



Deposited via The University of York.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/112654/>

Version: Accepted Version

Article:

Sollars, Elizabeth, Harper, Andrea Louise, Kelly, Laura et al. (2017) Genome sequence and genetic diversity of European ash trees. *Nature*. pp. 212-216. ISSN: 0028-0836

<https://doi.org/10.1038/nature20786>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Genome sequence and genetic diversity of European ash trees

Elizabeth SA Sollars, Andrea L Harper, Laura J Kelly, Christine Sambles, Ricardo H Ramirez-Gonzalez, David Swarbreck, Gemy Kaithakottil, Endymion D Cooper, Cristobal Uauy, Lenka Havlickova, Gemma Worswick, David J Studholme, Jasmin Zohren, Deborah L Salmon, Bernardo J Clavijo, Yi Li, Zhesi He, Alison Fellgett, Lea Vig McKinney, Lene Rostgaard Nielsen, Gerry C Douglas, Erik Dahl Kjær, J. Allan Downie, David Boshier, Steve Lee, Jo Clark, Murray Grant, Ian Bancroft, Mario Caccamo, Richard JA Buggs

Ash trees (genus *Fraxinus*, Oleaceae) are widespread throughout the Northern Hemisphere, but are being devastated in Europe by the fungus *Hymenoscyphus fraxineus*, causing ash dieback (ADB), and in North America by the Emerald Ash Borer (EAB), *Agrilus planipennis*^{1,2}. We sequenced the genome of a low-heterozygosity *Fraxinus excelsior* tree from Britain annotating 38,852 protein-coding genes of which 25% appear ash specific when compared with ten other plant species' genomes. Analyses of paralogous genes suggest a whole-genome duplication shared with olive (*Olea europaea*, Oleaceae). We resequenced 37 *F. excelsior* trees from Europe finding evidence for apparent long-term decline in effective population size. Using our reference sequence, we re-analysed association transcriptomic data³, yielding improved markers for reduced susceptibility to ADB. Surveys of these markers in British populations suggested that reduced ADB susceptibility may be more widespread in Great Britain than in Denmark. We also present evidence that susceptibility of trees to *H. fraxineus* is associated with their iridoid glycoside levels. This rapid, integrated, multidisciplinary research response to an emerging health threat in a non-model organism opens the way for mitigation of the epidemic.

We sequenced an European ash (*F. excelsior*) tree generated from self-pollination of a woodland tree in Gloucestershire, UK. The sequenced tree (Earth Trust accession 2451S) appeared free of ADB when sampled in 2013 and 2014, but showed symptoms in February 2016. Its genome size was measured by flow cytometry as 877.24 ± 1.41 Mbp. Total genomic DNA was sequenced to 192X coverage (see Supplementary Table 1). We assembled the genome into 89,514 nuclear scaffolds with an N₅₀ of 104 kbp, 26 mitochondrial scaffolds, and one plastid chromosome (Supplementary Tables 2-3), where the non-N assembly comprises 80.5% of the predicted genome size. RepeatMasker estimated 35.90% of the assembly to be repetitive elements, with LTR retrotransposons predominating (Supplementary Table 4). In comparison with other eudicot genomes of similar size^{4,5} this repeat content is low. The 17% of the assembly comprised of Ns likely contains additional repeats; 27% of reads that do not map to the assembly align to ash repeats (Supplementary Table 5). We generated ~160 million RNA-Seq read pairs from tree 2451S leaf tissue and from leaf, cambium, root, and flower tissue of its parent tree (Supplementary Table 6); low expression of repetitive elements was found in all tissues (Supplementary Table 7).

We annotated the genome using an evidence based workflow incorporating protein and RNA-Seq data, predicting 38,852 protein-coding genes and 50,743 transcripts (Supplementary Table 4). This gene count is within 12% that of tomato (v2.3)⁴, potato (v3.4)⁶ and hot pepper (v1.5)⁷ but higher than monkey flower (v2.0; 26,718 genes)⁸. Evidence for completeness and coherence of our models is shown in Extended Data Fig. 1. Of 38,852 predicted genes 97.67% (and 98.18% of transcripts) were supported by ash RNA-Seq data,

52 81.80% showed high similarity to plant proteins (> 50% high-scoring segment pair coverage)
53 (Supplementary Table 8), 97.05% had matches in the non-redundant (nr) databases
54 (excluding hits to ash), 82.74% generated hits to InterPro signatures, and 78.09% were
55 assigned Gene Ontology (GO) terms. We also identified 107 microRNA (miRNA), 792 tRNA
56 and 51 rRNA genes.

