVs). Genome sequence is depicted as a blue solid line; SVs by light blue dashed lines (representing deleted sequence relative to the reference in this illustration). Genes are depicted by black arrows. For illustration, each gene in the reference sequence (top line) is homologous to the gene immediately below them in the

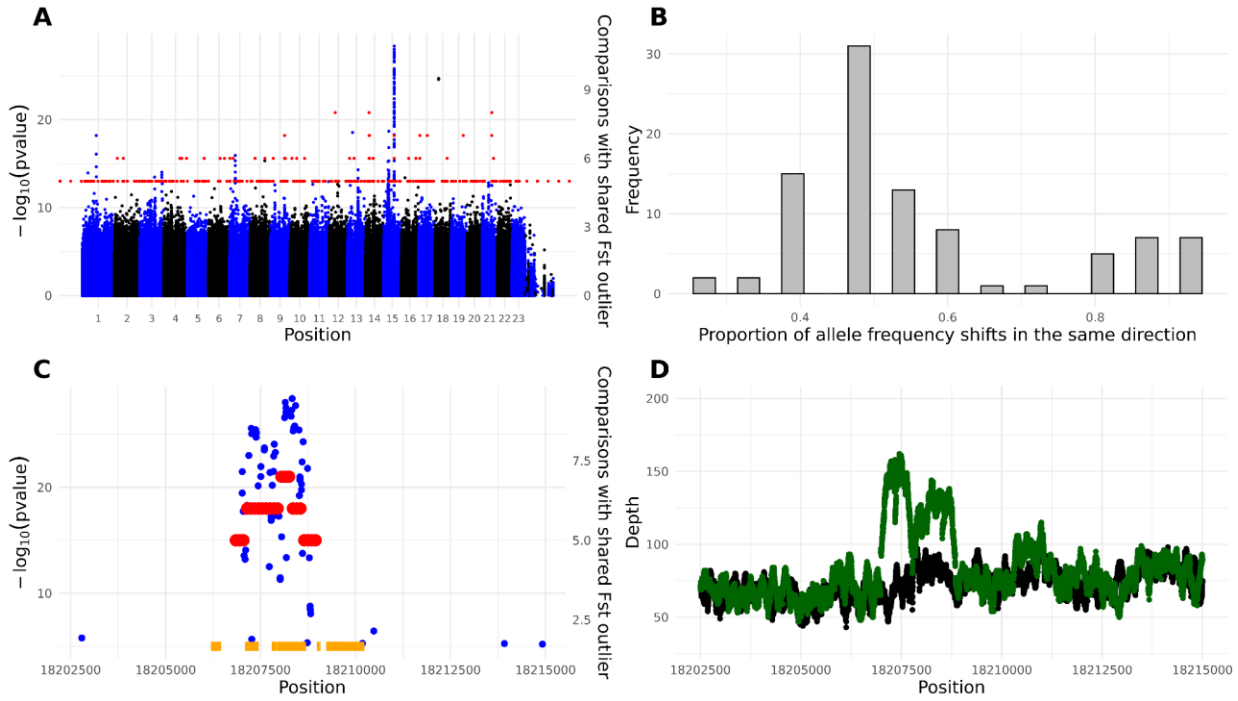
example sample (bottom line). For those present in the BATG-1.0 reference, genes can be present in every other sample (1) or sometimes absent (2, 3). The absence of a gene can either coincide with an SV overlapping the gene in at least one sample (2), or with no overlap with an SV (3). For genes absent from the reference but present in at least one sample, this presence can either coincide with an overlapping SV in at least one sample (4), or never coincide with an overlapping SV (5). B) Barplot of the number of genes in each category in the pangenome, where numbers correspond to those in part A. Following reclassification of genes in each category as invariant (orange), dispensable (blue), or discarded from the analysis (grey) according to how consistently they overlap with SVs based on an F1 score. Genes in category 2 with an F1 score > 0.8 are classified as dispensable (blue); those with a score below that threshold are categorised as invariant. Genes in category 3 are never associated with an SV so are all categorised as invariant. Genes in category 4 with an F1 score > 0.8 are classified as dispensable, those with a score below that are discarded from the analysis. Genes in category 5 are never associated with an SV so are discarded from the analysis. In an approach where genes absent in some individuals are treated as dispensable, all the genes in categories 2-5 would be treated as dispensable genes.



858
859
860 **Figure 4 - Annotation of putatively dispensable genes**

861 A) Barplots showing the percentage of genes annotated in the eggNOG, InterPro and UniProt
862 databases, for genes included in the pangenome after filtering (black), genes removed during
863 filtering (grey), included genes classified as dispensable (i.e. the high-confidence dispensable
864 genes; light blue) and included genes classified as invariant (orange). B) Results of Gene
865 Ontology (GO) enrichment analysis of the included dispensable genes, against a background of
866 the included dispensable genes plus the putatively dispensable genes removed during filtering
867 for not being consistently associated with an SV, with points coloured by significance in other
868 GO enrichment analyses. Values along the x-axis represent the ratio of the observed number of

genes in the included dispensable genes to the number of expected genes in the total background set. Y-axis values represent the $-\log_{10}$ transformed p-values from the GO enrichment test. The red dashed line indicates the significance threshold. Points are coloured according to whether those GO terms were enriched in GO enrichment tests for other sets of genes. Gray points represent GO terms identified as significantly enriched in the genes removed during filtering, against a background of excluded plus included genes in the pangenome. Blue points represent GO terms identified as significantly enriched in genes classified as variable after filtering, compared to a background of all genes included after filtering. Purple points represent GO terms identified as significantly enriched in both the previous comparisons described. The remaining GO terms are coloured in yellow.



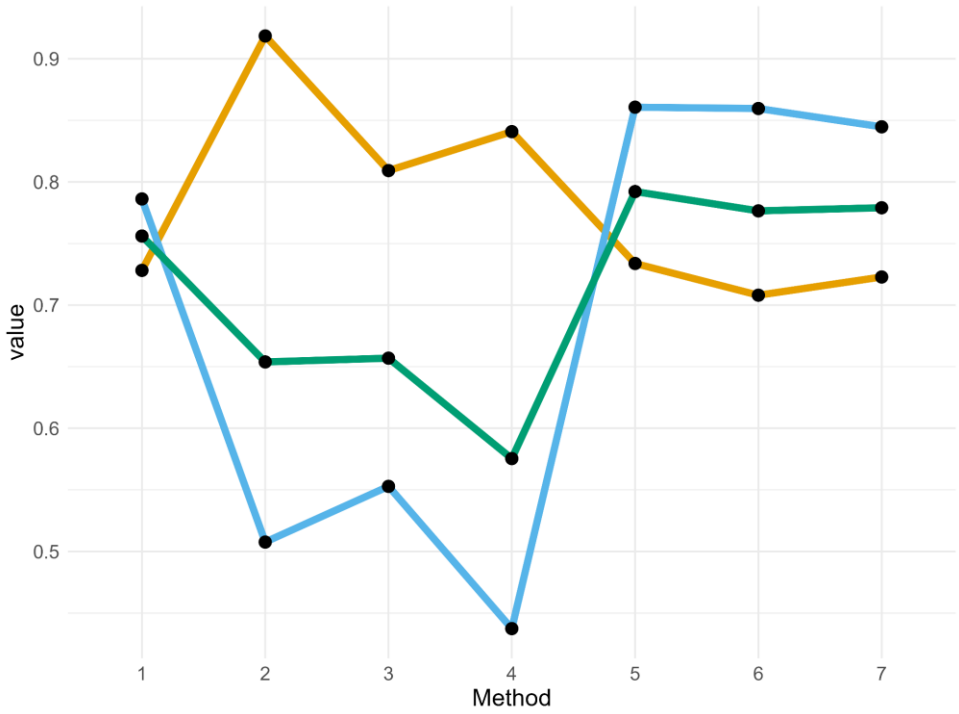
886

887 **Figure 5 - GWAS for low susceptibility to ash dieback**

888 A) Association of SNPs called from poolseq data from Stocks et al. 2019, with ash dieback
889 disease damage. Sites are ordered by position along scaffolds in the BATG-1.0 genome
890 assembly. Labels for the 23 largest scaffolds are shown. Black and blue coloured points
891 represent the -$\log_{10}$ p-values of a Cochran-Mantel Hahn (CMH) test, with alternating colours
892 representing the different scaffolds and values indicated by the left hand Y axis. Red points
893 indicate the number of 500bp sliding windows sharing an F_{st} window in the top 1% of values for
894 each chromosome, shared by a minimum number of comparisons as indicated on the right hand
895 Y axis (i.e. 5 corresponds to 5 comparisons having an Fst value in the top 1%; axes are scaled
896 so that 5 on the right hand axis corresponds to -$\log_{10}(p)$ = 13 [the significance threshold for the
897 CMH test] on the left hand axis). B) For significant SNPs ($p < 10^{-13}$), histogram of the proportion
898 of SNPs with a shift in the most common direction between healthy and unhealthy pools, across
899 all comparisons. C) Figure A, zoomed in on the only region in which SNPs with significant CMH
900 test values overlap with shared outlier F_{st} windows. Orange bars represent annotated repeats
901 overlapping the region. D: Depths of healthy (green) vs unhealthy (black) sites for the FTT-SO

provenance, for the region displayed in panel C. See Extended Data Figure 10 for equivalent plots for each of the 15 pairs of pools considered.