57
58 Past whole genome duplication (WGD) events are commonly inferred from the distributions
59 of pairwise synonymous site divergence (K_s) within paralogous gene groups⁹. We plotted
60 these for ash and six other plant species (Fig. 1a, Supplementary Table 9). Ash and olive
61 shared a peak near $K_s = 0.25$, suggesting an Oleaceae-specific WGD. A peak near $K_s =$
62 0.6 shared by ash, olive, monkey flower and tomato but not by bladderwort, coffee and
63 grape does not fit a common origin hypothesis, unless bladderwort has an accelerated
64 substitution rate and the tomato peak is not restricted to the Solanales as evidenced
65 previously⁴. Synteny analysis between ash and monkey flower did not provide conclusive
66 evidence for shared WGD (Extended Data Fig. 2). Duplicated genes in the ash genome that
67 were not locally duplicated (i.e. within 10 genes of each other in our assembly) show no
68 significantly enriched GO terms at an FDR level of 0.05. In contrast 1,005 locally duplicated
69 genes showed significant enrichment of terms relating to oxidoreductase, catalytic and
70 monooxygenase activity compared to all other genes, suggesting evolution of secondary
71 metabolism by local duplications.

72
73 We analysed gene families shared between ash and 10 other species (Supplementary Table
74 10). In total, 279,603 proteins (77.14% of the input sequences) clustered into 27,222 groups,
75 of which 4,292 contained sequences from all species, 3,266 were angiosperm-specific and
76 462 Eudicot-specific. Patterns of gene-family sharing among Asterids and among woody
77 species are shown in Figures 1b and c. For 38,852 ash proteins, 30,802 clustered into
78 14,099 groups, of which 643 were ash-specific, containing 1,554 proteins. There were also
79 8,050 singleton proteins unique to ash. Of the 9,604 ash-specific proteins, 6,405 matched ≥ 1
80 InterPro signature. The 20 largest groups in ash are listed in Extended Data Table 1: several
81 are putatively associated with disease resistance.

82
83 To investigate genomic diversity in *F. excelsior*, we sequenced 37 ash trees from central,
84 northern and western Europe (Fig. 2 and Supplementary Table 11), to an average of 8.4X
85 genome coverage by trimmed and filtered reads. Together with reads from Danish 'Tree35'
86 (<http://oadb.tsl.ac.uk/>), these were mapped to the reference genome. We found 12.48M
87 polymorphic sites with a variant of high confidence in at least one individual (qual > 300
88 using FreeBayes¹⁰): we refer to these as the 'genome-wide SNP set' in the 'European
89 Diversity Panel'. Of these, 6.85M (54.88%) occur inside or within 5kbp of genes
90 (Supplementary Table 12). We found 259,946 amino-acid substitutions and 71,513 variants
91 that affect stop or start codons, or splice sites. We selected 23 amino-acid variants, and 26
92 non-coding variants with a range of call qualities for validation using KASP: individual
93 genotype calls with quality > 300 have a false positive rate of 6% and those with quality >
94 1000 have a false positive rate of zero (Supplementary Table 13). We ran a more stringent
95 variant calling restricted to regions of the genome with between 5 and 30X coverage in all 38
96 samples. These totalled 20.6 Mbp (2.3% of the genome), within which 529,812 variants were
97 called with CLC Genomics Workbench. Of these, 394,885 were biallelic SNPs with minimum
98 allele frequency above 0.05, which we refer to as the 'reduced SNP set'. We also found c.
99 31,300 singleton simple sequence repeat (SSR) loci in the ash genome, and designed
100 primers for 664 (Supplementary Data 1). In a sample of 366 of these, 48% were polymorphic
101 in the European Diversity Panel sequences. We PCR tested 48 of these in multiplexes with
102 European Diversity Panel genomic DNA and found that 41 amplified successfully
103 (Supplementary Data 1).

105 We analysed population structure of the European Diversity Panel using: a plastid haplotype
106 network, STRUCTURE¹¹ runs on genomic SNPs and principal components analysis of the
107 'reduced SNP set' (Fig. 2a-d, Extended Data Fig. 3). Clearest differentiation was found in the
108 plastid network, with four distinct haplotype groups each separated from each other by at
109 least 20 substitutions. One group was more frequent in Great Britain than on the continent.
110 The second and third principal components of the PCA corresponded with the plastid data
111 somewhat (Fig. 2c). Previous analyses of SSRs in plastids identified variants unique to the
112 British Isles and Iberia¹². Linkage disequilibrium (LD) in the European Diversity Panel
113 decayed logarithmically, with an average r^2 of 0.15 at 100 bp between SNPs, reaching an r^2
114 0.05 at ~40 kbp (Fig. 2e). This is similar to long-range LD estimates found in *Populus*
115 *tremuloides*¹³. Apparent long-term effective population size decline of *F. excelsior* in Europe
116 was shown by analyses based on heterozygosity in the reference genome (using PSMC¹⁴
117 Fig. 2f). Such patterns may also reflect a complex history of population subdivision in ash¹⁵.