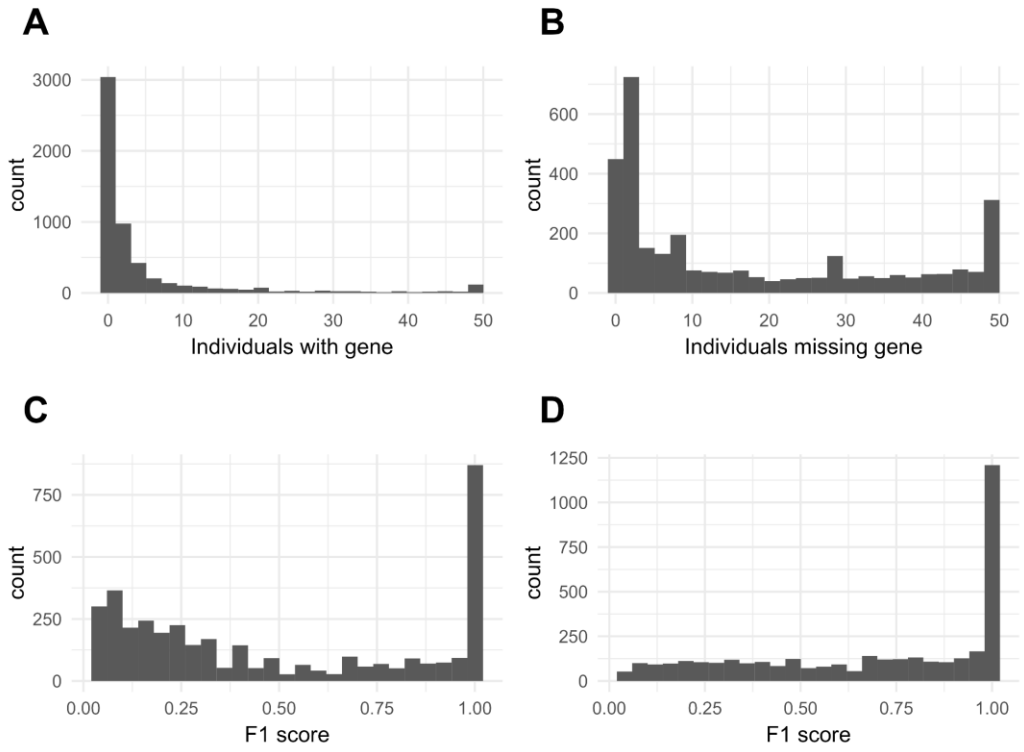
Extended Data



Extended Data Figure 1 - Precision, recall and F1 scores for different SV calling methods

Values of the precision (orange), recall (blue) and F1 score (green) of SV calling for different methods. Values were based on comparing SVs called using the long-read sequencing data from the reference individual, mapped to BATG-1.0 either as individual reads or assembled contigs, with the “truth” set comprising variants called by mapping the second haplotype of the reference genome to BATG-1.0 and calling SVs with svim-asm. The methods considered were: 1) SVs called by both cuteSV and Sniffles2; 2) SVs called by svim-asm with a mapping of the *de novo* assembly produced by shasta; 3) SVs called by svim-asm with a mapping the *de novo* assembly produced by Flye; 4) SVs called by svim-asm with a mapping of *de novo* assembly

929 produced by nextDeNovo; 5) SVs called by methods 1 and/or 2; 6) SVs called by methods 1
930 and/or 3; 7) SVs called by methods 1 and/or 4.
931
932
933
934
935
936
937
938
939
940
941
942
943

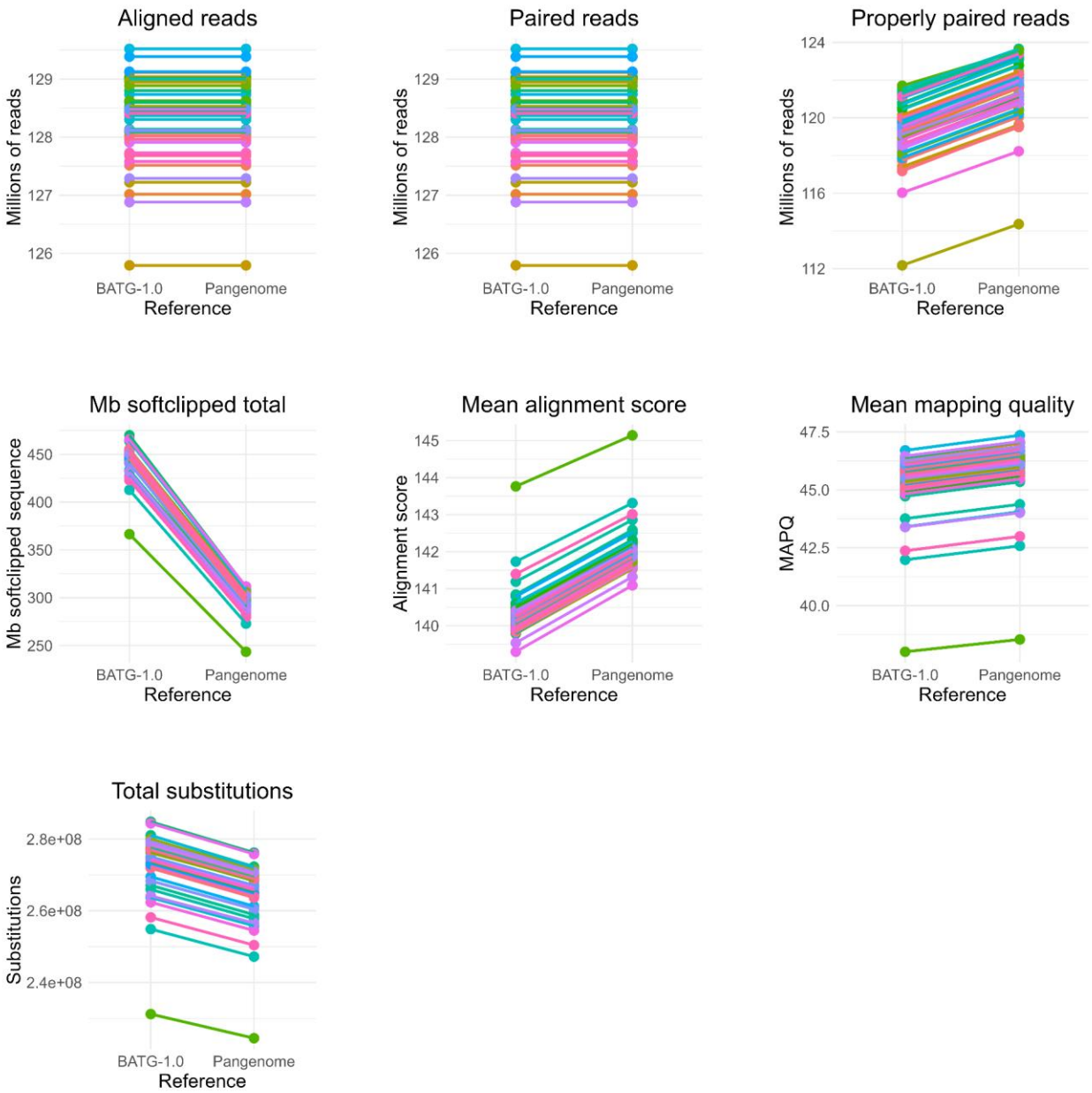


944

945 Extended Data Figure 2 - F1 scores for putatively variable genes association with SVs

946 Histograms showing A) for genes absent from the linear reference annotation but present in at
947 least one other individual, but not overlapping with an SV, the number of individuals where this
948 gene was present; B) for genes present in the linear reference annotation but absent from at
949 least one other individual, but not overlapping with an SV the number of individuals where the
950 gene was absent, C) for genes absent from the linear reference annotation but present in at
951 least one individual with an SV overlapping the gene location, the best F1 score of the SVs
952 overlapping the gene, D) for genes present in the linear reference individual but absent from at
953 least one individual where an SV overlaps with the gene sequence, the best F1 score of the
954 SVs overlapping the gene.