118
119 We used associative transcriptomics to predict ADB damage in Great Britain. We used the
120 full CDS models from our genome annotation as a mapping reference for previously
121 generated³ RNA-Seq reads from 182 Danish ash accessions ('Danish Scored Panel') that
122 have been exposed to *H. fraxineus*, and scored for damage (Supplementary Data 2). This
123 yielded 40,133 gene expression markers (GEMs; Supplementary Data 3) and 394,006 SNPs
124 (Supplementary Data 4). Twenty GEMs were associated with ADB damage scores, including
125 eight MADS-box proteins, and two cinnamoyl-CoA reductase 2 genes that may be involved
126 in the hypersensitive response (Supplementary Data 5). Four assays representing the top
127 five GEMs were applied to 58 Danish accessions ('Danish Test Panel') to validate the top
128 markers. Results were combined into a single predicted damage score for each tree
129 (Supplementary Data 6), which was compared to the observed damage scores (Fig. 3;
130 $R^2=0.25$, $P=6.9 \times 10^{-5}$): predictions of damage < 50% consistently detected trees with very
131 low observed damage scores. The same assays were also applied to 130 accessions from
132 across the British range of *F. excelsior* ('British Screening Panel'; Supplementary Data 6).
133 Strikingly, this provided lower predictions for ADB damage in the British Screening Panel:
134 25% were predicted to have <25% canopy damage, compared to 9% of the Danish Test
135 Panel. Trees with low predicted damage are scattered throughout Britain (Fig. 3).

136
137 We also examined expression of the top five GEM loci using RPKM values from our shotgun
138 Illumina read data for the reference tree (Extended Data Fig. 4), comparing these with
139 RPKM values from the Danish Scoring Panel. Expression patterns in the reference tree were
140 highly correlated with those of the most susceptible Danish quartile ($R^2=0.995$, $p<0.001$), but
141 not the least susceptible ($p=0.24$), consistent with observations that the reference tree is
142 now succumbing to the disease. We correlated the expression of all 20 top GEM markers in
143 leaf, flower, cambium and root transcriptomes of the parent of the reference tree. This
144 revealed that leaf expression levels were positively correlated with those in the cambium
145 ($R^2=0.65$, $p<0.001$) and flower ($R^2=0.38$, $p=0.0041$), but not with the root ($p=0.3594$).

146
147 We identified putative orthologues of the five GEM loci using our OrthoMCL results
148 (Supplementary Data 5) and BLAST searches of GenBank, and conducted maximum
149 likelihood and Bayesian analyses of relevant hits (Extended Data Fig. 5).
150 FRAEX38873_v2_000173540.4, FRAEX38873_v2_000048340.1 and
151 FRAEX38873_v2_000048360.1 clustered into the SVP/StMADS11 group¹⁶ of type II MADS-
152 box genes. FRAEX38873_v2_000261470.1 and FRAEX38873_v2_000199610.1 clustered
153 into the SOC1/TM3 group of type II MADS-box proteins^{16,17}. Both groups have roles in flower
154 development¹⁸⁻²¹, and appear to be involved in stress response in *Brassica rapa*²². Many
155 genes involved in regulation of flowering time in *A. thaliana* are involved in controlling
156 phenology in perennial trees species²³ and genes belonging to the SVP/StMADS11 clade

157 have potential roles in growth cessation, bud set and dormancy²³. In *A. thaliana*, *AGL22/SVP*
158 may be required for age-related resistance (ARR)²⁴.

159

160 One mechanism by which transcriptional cascades, such as those involving MADS box
161 genes, might be involved in tolerance or resistance to pathogens is via modulation of
162 secondary metabolite concentrations. For five high-susceptibility and five low-susceptibility
163 Danish trees, we profiled methanol-extracted leaf samples by liquid chromatography mass
164 spectrometry on a quadrupole time-of-flight mass spectrometer. Partial Least Squares
165 Discriminant Analysis (PLS-DA) clearly discriminated high and low susceptibility trees (Fig.
166 4a). By using accurate mass to identify the chemical nature of discriminant features, we
167 found greater abundance (Fig. 4b) of iridoid glycosides (for details see Extended Data
168 Figures 6-9, and Supplementary Data 9) in high ADB susceptibility genotypes than in low
169 susceptibility genotypes. A MS/MS fragmentation network identified a number of product
170 ions expected from fragmentation of iridoid glycosides (Fig. 4c). Iridoid glycosides are a well-
171 known anti-herbivore defense mechanism in the Oleaceae²⁵⁻²⁷. They can also enhance
172 fungal growth *in vitro*²⁸, although their aglycone hydrolysis product formed following tissue
173 damage can also mediate fungal resistance²⁹. Our data suggest there may be a trade-off
174 between ADB susceptibility and herbivore susceptibility. This is of particular concern given
175 the threat of the herbivore EAB to ash in both North America¹ and Europe³⁰ and may hamper
176 efforts to breed trees with low susceptibility to both threats.

177

178

179

180

181 Literature cited

- 182 1. Poland, T. M. & McCullough, D. G. Emerald ash borer: invasion of the urban forest and
183 the threat to North America's ash resource. *J. For.* **104**, 118–124 (2006).
- 184 2. Pautasso, M., Aas, G., Queloz, V. & Holdenrieder, O. European ash (*Fraxinus excelsior*)
185 dieback – A conservation biology challenge. *Biol. Conserv.* **158**, 37–49 (2013).
- 186 3. Harper, A. L. *et al.* Molecular markers for tolerance of European ash (*Fraxinus*
187 *excelsior*) to dieback disease identified using Associative Transcriptomics. *Sci. Rep.* **6**,
188 19335 (2016).
- 189 4. Tomato Genome Consortium. The tomato genome sequence provides insights into
190 fleshy fruit evolution. *Nature* **485**, 635–641 (2012).
- 191 5. Ming, R. *et al.* Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.).
192 *Genome Biol.* **14**, R41 (2013).
- 193 6. Potato Genome Sequencing Consortium *et al.* Genome sequence and analysis of the
194 tuber crop potato. *Nature* **475**, 189–195 (2011).
- 195 7. Kim, S. *et al.* Genome sequence of the hot pepper provides insights into the evolution of
4

- 196 pungency in *Capsicum* species. *Nat. Genet.* **46**, 270–278 (2014).
- 197 8. Hellsten, U. *et al.* Fine-scale variation in meiotic recombination in *Mimulus* inferred from
198 population shotgun sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19478–19482
199 (2013).
- 200 9. Blanc, G. & Wolfe, K. H. Widespread paleopolyploidy in model plant species inferred
201 from age distributions of duplicate genes. *Plant Cell* **16**, 1667–1678 (2004).
- 202 10. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing.
203 *arXiv [q-bio.GN]* (2012).
- 204 11. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using
205 multilocus genotype data. *Genetics* **155**, 945–959 (2000).
- 206 12. Heuertz, M. *et al.* Chloroplast DNA phylogeography of European ashes, *Fraxinus* sp.
207 (Oleaceae): roles of hybridization and life history traits. *Mol. Ecol.* **15**, 2131–2140
208 (2006).
- 209 13. Wang, J., Street, N. R., Scofield, D. G. & Ingvarsson, P. K. Natural Selection and
210 Recombination Rate Variation Shape Nucleotide Polymorphism Across the Genomes of
211 Three Related *Populus* Species. *Genetics* **202**, 1185–1200 (2016).
- 212 14. Li, H. & Durbin, R. Inference of human population history from individual whole-genome
213 sequences. *Nature* **475**, 493–496 (2011).
- 214 15. Mazet, O., Rodríguez, W., Grusea, S., Boitard, S. & Chikhi, L. On the importance of
215 being structured: instantaneous coalescence rates and human evolution-lessons for
216 ancestral population size inference? *Heredity* (2015). doi:10.1038/hdy.2015.104
- 217 16. Smaczniak, C., Immink, R. G. H., Angenent, G. C. & Kaufmann, K. Developmental and
218 evolutionary diversity of plant MADS-domain factors: insights from recent studies.
219 *Development* **139**, 3081–3098 (2012).
- 220 17. Wells, C. E., Vendramin, E., Jimenez Tarodo, S., Verde, I. & Bielenberg, D. G. A
221 genome-wide analysis of MADS-box genes in peach [*Prunus persica* (L.) Batsch]. *BMC*
222 *Plant Biol.* **15**, 41 (2015).

- 223 18. Liu, C. *et al.* Direct interaction of AGL24 and SOC1 integrates flowering signals in
224 *Arabidopsis*. *Development* **135**, 1481–1491 (2008).
- 225 19. Li, D. *et al.* A repressor complex governs the integration of flowering signals in
226 *Arabidopsis*. *Dev. Cell* **15**, 110–120 (2008).
- 227 20. Dorca-Fornell, C. *et al.* The *Arabidopsis* SOC1-like genes AGL42, AGL71 and AGL72
228 promote flowering in the shoot apical and axillary meristems. *Plant J.* **67**, 1006–1017
229 (2011).
- 230 21. Gregis, V., Sessa, A., Colombo, L. & Kater, M. M. AGAMOUS-LIKE24 and SHORT
231 VEGETATIVE PHASE determine floral meristem identity in *Arabidopsis*. *Plant J.* **56**,
232 891–902 (2008).
- 233 22. Saha, G. *et al.* Genome-wide identification and characterization of MADS-box family
234 genes related to organ development and stress resistance in *Brassica rapa*. *BMC*
235 *Genomics* **16**, 178 (2015).
- 236 23. Ding, J. & Nilsson, O. Molecular regulation of phenology in trees — because the
237 seasons they are a-changin'. *Curr. Opin. Plant Biol.* **29**, 73–79 (2016/2).
- 238 24. Wilson, D. C., Carella, P., Isaacs, M. & Cameron, R. K. The floral transition is not the
239 developmental switch that confers competence for the *Arabidopsis* age-related
240 resistance response to *Pseudomonas syringae* pv. *tomato*. *Plant Mol. Biol.* **83**, 235–246
241 (2013).
- 242 25. Jensen, S. R., Franzyk, H. & Wallander, E. Chemotaxonomy of the Oleaceae: iridoids
243 as taxonomic markers. *Phytochemistry* **60**, 213–231 (2002).
- 244 26. Kubo, I., Matsumoto, A. & Takase, I. A multichemical defense mechanism of bitter olive
245 *Olea europaea* (Oleaceae) : Is oleuropein a phytoalexin precursor? *J. Chem. Ecol.* **11**,
246 251–263 (1985).
- 247 27. Eyles, A. *et al.* Comparative phloem chemistry of Manchurian (*Fraxinus mandshurica*)
248 and two North American ash species (*Fraxinus americana* and *Fraxinus pennsylvanica*).
249 *J. Chem. Ecol.* **33**, 1430–1448 (2007).