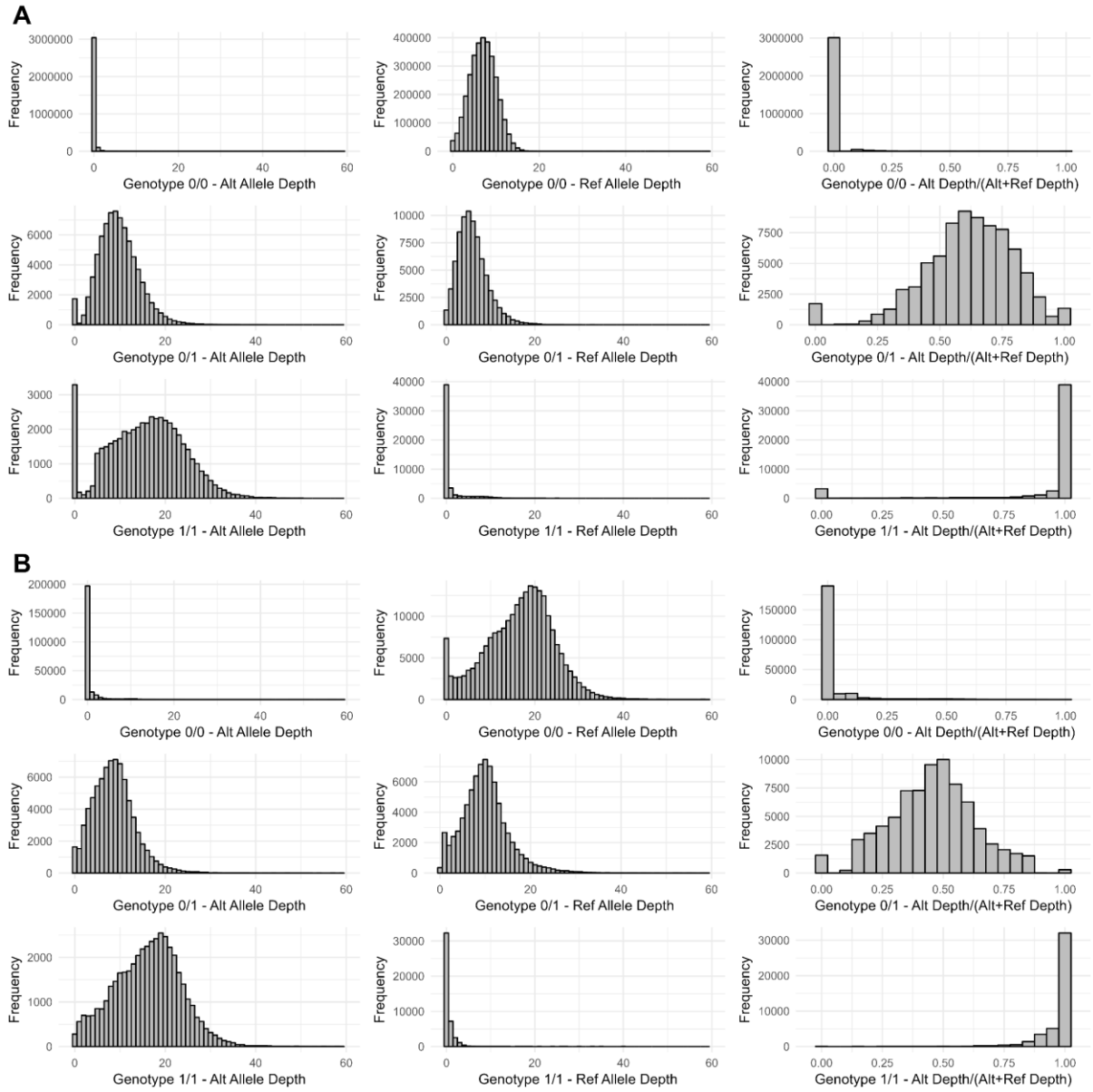
955



Extended Data Figure 3 - Read mapping statistics for the linear reference vs. the pangenome

Read mapping statistics for short reads from 42 individuals sequenced in Stocks et al. (2019), mapped using giraffe either to the linear reference genome in pangenome format (BATG-1.0), or to the full pangenome (Pangenome). Each individual is represented by a different colour -

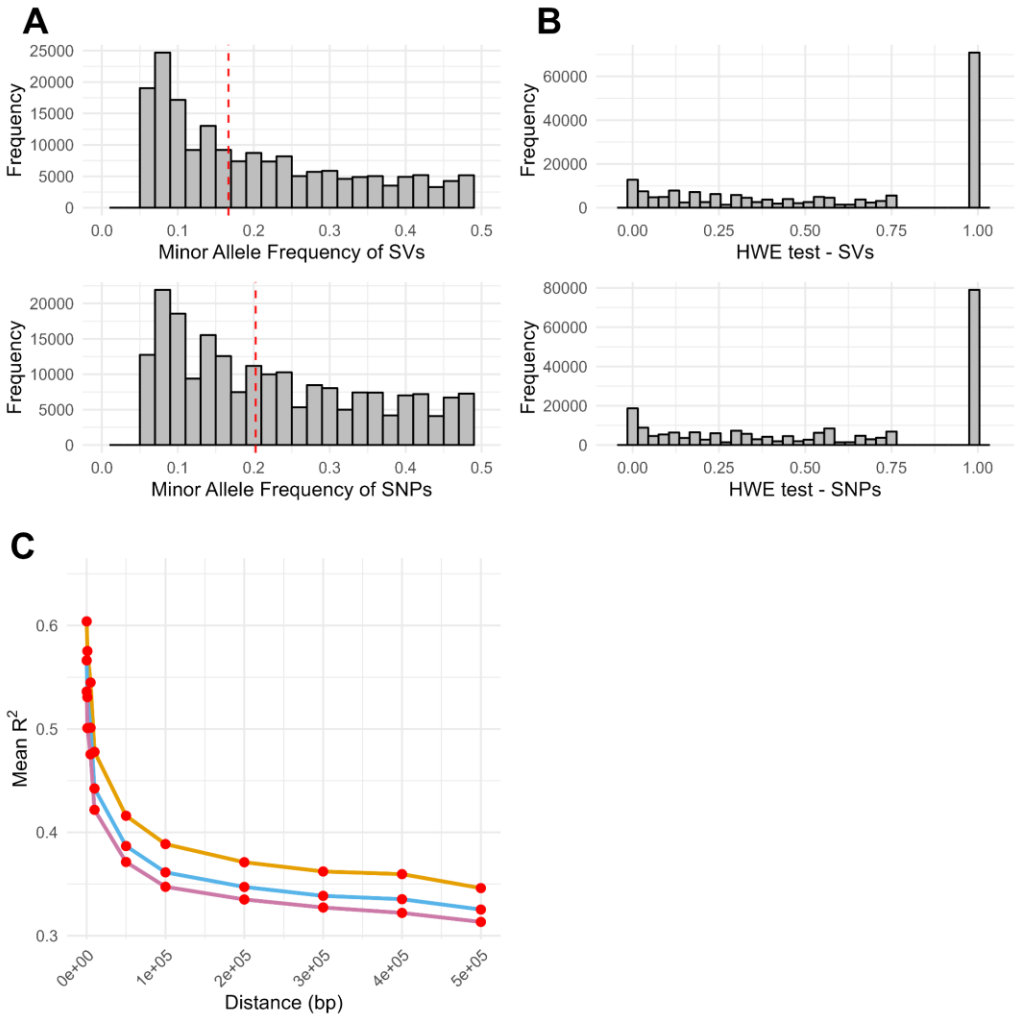
962 lines connect values in the reference-only “pangenome” to the full pangenome in the individual
963 panels. The panels outline i) the total number of aligned reads, ii) the total number of paired
964 reads, iii) the total number of properly paired reads, iv) the number of megabases of softclipped
965 sequences, v) the mean alignment score, vi) the mean mapping quality and vii) the total number
966 of substitutions, as estimated by the samtools stats function.
967
968
969