- 250 28. Marak, H. B., Biere, A. & Van Damme, J. M. M. Systemic, genotype-specific induction of
251 two herbivore-deterrent iridoid glycosides in *Plantago lanceolata* L. in response to fungal
252 infection by *Diaporthe adunca* (Rob.) Niessel. *J. Chem. Ecol.* **28**, 2429–2448 (2002).
- 253 29. Biere, A., Marak, H. B. & van Damme, J. M. M. Plant chemical defense against
254 herbivores and pathogens: generalized defense or trade-offs? *Oecologia* **140**, 430–441
255 (2004).
- 256 30. Valenta, V., Moser, D., Kuttner, M., Peterseil, J. & Essl, F. A High-Resolution Map of
257 Emerald Ash Borer Invasion Risk for Southern Central Europe. *For. Trees Livelihoods* **6**,
258 3075–3086 (2015).

259

260 **Figure Legends**

261

262 **Figure 1 | Gene sharing within and among plant genomes.** **a**, Distribution of Ks values
263 between paralogous gene pairs within the genomes of ash (*Fraxinus excelsior*), tomato
264 (*Solanum lycopersicum*), coffee (*Coffea canephora*), bladderwort (*Utricularia gibba*), grape
265 (*Vitis vinifera*) and monkey flower (*Mimulus guttatus*), and transcriptome of olive (*Olea*
266 *europaea*). **b**, Venn diagram of gene sharing by five Asterid species. **c**, Venn diagram of
267 gene sharing by six woody species. Numbers in parentheses are the total number of
268 OrthoMCL groups found for that species; numbers in intersections show the total number of
269 groups shared between given combinations of taxa.

270

271 **Figure 2 | Genome diversity of *F. excelsior* in Europe.** **a**, Map showing the distribution of
272 plastid haplotypes (n=37), based on a median-joining plastid haplotype network for the
273 European Diversity Panel (inset). **b**, Map showing diversity structure of genomic SNPs,
274 based on average Q-value for each individual (inset), from three runs of STRUCTURE with
275 different sets of 8,955 SNPs and k=3. **c**, Principal component analysis of 34,607 nuclear
276 SNPs in the European Diversity Panel, PC2 plotted against PC3, with points coloured by
277 plastid haplotype. **d**, From the same PCA, PC1 plotted against PC2, with points coloured by
278 groupings found by STRUCTURE using genomic SNPs. **e**, Linkage disequilibrium decay
279 between SNPs in the European Diversity Panel. **f**, Effective population size history estimated
280 using the PSMC method on the reference genome, with 100 bootstraps (shown in light blue).

281

282 **Figure 3 | Predicted ash dieback damage scores in Britain and Denmark.** Map points
283 are scaled by hue (high predicted damage scores in brown, low in green) and plotted
284 according to the geographical origin of the parent trees of the British Screening Panel
285 (n=130) and the Danish Test Panel (n=58). Single leaf samples taken from grafts of each
286 individual tree were used for predicting damage scores. Inset: Damage predictions for the
287 Danish Test Panel (n=58) correlated with log mean observed damage scores from 2013-14
288 ($R^2=0.25$, $P=6.9 \times 10^{-5}$).

289

290 **Figure 4 | Putative iridoid glycosides as discriminatory features between *F. excelsior***
291 **genotypes with differential susceptibility to ADB.** **a**, Multivariate analysis PLS-DA score
292 plot of metabolic profiles of five high and five low susceptibility trees (n=3 per genotype). **b**,

293 Boxplots from these profiles showing normalised (internal standard) intensity (log₂
294 transformed) of five discriminatory features observed in negative mode; m/z and retention
295 time (RT) are given for each feature. **c**, Fragmentation network of discriminatory features,
296 highlighted in black (positive mode) and grey (negative mode). Each product ion is labelled
297 with its size (m/z), also depicted by its circle size. Blue shading increases with the number of
298 times each ion is present in the precursor discriminatory features. Product ions not shared
299 among precursors are shown as unlabelled tips. The edges are in shades of red based on
300 retention time; the paler the colour the earlier the retention time. Those fragment masses
301 shaded in green have been previously reported from fragmentation of iridoid glycosides.