Extended Data Figure 4 - Allele depths for SNP and SV calls from data for an example individual

Histograms outlining allelic depths for A) SNPs and B) SVs called from an example individual - Individual 164 from Stocks et al. (2019). The panels indicate - for homozygous reference calls (top row), heterozygote calls (middle row) and homozygous alternate calls (bottom row) - the

977 alternate allele depth (first column), reference allele depth (second column) and the alternate
978 allele depth as a proportion of the total allelic depth (last column).
979
980
981
982
983
984

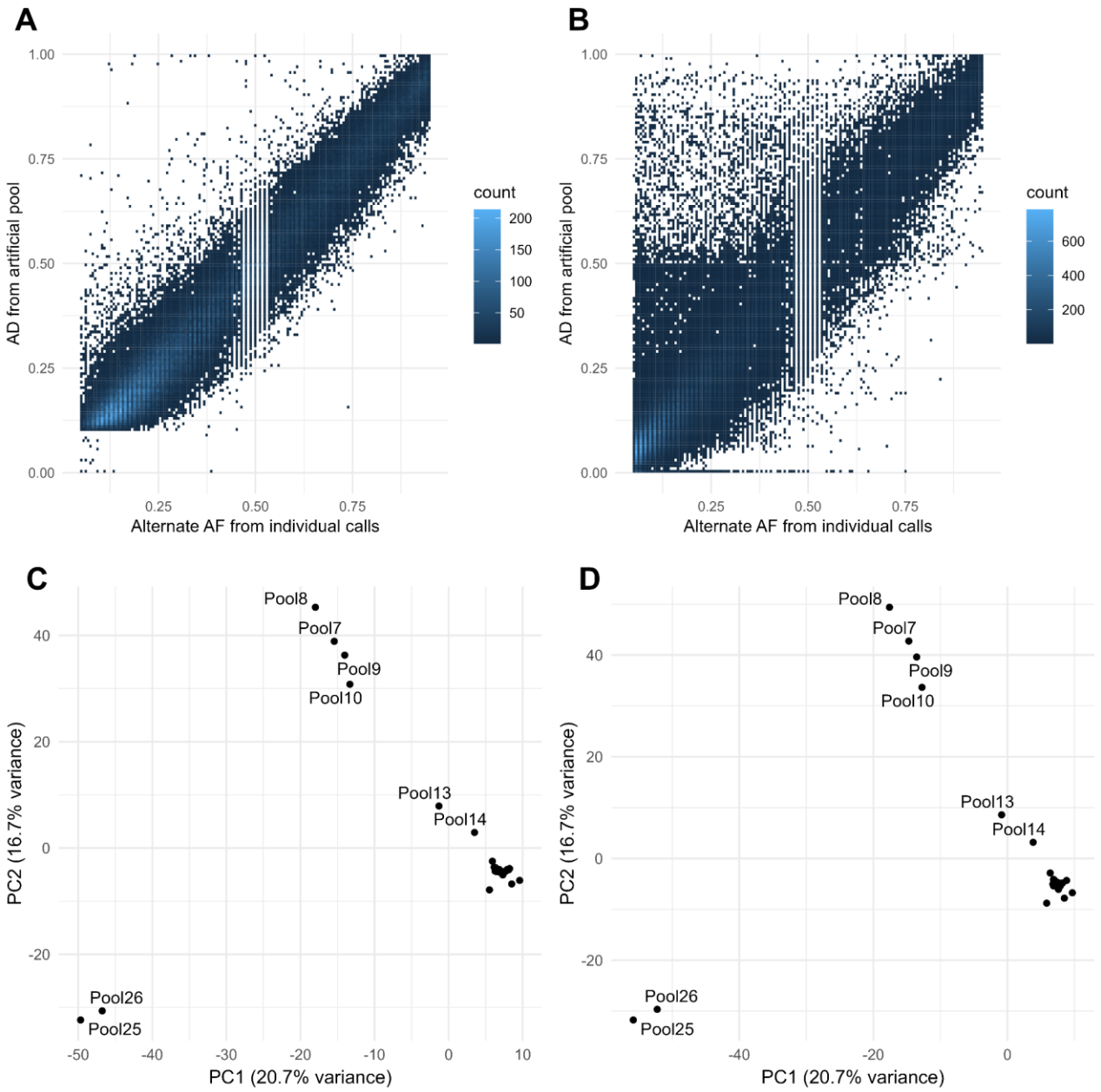


985

986 Extended Data Figure 5 - Statistics of SNP and SV calls combined across individuals

987 A) Histograms of minor allele frequencies for SNPs (top panel) and SVs (bottom panel) called
988 from a set of individually sequenced ash trees from Stocks et al. (2019). Red vertical dotted line
989 indicates the median value of minor allele frequency for both datasets. B) For the same dataset,
990 estimates of the probability of deviation from Hardy-Weinburg equilibrium (HWE) calculated by
991 bcftools. C) Estimates of linkage disequilibrium decay by distance between SNPs (orange line),
992 SVs (purple line) and between SNPs and SVs (blue line). Pairwise linkage disequilibrium was
993 estimated using plink2, with mean R^2 values in predefined bins (red points) estimated in R.

994

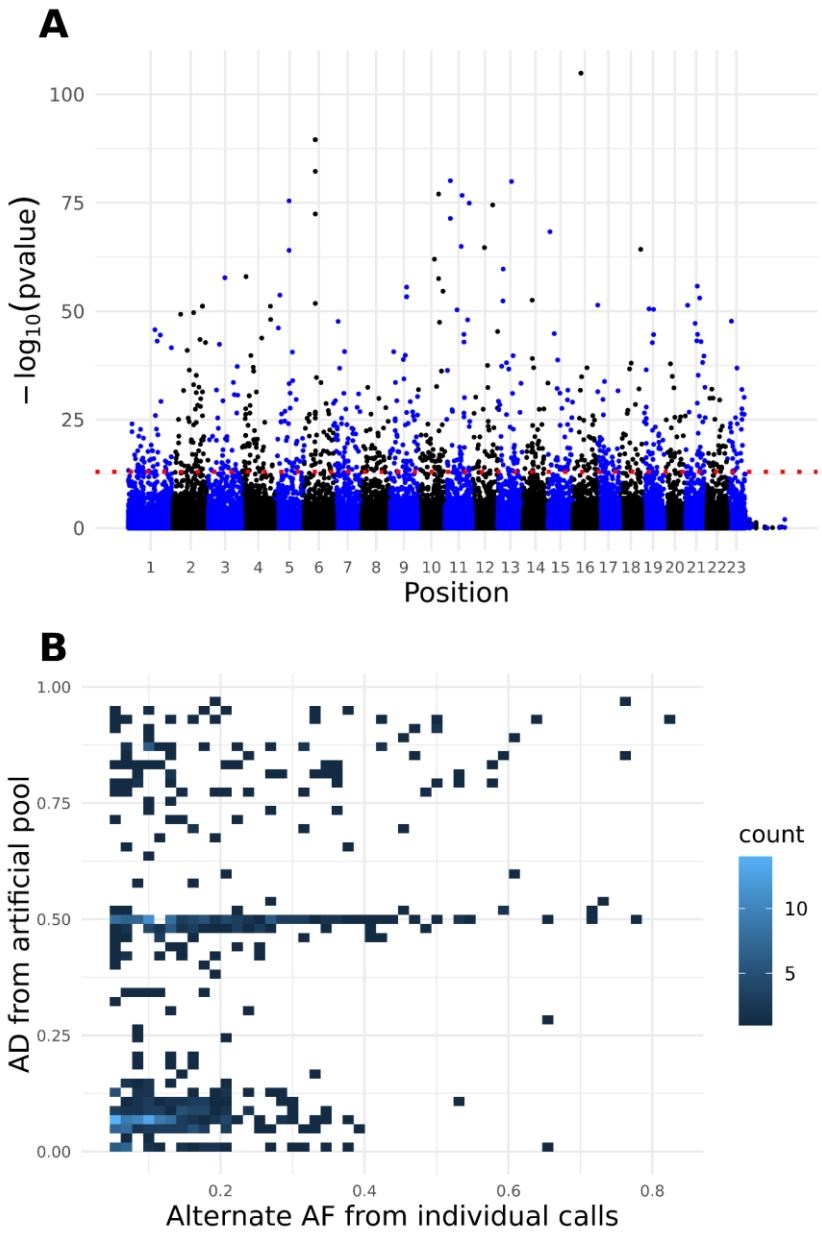


995

996 Extended Data Figure 6 - SV and SNP calling from Poolseq data

997 A), B) Heatmaps of alternate allele frequencies for SNPs (A) and SVs (B) estimated from
998 individual calls (X axis) from a subset of individually sequenced ash trees from Stocks et al.
999 (2019) vs. an allele frequency estimate from *in silico* pools generated by randomly
1000 downsampling reads for these individuals, and measuring the allelic depth (AD; Y axis). Colour