302

303

304 **Methods**

305

306 **Tree Material**

307

308 *Reference tree*: In 2013 twig material was collected from tree 2451S growing at Paradise
309 Wood, Earth Trust, Oxfordshire. This tree was produced via self pollination of an
310 hermaphroditic *F. excelsior* tree growing in woodland in Gloucestershire (Lat. 52.020592,
311 Long. -1.832804), UK, in 2002 as part of the FRAXIGEN project³¹. The parent tree was one
312 of 19 trees that produced seed from self-pollination, and had lower heterozygosity at four
313 microsatellite loci than the other 18 trees (D. Boshier, unpubl. data). DNA was extracted from
314 bud, cambial and wood tissues using CTAB³² and Qiagen DNeasy protocols. RNA was
315 extracted using the Qiagen RNeasy protocol from leaf tissue of tree 2451S and from leaf,
316 cambium, root, and flower tissue of its parent tree in Gloucestershire.

317

318 *European Diversity Panel*: In 2014, twig material was collected from 37 trees representing 37
319 European provenances in a trial of *F. excelsior* established in 2004 at Paradise Wood, Earth
320 Trust, Oxfordshire, UK, as part of the Realising Ash's Potential project. DNA was extracted
321 from cambial tissue of the twigs using a CTAB protocol.

322

323 *British Screening Panel*: In 2015, freshly flushed leaf material was collected from a clonal
324 seed orchard of *F. excelsior* growing at Paradise Wood, Earth Trust, Oxfordshire, UK for
325 RNA extraction and cDNA synthesis as in Harper et al.³. Single whole leaves were
326 harvested from four ramets of each of 130 ash trees selected from phenotypically superior
327 parents throughout Britain, that had been cloned by grafting.

328

329 **2451S DNA Sequencing and Genome Assembly**

330

331 The genome size of 2451S was estimated by flow cytometry with propidium iodide (PI)
332 staining of nuclei, using leaf tissue co-chopped with an internal standard using a razor blade.
333 Three preparations were made, two with *Petroselinum crispum* 'Curled Moss' parsley as
334 standard (2C = 4.50 pg)³³ and one with *Solanum lycopersicum* 'Stupicke polni rane' (2C =
335 1.96 pg)³⁴ as standard. The Partec CyStain Absolut P protocol was used (Partec GmbH,
336 Germany). Each preparation was measured six times, with the relative fluorescence of over
337 5000 particles per replicate recorded on a Partec Cyflow SL3 (Partec GmbH, Germany) flow
338 cytometer fitted with a 100-mW green solid state laser (Cobolt Samba; Cobolt, Sweden). The
339 resulting histograms were analysed with the Flow-Max software (v. 2.4, Partec GmbH). The
340 measurement with the tomato internal standard was used as the best estimate of genome
341 size, because the tomato genome size is closest to that of 2451S, yielding a more accurate
342 result.

343

344 Genomic DNA of 2451S was sequenced using the following methods: (1) HiSeq 2000
345 (Illumina, San Diego, CA) at Eurofins, Ebersberg, Germany, with 100 bp reads and shotgun
346 libraries with fragment sizes of 200 bp, 300 bp, and 500 bp, and long jumping distance (LJD)
347 libraries with 3 kbp, 8 kbp, 20 kbp and 40 kbp insert sizes, generating 188X genome
348 coverage; (2) 454 FLX+ (Roche, Switzerland) at Eurofins with shotgun libraries and
349 maximum read length of 1,763 bp and mean length of 642 bp giving 4.3X genome coverage;
350 and (3) MiSeq (Illumina, San Diego, CA) at The Genome Analysis Centre, Norwich, UK, with
351 300 bp paired-end reads from a Nextera library with ~5 kbp insert size, giving 16X genome
352 coverage (see Supplementary Table 1). We assembled and released five genome assembly
353 versions over the course of 3 years, details of which can be found in Supplementary Table 3.
354 The most recent version assembled first into 235,463 contigs with a total size of 663 Mbp
355 and an N50 of 5.7 kbp (Supplementary Table 2), and after scaffolding and removing
356 organellar scaffolds, the assembly comprised 89,487 scaffolds totaling 867 Mbp (17% "N")
357 with an N50 of 104 kbp (Supplementary Table 2). The plastid genome was assembled
358 separately into one circular contig of 155,498 bp, including an inverted repeat region of
359 approximately 25,700 bp. The mitochondrial genome initially assembled into 296 contigs
360 totaling 232 kbp. After several rounds of contig extension using overlaps of mapped 454
361 reads the final assembly consisted of 26 contigs totaling 581 kbp with an N50 of 60.6 kbp.
362