scale represents the number of sites in a square. C), D) Plots of the first two principal components (PCs) from principal components analysis from allele frequencies estimated for SNPs (C) and SVs (D) from the poolseq data from Stocks et al.; axis labels represent the percentage of total variance explained by the PCs. Point labels indicate pool designations from Stocks et al. (2019) for a subset of pools (i.e. those outside the main dense cluster).



1011

1012 Extended Data Figure 7 - GWAS using SV allele frequencies estimated from poolseq

1013 data

1014 A) Manhattan plot for SVs called from poolseq data from Stocks et al. 2019, identifying

1015 association with ash dieback disease damage. Y axis values represent the -$\log_{10}$ p-values of a

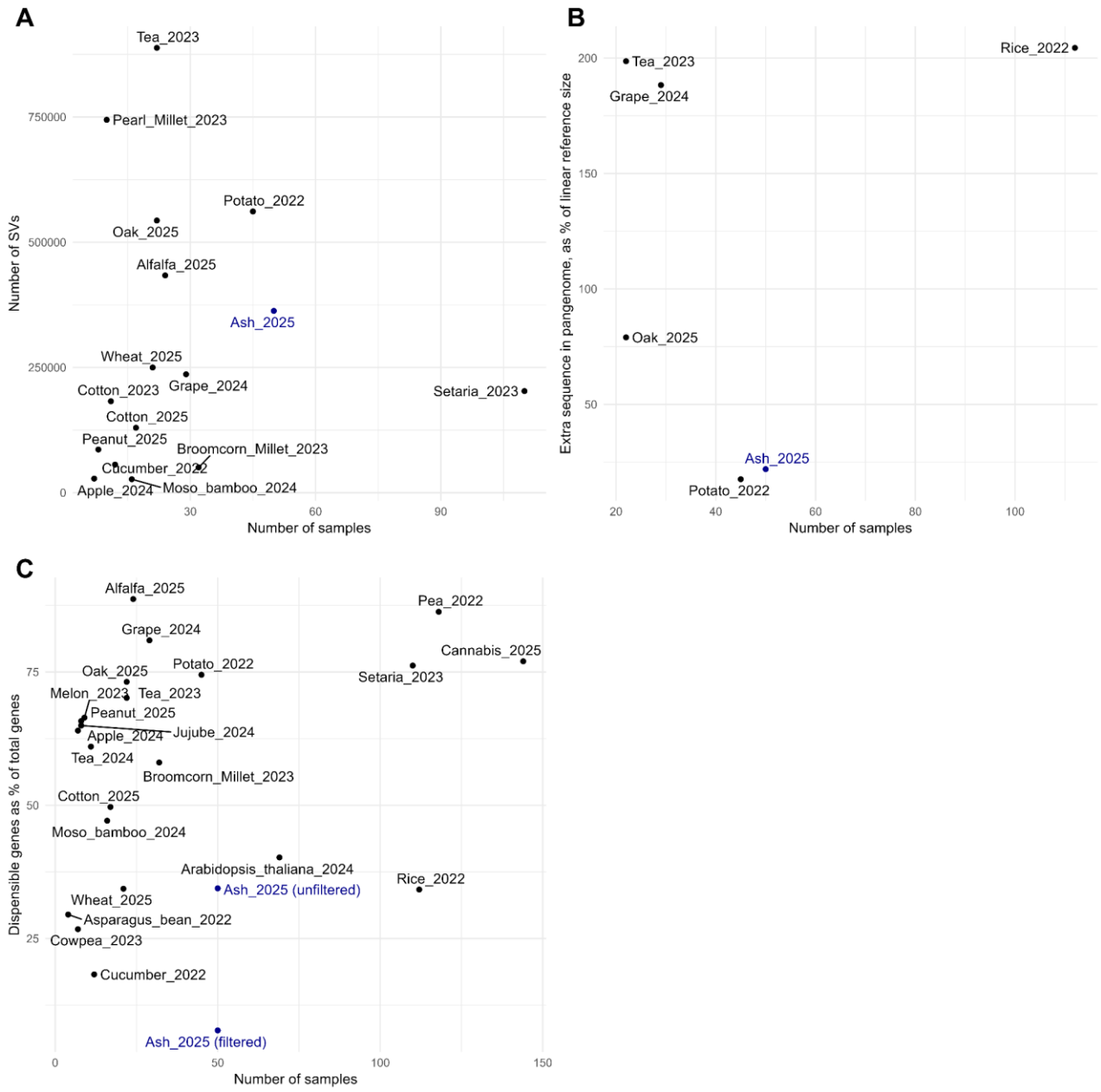
1016 Cochran-Mantel Hahn (CMH) test. Sites are ordered by position along scaffolds in the BATG-

1017 1.0 genome assembly, with alternating black and blue colours representing the different

1018 scaffolds. Scaffolds are listed in order of size, with the largest on the right and smallest on the
1019 left. B) For SVs identified as significant in the CMH test, heatmap of alternate allele frequencies
1020 estimated from individual calls (X axis) from a subset of individually sequenced ash trees from
1021 Stocks et al. vs. an allele frequency estimate from *in silico* pools generated by randomly down
1022 sampling reads for these individuals and measuring the allelic depth (AD; Y axis). Colour scale
1023 represents the number of sites in a square.

Extended Data Figure 8 - Read depths across chromosome 15 outlier region

Total read depths of reads for each of the 15 pairs of pools, surrounding the region with overlapping significant CMH SNPs and F_{st} outlier windows on Chromosome 15. Green points represent depths in the healthy pools; black the unhealthy pools. Biological replicates from NSZ-107 and NSZ-106 are plotted separately.



Extended Data Figure 9 - Comparison of key statistics with other recent pangenome studies.

Results from the present study compared with a selection of plant pangenome studies published since 2022; pangenome studies using primarily short-read data, or incorporating multiple species, were excluded. Labels correspond to study organism and publication date, detailed in

Supplementary Table 6. Number of samples included in the pangenome is on the X axis. Y axis numbers correspond to total number of SVs discovered (A), additional sequence included in the pangenome, as a percentage of the reference genome size of that species (B), or the number dispensable genes as a proportion of the total number of genes in the pangenome. Results from the present study are in blue; in panel C, Ash_2025 (unfiltered) refers to the proportion of dispensable genes prior to filtering, Ash_2025 (filtered) refers to the proportion after filtering for consistent association with SVs - see Results for further details.

1047

1048 Extended Data Figure 10 - Explanatory diagram for the pangenome annotation

1049 Explanatory diagram for annotation of the pangenome - see Methods section “Linear reference
1050 and pangenome annotation” for full description. For the pangenome annotation, the BATG-1.0
1051 linear reference was transformed using the vcf file of SV calls from each of the 50 ONT samples
1052 (1 - S1, S2 and S3 correspond to 3 example samples). For each sample, calls from the vcf used
1053 to transform the BATG-1.0 fasta; where an insertion was called, the consensus insertion
1054 sequence was inserted at the appropriate position, for deletions the relevant sequence from the
1055 BATG-1.0 fasta was deleted. This resulted in a transformed fasta for each sample (2). Using the
1056 RNA-seq data from the reference individual, each of the transformed fasta files was annotated
1057 (3). Based on sequence homology and position in the genome, genes were clustered across
1058 each of the transformed fasta files (4). For each gene cluster, the strength of association of
1059 genes with each SV was assessed (5): genes that were consistently associated with an SV
1060 were kept as variable genes, whereas those not showing a consistent pattern with an
1061 overlapping SV were not.

References

1. Ellegren, H. Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* **29**, 51–63 (2014).
2. Field, M. A. *et al.* Recurrent miscalling of missense variation from short-read genome sequence data. *BMC Genomics* **20**, 546 (2019).
3. Tang, D. *et al.* Genome evolution and diversity of wild and cultivated potatoes. *Nature* **606**, 535–541 (2022).
4. Zhang, F. *et al.* Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. *Genome Res.* **32**, 853–863 (2022).
5. Chen, S. *et al.* Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant *Camellia sinensis*. *Nat. Plants* **9**, 1986–1999 (2023).
6. Liu, Z. *et al.* Grapevine pangenome facilitates trait genetics and genomic breeding. *Nat. Genet.* **56**, 2804–2814 (2024).
7. Liang, Y.-Y. *et al.* Pan-genome analysis reveals local adaptation to climate driven by introgression in oak species. *Mol. Biol. Evol.* **42**, msaf088 (2025).
8. Lynch, R. C. *et al.* Domesticated cannabinoid synthases amid a wild mosaic cannabis pangenome. *Nature* 1–10 (2025).
9. He, F. *et al.* Pan-genomic analysis highlights genes associated with agronomic traits and enhances genomics-assisted breeding in alfalfa. *Nat. Genet.* **57**, 1262–1273 (2025).
10. Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J. & Edwards, D. Plant pan-genomes are the new reference. *Nat Plants* **6**, 914–920 (2020).
11. Crysnanto, D. & Pausch, H. Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. *Genome Biol.* **21**, 184 (2020).
12. Bayer, P. E. *et al.* Modelling of gene loss propensity in the pangenomes of three Brassica

1087 species suggests different mechanisms between polyploids and diploids. *Plant Biotechnol.*
1088 *J.* **19**, 2488–2500 (2021).
1089 13. Mackenzie, A. *et al.* Wheat stripe rust resistance locus YR63 is a hot spot for evolution of
1090 defence genes - a pangenome discovery. *BMC Plant Biol.* **23**, 590 (2023).
1091 14. Zhou, H. *et al.* Pan-genome and transcriptome analyses provide insights into genomic
1092 variation and differential gene expression profiles related to disease resistance and fatty
1093 acid biosynthesis in eastern black walnut (*Juglans nigra*). *Hortic. Res.* **10**, uhad015 (2023).
1094 15. Su, Y. *et al.* Phased telomere-to-telomere reference genome and pangenome reveal an
1095 expansion of resistance genes during apple domestication. *Plant Physiol.* **195**, 2799–2814
1096 (2024).
1097 16. Lovell, J. T. *et al.* Four chromosome scale genomes and a pan-genome annotation to
1098 accelerate pecan tree breeding. *Nat. Commun.* **12**, 4125 (2021).
1099 17. Hämälä, T. *et al.* Genomic structural variants constrain and facilitate adaptation in natural
1100 populations of *Theobroma cacao*, the chocolate tree. *Proc. Natl. Acad. Sci. U. S. A.* **118**,
1101 (2021).
1102 18. Zhou, Y. *et al.* Graph pangenome captures missing heritability and empowers tomato
1103 breeding. *Nature* **606**, 527–534 (2022).
1104 19. Khan, A. W. *et al. Cicer* super-pangenome provides insights into species evolution and
1105 agronomic trait loci for crop improvement in chickpea. *Nat. Genet.* **56**, 1225–1234 (2024).
1106 20. Huang, Y. *et al.* Pangenome analysis provides insight into the evolution of the orange
1107 subfamily and a key gene for citric acid accumulation in citrus fruits. *Nat. Genet.* **55**, 1964–
1108 1975 (2023).
1109 21. Yan, H. *et al.* Pangenomic analysis identifies structural variation associated with heat
1110 tolerance in pearl millet. *Nat. Genet.* **55**, 507–518 (2023).
1111 22. Lei, L. *et al.* Plant Pan-Genomics Comes of Age. *Annu. Rev. Plant Biol.* **72**, 411–435
1112 (2021).

23. Gross, A., Holdenrieder, O., Pautasso, M., Queloz, V. & Sieber, T. N. *Hymenoscyphus pseudoalbidus*, the causal agent of European ash dieback: H. pseudoalbidus, the causal agent of ash dieback. *Mol. Plant Pathol.* **15**, 5–21 (2014).
24. Pautasso, M., Aas, G., Queloz, V. & Holdenrieder, O. European ash (*Fraxinus excelsior*) dieback – A conservation biology challenge. *Biol. Conserv.* **158**, 37–49 (2013).
25. Coker, T. L. R. *et al.* Estimating mortality rates of European ash (*Fraxinus excelsior*) under the ash dieback (*Hymenoscyphus fraxineus*) epidemic. *Plants People Planet* **1**, 48–58 (2019).
26. Pacia, A. *et al.* Ash dieback in forests and rural areas—history and predictions. *Forests* **14**, 2151 (2023).
27. George, J.-P., Sanders, T. G. M., Timmermann, V., Potočić, N. & Lang, M. European-wide forest monitoring substantiate the neccessity for a joint conservation strategy to rescue European ash species (Fraxinus spp.). *Sci. Rep.* **12**, 4764 (2022).
28. Hill, L. *et al.* The £15 billion cost of ash dieback in Britain. *Curr. Biol.* **29**, R315–R316 (2019).
29. Mitchell, R. J. *et al.* Cumulative impact assessments of multiple host species loss from plant diseases show disproportionate reductions in associated biodiversity. *J. Ecol.* **110**, 221–231 (2022).
30. Combes, M., Webber, J. & Boddy, L. Current understanding and future prospects for ash dieback disease with a focus on Britain. *Forestry (Lond.)* **97**, 678–691 (2024).
31. Fellenor, J. *et al.* 'I'd like to report a suspicious looking tree': Public concern, public attention and the nature of reporting about ash dieback in the United Kingdom. *Public Underst. Sci.* **28**, 339–356 (2019).
32. Jepson, P. R. & Arakelyan, I. Developing publicly acceptable tree health policy: public perceptions of tree-breeding solutions to ash dieback among interested publics in the UK. *For. Policy Econ.* **80**, 167–177 (2017).

33. Rallapalli, G. *et al.* Lessons from Fraxinus, a crowd-sourced citizen science game in genomics. *Elife* **4**, e07460 (2015).

34. Tomlinson, I. The discovery of ash dieback in the UK: the making of a focusing event. *Env. Polit.* **25**, 709–728 (2016).

35. Stocks, J. J., Buggs, R. J. A. & Lee, S. J. A first assessment of Fraxinus excelsior (common ash) susceptibility to *Hymenoscyphus fraxineus* (ash dieback) throughout the British Isles. *Sci. Rep.* **7**, 16546 (2017).

36. Pliūra, A., Lygis, V., Suchockas, V. & Bartkevičius, E. Performance of twenty four European *Fraxinus excelsior* populations in three Lithuanian progeny trials with a special emphasis on resistance to *Chalara fraxinea*. *Baltic Forestry* **17**, 17–163 (2011).

37. Clark, J. The Living Ash Project - 10 Years On. Preprint at https://www.futuretrees.org/wp-content/uploads/2024/10/clark-october-2024-qjf.pdf (2024).

38. McKinney, L. V., Nielsen, L. R., Hansen, J. K. & Kjær, E. D. Presence of natural genetic resistance in *Fraxinus excelsior* (Oleraceae) to *Chalara fraxinea* (Ascomycota): an emerging infectious disease. *Heredity (Edinb.)* **106**, 788–797 (2011).

39. Seidel, H., Šeho, M. & Fussi, B. Hope for ash conservation and propagation—single individuals can be highly resistant to an invasive pathogen. *J. Plant Dis. Prot. (2006)* **132**, 1–15 (2025).

40. Plumb, W. J. *et al.* The viability of a breeding programme for ash in the British Isles in the face of ash dieback. *Plants People Planet* **2**, 29–40 (2020).