363 All Illumina reads from 2451S were trimmed using CLC Genomics Workbench (QIAGEN
364 Aarhus, Denmark) versions 6-8 (depending on when the data was received) to a minimum
365 quality score of 0.01 (equivalent to Phred quality score of 20), a minimum length of 50 bp,
366 and were also trimmed of any adaptor and repetitive telomere sequences. The MiSeq
367 Nextera reads were also run through FLASH³⁵ to merge overlapping paired reads, and
368 NextClip³⁶ to remove adaptor sequences, both used with default parameters. Roche 454
369 reads were trimmed to a minimum Phred score of 0.05, and minimum length of 50 bp. *De*
370 *novovo* assembly was performed with the CLC Genomics Workbench, using the 200 bp, 300
371 bp, 500 bp, and 5 kbp insert size Illumina library reads to build the De Bruijn graphs. The
372 remaining Illumina reads and the 454 reads were used as 'guidance only reads' to help
373 select the most supported path through the De Bruijn graphs. A word size (k-mer) of 50 and
374 maximum bubble size of 5000 were used to assemble the reads into contigs with a minimum
375 length of 500 bp. Contigs were then scaffolded with the stand-alone tool SSPACE³⁷ Basic
376 v2.0 using all paired Illumina reads, with the '-k' parameter (number of mapped paired reads
377 required to join contigs) set to 7. Gaps in the scaffolds were closed using the GapCloser
378 v1.12 program using all paired reads (except for LJD libraries), with pair_num_cutoff
379 parameter set at 7. 454 reads were mapped to the assembly and used to join overlapping
380 scaffolds using the Jelly.py script from PBSuite³⁸ v14.7.14 with blasr parameters: -minMatch
381 11 -minPctIdentity 70 -bestn 1 -nCandidates 10 -maxScore -500 -noSplitSubreads.
382 Contig57544 was removed from the assembly because it aligned fully to the PhiX
383 bacteriophage genome, indicating it derived from the PhiX control library added to Illumina
384 sequencing runs.
385

386 To assemble the plastid and mitochondrial genomes, high read depth 50 bp k-mers were
387 extracted from the 200, 300, and 500 bp read libraries. Jellyfish³⁹ v2.1.1 was used to count
388 the depth for each k-mer, and these values were plotted in a scatterplot to identify peaks that
389 could correspond to the organellar genomes. Every k-mer over 600x coverage was used in a
390 BLAST search against the NCBI non-redundant (nr) database with a filter allowing only plant
391 sequences. K-mers were then extracted based on whether their first hit contained a
392 'mitochondrion' or 'plastid / chloroplast' related description. Reads from the 200, 300 and
393 500 bp libraries were then filtered against the k-mer sets, and were kept if the first and last
394 50 bp matched k-mers from the extracted sets (reads were at most 90 bp long). Each set of
395 reads (mitochondrial and plastid) were then assembled *de novo* using the CLC Genomics
396 Workbench. The plastid genome assembled initially into two contigs, which were joined

397 using an alignment to the *Olea europaea* plastid genome (GenBank accession
398 NC_015401.1), with the inverted repeat region being identified also. Reads from the 454
399 library were mapped to the assembly to check the sequence and especially the join region.
400 The mitochondrial genome assembled first into 296 contigs. To fill in gaps and join the
401 contigs together, 454 reads were mapped against the assembly and contig ends were
402 extended using the Extend Contigs tool in the CLC Genome Finishing Module. The Join
403 Contigs tool was then used to join overlapping ends together, and 454 reads were mapped
404 to the resulting assembly to check any joined regions. Using this method of “Map-Extend-
405 Join” iteratively (approximately ten times in total), a more contiguous assembly of 26 contigs
406 was obtained.

407

408 **RNA Sequencing**

409

410 The five RNA samples (see “Tree Material” above) were sequenced paired-end on Illumina
411 HiSeq 2000 with 200 bp insert sizes, and a read length of 100 bp at the QMUL Genome
412 Centre, London, UK. Reads were trimmed using CLC Genomics Workbench to a minimum
413 quality score of 0.01 (equivalent to Phred score of 20) and minimum length of 50 bp, and
414 adaptors were also removed (Supplementary Table 6).

415

416 **Analysis of repetitive DNA**

417

418 The repetitive element (transposable elements, TEs, and tandem repeats) content of the ash
419 genome was analysed via two approaches: (1) *de novo* identification of the most abundant
420 repeat families from unassembled 454 and Illumina reads; (2) *de novo* and similarity-based
421 identification of repeats from the ash genome assembly.