41. Sniezko, R. A. & Koch, J. Breeding trees resistant to insects and diseases: putting theory into application. *Biol. Invasions* **19**, 3377–3400 (2017).

42. Stocks, J. J. *et al.* Genomic basis of European ash tree resistance to ash dieback fungus. *Nat Ecol Evol* **3**, 1686–1696 (2019).

43. Meger, J., Ulaszewski, B., Pałucka, M., Kozioł, C. & Burczyk, J. Genomic prediction of resistance to Hymenoscyphus fraxineus in common ash (*Fraxinus excelsior* L.) populations.

1165 *Evol. Appl.* **17**, e13694 (2024).

1166 44. Doonan, J. M. *et al.* Multiple, single trait GWAS and supervised machine learning reveal the
1167 genetic architecture of *Fraxinus excelsior* tolerance to ash dieback in Europe. *Plant Cell*
1168 *Environ.* **48**, 3793–3809 (2025).

1169 45. Stener, L.-G. Clonal differences in susceptibility to the dieback of Fraxinus excelsior in
1170 southern Sweden. *Scand. J. For. Res.* **28**, 205–216 (2013).

1171 46. Stener, L.-G. Genetic evaluation of damage caused by ash dieback with emphasis on
1172 selection stability over time. *For. Ecol. Manage.* **409**, 584–592 (2018).

1173 47. Metheringham, C. L. *et al.* Rapid polygenic adaptation in a wild population of ash trees
1174 under a novel fungal epidemic. *Science* **388**, 1422–1425 (2025).

1175 48. Szukala, A. *et al.* Limitations of ad hoc genotyping in detecting ash dieback tolerance in
1176 Fraxinus excelsior. *bioRxiv* 2024.10.11.617393 (2024) doi:10.1101/2024.10.11.617393.

1177 49. McKinney, L. V. *et al.* The ash dieback crisis: genetic variation in resistance can prove a
1178 long-term solution. *Plant Pathol.* **63**, 485–499 (2014).

1179 50. Liu, J. N. *et al.* Pan-genome analyses of 11 *Fraxinus* species provide insights into salt
1180 adaptation in ash trees. *Plant Commun.* **0**, 101137 (2024).

1181 51. Sollars, E. S. A. *et al.* Genome sequence and genetic diversity of European ash trees.
1182 *Nature* **541**, 212–216 (2017).

1183 52. Sax, K. & Abbe, E. C. Chromosome numbers and the anatomy of the secondary xylem in
1184 the Oleaceae. *J. Arnold Arbor.* (1932) doi:10.5962/p.185244.

1185 53. Franco Ortega, S. *et al. Fraxinus excelsior* updated long-read genome reveals the
1186 importance of MADS-box genes in tolerance mechanisms against ash dieback. *G3*
1187 *(Bethesda)* **15**, 2024.12.20.629733 (2025).

1188 54. Huff, M. *et al.* A high-quality reference genome for *Fraxinus pennsylvanica* for ash species
1189 restoration and research. *Mol. Ecol. Resour.* **22**, 1284–1302 (2022).

1190 55. Julca, I., Marcet-Houben, M., Vargas, P. & Gabaldón, T. Phylogenomics of the olive tree

1191 (*Olea europaea*) reveals the relative contribution of ancient allo- and autopolyploidization
1192 events. *BMC Biol.* **16**, 15 (2018).
1193 56. Rao, G. *et al.* De novo assembly of a new *Olea europaea* genome accession using
1194 nanopore sequencing. *Hortic Res* **8**, 64 (2021).
1195 57. Jouffroy, O., Saha, S., Mueller, L., Quesneville, H. & Maumus, F. Comprehensive
1196 repeatome annotation reveals strong potential impact of repetitive elements on tomato
1197 ripening. *BMC Genomics* **17**, 624 (2016).
1198 58. Su, X. *et al.* A high-continuity and annotated tomato reference genome. *BMC Genomics* **22**,
1199 898 (2021).
1200 59. Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2.
1201 *Nat. Biotechnol.* (2024) doi:10.1038/s41587-023-02024-y.
1202 60. Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV.
1203 *Genome Biol.* **21**, 189 (2020).
1204 61. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads
1205 using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
1206 62. Shafin, K. *et al.* Nanopore sequencing and the Shasta toolkit enable efficient de novo
1207 assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020).
1208 63. Hu, J. *et al.* NextDenovo: an efficient error correction and accurate assembly tool for noisy
1209 long reads. *Genome Biol.* **25**, 107 (2024).
1210 64. Landolt, J., Gross, A., Holdenrieder, O. & Pautasso, M. Ash dieback due to
1211 *Hymenoscyphus fraxineus*: what can be learnt from evolutionary ecology? *Plant Pathol.* **65**,
1212 1056–1070 (2016).
1213 65. Kofler, R., Pandey, R. V. & Schlötterer, C. PoPoolation2: identifying differentiation between
1214 populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27**,
1215 3435–3436 (2011).
1216 66. Sirén, J. *et al.* Pangenomics enables genotyping of known structural variants in 5202

1217 diverse genomes. *Science* **374**, abg8871 (2021).

1218 67. Mérot, C. *et al.* Genome assembly, structural variants, and genetic differentiation between
1219 lake whitefish young species pairs (*Coregonus* sp.) with long and short reads. *Mol. Ecol.*
1220 **32**, 1458–1477 (2023).

1221 68. Cochran, W. G. Some methods for strengthening the common χ 2 tests. *Biometrics* **10**, 417
1222 (1954).

1223 69. Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective
1224 studies of disease. *J. Natl. Cancer Inst.* **22**, 719–748 (1959).

1225 70. Wiberg, R. A. W., Gaggiotti, O. E., Morrissey, M. B. & Ritchie, M. G. Identifying consistent
1226 allele frequency differences in studies of stratified populations. *Methods Ecol. Evol.* **8**,
1227 1899–1909 (2017).

1228 71. Micheletti, S. J. & Narum, S. R. Utility of pooled sequencing for association mapping in
1229 nonmodel organisms. *Mol. Ecol. Resour.* **18**, 825–837 (2018).

1230 72. Gazda, M. A. *et al.* A genetic mechanism for sexual dichromatism in birds. *Science* **368**,
1231 1270–1274 (2020).

1232 73. Brennan, R. S. *et al.* Experimental evolution reveals the synergistic genomic mechanisms
1233 of adaptation to ocean warming and acidification in a marine copepod. *Proc. Natl. Acad.*
1234 *Sci. U. S. A.* **119**, e2201521119 (2022).

1235 74. Audet, T., Krol, J., Pelletier, K., Stewart, A. D. & Dworkin, I. Sexually discordant selection is
1236 associated with trait-specific morphological changes and a complex genomic response.
1237 *Evolution* **78**, 1426–1440 (2024).

1238 75. Ou, L. *et al.* Pan-genome of cultivated pepper (*Capsicum*) and its use in gene presence-
1239 absence variation analyses. *New Phytol.* **220**, 360–363 (2018).

1240 76. Barchi, L. *et al.* Improved genome assembly and pan-genome provide key insights into
1241 eggplant domestication and breeding. *Plant J.* **107**, 579–596 (2021).

1242 77. Li, H. *et al.* Graph-based pan-genome reveals structural and sequence variations related to

1243 agronomic traits and domestication in cucumber. *Nat. Commun.* **13**, 682 (2022).

1244 78. Liang, L. *et al.* Genome and pan-genome assembly of asparagus bean (*Vigna unguiculata*
1245 *ssp. sesquipedialis*) reveal the genetic basis of cold adaptation. *Front. Plant Sci.* **13**,
1246 1059804 (2022).

1247 79. He, Q. *et al.* A graph-based genome and pan-genome variation of the model plant *Setaria*.
1248 *Nat. Genet.* **55**, 1232–1242 (2023).

1249 80. Chen, J. *et al.* Pangenome analysis reveals genomic variations associated with
1250 domestication traits in broomcorn millet. *Nat. Genet.* **55**, 2243–2254 (2023).

1251 81. Lyu, X. *et al.* Pan-genome analysis sheds light on structural variation-based dissection of
1252 agronomic traits in melon crops. *Plant Physiol.* **193**, 1330–1348 (2023).

1253 82. He, X. *et al.* Pangenome analysis reveals transposon-driven genome evolution in cotton.
1254 *BMC Biol.* **22**, 92 (2024).

1255 83. Guo, M. *et al.* Analyzes of pan-genome and resequencing atlas unveil the genetic basis of
1256 jujube domestication. *Nat. Commun.* **15**, 9320 (2024).

1257 84. Fang, Y. *et al.* Pan-genome and phylogenomic analyses highlight *Hevea* species
1258 delineation and rubber trait evolution. *Nat. Commun.* **15**, 7232 (2024).

1259 85. Lian, Q. *et al.* A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved
1260 genome structure throughout the global species range. *Nat. Genet.* **56**, 982–991 (2024).

1261 86. Guo, N. *et al.* A graph-based pan-genome of *Brassica oleracea* provides new insights into
1262 its domestication and morphotype diversification. *Plant Commun.* **5**, 100791 (2024).

1263 87. Liang, Q. *et al.* A view of the pan-genome of domesticated Cowpea (*Vigna unguiculata* [L.]
1264 Walp.). *Plant Genome* **17**, e20319 (2024).

1265 88. Glick, L. & Mayrose, I. The effect of methodological considerations on the construction of
1266 gene-based plant pan-genomes. *Genome Biol. Evol.* **15**, evad121 (2023).

1267 89. Zhao, K. *et al.* Pangenome analysis reveals structural variation associated with seed size
1268 and weight traits in peanut. *Nat. Genet.* **57**, 1250–1261 (2025).

1269 90. Hou, Y. *et al.* Haplotype-based pangenomes reveal genetic variations and climate
1270 adaptations in moso bamboo populations. *Nat. Commun.* **15**, 8085 (2024).
1271 91. Meng, Q. *et al.* Pangenome analysis reveals yield- and fiber-related diversity and
1272 interspecific gene flow in *Gossypium barbadense* L. *Nat. Commun.* **16**, 4995 (2025).
1273 92. Ebler, J. *et al.* Pangenome-based genome inference allows efficient and accurate
1274 genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
1275 93. Eggertsson, H. P. *et al.* GraphTyper2 enables population-scale genotyping of structural
1276 variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).
1277 94. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome Biol.* **20**,
1278 246 (2019).
1279 95. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo
1280 assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
1281 96. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for
1282 long-range linkage. *Genome Res.* **26**, 342–350 (2016).
1283 97. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
1284 transform. *Bioinformatics* **25**, 1754–1760 (2009).
1285 98. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for
1286 genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
1287 99. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update:
1288 Novel and streamlined workflows along with broader and deeper phylogenetic coverage for
1289 scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654
1290 (2021).
1291 100. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
1292 (2009).
1293 101. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–
1294 3100 (2018).

1295 102. Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. BlobToolKit - interactive
1296 quality assessment of genome assemblies. *G3 (Bethesda)* **10**, 1361–1374 (2020).
1297 103. Zhou, C. *et al.* Oatk: a de novo assembly tool for complex plant organelle genomes. *bioRxiv*
1298 2024.10.23.619857 (2024) doi:10.1101/2024.10.23.619857.
1299 104. Tillich, M. *et al.* GeSeq - versatile and accurate annotation of organelle genomes. *Nucleic*
1300 *Acids Res.* **45**, W6–W11 (2017).
1301 105. Using local stock for planting native trees and shrubs. *Forest Research*
1302 https://www.forestresearch.gov.uk/publications/using-local-stock-for-planting-native-trees-
1303 and-shrubs/ (1999).
1304 106. Caudullo, G., Welk, E. & San-Miguel-Ayanz, J. Chorological maps for the main European
1305 woody species. *Data Brief* **12**, 662–666 (2017).
1306 107. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
1307 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
1308 108. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack:
1309 visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669
1310 (2018).
1311 109. De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read
1312 sequencing data. *Bioinformatics* **39**, btad311 (2023).
1313 110. Andrews, S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput
1314 Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
1315 111. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for
1316 multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
1317 112. Zheng, Z. *et al.* Symphonizing pileup and full-alignment for deep learning-based long-read
1318 variant calling. *Nat Comput Sci* **2**, 797–803 (2022).
1319 113. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
1320 114. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer

1321 datasets. *Gigascience* **4**, 7 (2015).
1322 115. Zhou, A., Lin, T. & Xing, J. Evaluating nanopore sequencing data processing pipelines for
1323 structural variation identification. *Genome Biol.* **20**, 237 (2019).
1324 116. De Coster, W. *et al.* Structural variants identified by Oxford Nanopore PromethION
1325 sequencing of the human genome. *Genome Res.* **29**, 1178–1187 (2019).
1326 117. Bolognini, D. & Magi, A. Evaluation of Germline Structural Variant Calling Methods for
1327 Nanopore Sequencing Data. *Front. Genet.* **12**, 761791 (2021).
1328 118. Yildiz, G. *et al.* Benchmarking Oxford Nanopore Read Alignment-Based Structural Variant
1329 Detection Tools in Crop Plant Genomes. *bioRxiv* 2022.09.23.508909 (2022)
1330 doi:10.1101/2022.09.23.508909.
1331 119. Jiang, T. *et al.* Long-read sequencing settings for efficient structural variation detection
1332 based on comprehensive evaluation. *BMC Bioinformatics* **22**, 552 (2021).
1333 120. Duan, X., Pan, M. & Fan, S. Comprehensive evaluation of structural variant genotyping
1334 methods based on long-read sequencing data. *BMC Genomics* **23**, 324 (2022).
1335 121. Liu, Y. *et al.* Comparison of structural variants detected by PacBio-CLR and ONT
1336 sequencing in pear. *BMC Genomics* **23**, 830 (2022).
1337 122. Dierckxsens, N., Li, T., Vermeesch, J. R. & Xie, Z. A benchmark of structural variation
1338 detection by long reads through a realistic simulated model. *Genome Biol.* **22**, 342 (2021).
1339 123. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid
1340 genome assemblies. *Bioinformatics* **36**, 5519–5521 (2021).
1341 124. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits
1342 and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
1343 125. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
1344 features. *Bioinformatics* **26**, 841–842 (2010).
1345 126. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable
1346 element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).

1347 127. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in
1348 genomic sequences. *Curr. Protoc. Bioinformatics* **Chapter 4**, Unit 4.10 (2009).
1349 128. Formenti, G. *et al.* Gfastats: conversion, evaluation and manipulation of genome sequences
1350 using assembly graphs. *Bioinformatics* **38**, 4214–4216 (2022).
1351 129. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
1352 130. Lovell, J. T. *et al.* The genomic landscape of molecular responses to natural drought stress
1353 in *Panicum hallii*. *Nat. Commun.* **9**, 5213 (2018).
1354 131. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-Seq and
1355 protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *bioRxiv*
1356 doi:10.1101/2023.06.10.544449.
1357 132. Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of
1358 organismal diversity. *Nucleic Acids Res.* **51**, D445–D451 (2023).
1359 133. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative
1360 genomics. *Genome Biol.* **20**, 238 (2019).
1361 134. Unver, T. *et al.* Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad.*
1362 *Sci. U. S. A.* **114**, E9413–E9422 (2017).
1363 135. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy
1364 fruit evolution. *Nature* **485**, 635–641 (2012).
1365 136. Hellsten, U. *et al.* Fine-scale variation in meiotic recombination in *Mimulus* inferred from
1366 population shotgun sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19478–19482 (2013).
1367 137. Wang, L. *et al.* Genome sequencing of the high oil crop sesame provides insight into oil
1368 biosynthesis. *Genome Biol.* **15**, R39 (2014).
1369 138. Pan, X. *et al.* Chromosome-level genome assembly of *Salvia miltiorrhiza* with orange roots
1370 uncovers the role of Sm2OGD3 in catalyzing 15,16-dehydrogenation of tanshinones. *Hortic.*
1371 *Res.* **10**, uhad069 (2023).
1372 139. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale

using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).

140. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

141. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. EggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).

142. Alexa A, R. J. *topGO: Enrichment Analysis for Gene Ontology*. http://bioconductor.org/packages/topGO/ (2024).

143. Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S. & Prins, P. A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Comput. Biol.* **18**, e1009123 (2022).

144. 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).

145. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**, e0163962 (2016).

146. R Core Team. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria*. (2023).

147. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

148. vcd: Visualizing Categorical Data. https://cran.r-universe.dev/vcd.