422

423 *De novo identification of repeat families from unassembled reads.* Individual 454 reads and
424 Illumina read pairs from the 500 bp insert library (post adaptor trimming, but prior to any
425 further quality control or filtering – see above) were used for *de novo* repeat identification.
426 Reads were quality filtered and trimmed using the FASTX-Toolkit v. 0.0.13
427 (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Using *fastx_trimmer*, the first 10 bp of all
428 reads (454 and Illumina) was removed (due to skewed base composition). 454 reads were
429 clipped to a maximum of 250 bp and Illumina reads to a maximum of 90 bp; all shorter reads
430 were removed using a custom Perl script. Reads were then quality filtered with the
431 *fastq_quality_filter* tool to retain only those where 90% of bases had a phred score of ≥ 20 .
432 Exact duplicates (which are likely artifacts from the emulsion PCR⁴⁰) were removed from the
433 454 reads using the *fastx_collapser* tool.

434

435 The complete set of quality filtered and trimmed 454 reads (3,330,483) was used as input for
436 the RepeatExplorer pipeline on Galaxy⁴¹, with a minimum of 138 bp overlap for clustering
437 and a minimum of 100 bp overlap for assembly. All clusters containing $\geq 0.01\%$ of the input
438 reads were examined manually in order to identify clusters that required merging (i.e. where
439 there was evidence that a single repeat family had been split over multiple clusters). Clusters
440 were merged if they met the following three criteria: (1) they shared a significant number of
441 similarity hits (e.g. in a pair of clusters, 10% of the reads in the smaller cluster had BLAST
442 hits to reads in the larger cluster); (2) they were the same repeat type (e.g. LINEs); (3) they
443 could be merged in a logical position (e.g. for repetitive elements containing conserved
444 domains these domains would be joined in the correct order). The re-clustering pipeline was
445 run with a minimum of 100 bp overlap for assembly; merged clusters were examined
446 manually to verify that all domains were in the correct orientation.

447

448 Quality filtered and trimmed Illumina reads were paired using the FASTA interlacer tool (v.
449 1.0.0) in RepeatExplorer, resulting in 111,230,011 pairs; unpaired reads were discarded. An

450 initial run of RepeatExplorer with a sample of 100,000 read pairs was performed to obtain an
451 estimate of the maximum number of reads that could be handled by the pipeline. A random
452 sample of 3.5 million read pairs was then taken using the sequence sampling tool (v. 1.0.0)
453 in RepeatExplorer and used as input for the clustering pipeline, which further randomly
454 subsampled the reads down to 3,370,186 pairs. The pipeline was run with a minimum of 50
455 bp overlap for clustering and a minimum of 36 bp overlap for assembly. Clusters containing
456 $\geq 0.01\%$ of the input reads were merged if $k_{x,y}$ passed the 0.2 cut-off (for clusters x and y , $k_{x,y}$
457 is defined as: $k_{1,2} = 2 * W / (n_1 + n_2)$ where W is the number of read pairs shared between clusters
458 x and y and n_x is the number of reads in cluster x which does not include the other read from
459 its pair within the same cluster); clusters that passed this threshold but which had no
460 similarity hits to each other were not merged. The re-clustering pipeline was run with a
461 minimum of 36 bp overlap for assembly.

462
463 Repeat families identified by RepeatExplorer were annotated according to the results of
464 BLAST searches to the Viridiplantae RepeatMasker library, to a database of conserved
465 protein coding domains from transposable elements and to a custom RepeatMasker library
466 comprising all *Fraxinus* sequences (excluding shotgun sequences), all mitochondrial
467 genome sequences from Asterids and all plastid genome sequence from Oleaceae available
468 from NCBI (downloaded on 13.02.2014); these BLAST searches were performed as part of
469 the RepeatExplorer pipeline. For repeat families that were not annotated in RepeatExplorer
470 (i.e. no significant BLAST hits), or where only very few reads ($< 2\%$) had a BLAST hit or
471 separate reads matched different repeat types (i.e. inconsistent BLAST hits), contigs were
472 also searched against the nr/nt database in GenBank using BLASTN with an E-value cut-
473 off⁴² of $1e-10$, against the nr database using BLASTX with an E-value cut-off of $1e-05$, and
474 submitted to Tandem Repeat Finder v. 4.07b with default parameters⁴³. Annotation of repeat
475 families from the clustering of the 454 and Illumina data was cross-validated by BLAST
476 searching the contigs from each analysis against each other using the blastn program in the
477 BLAST+ package (v. 2.2.28+) with an E-value cut-off of $1e-10$ and the DUST filter switched
478 off. Any repeat families annotated as plastid or mitochondrial DNA were removed prior to
479 downstream analyses (see below).

480
481 *Identification of repeats from the genome assembly. De novo* identification of repetitive
482 elements from the assembled ash genome sequence was conducted with RepeatModeler v.
483 1.0.7 (www.repeatmasker.org/RepeatModeler.html) using RMBlast as the search engine. All
484 unannotated ('unknown') repeat families from the RepeatModeler library were searched
485 against a custom BLAST database of organellar genomes (see above) using BLASTN with
486 an E-value cutoff of $1e-10$ in the BLAST+ package (v. 2.2.28+⁴⁴). Any repeat families
487 matching plastid or mitochondrial DNA were removed.

488
489 To prevent any captured gene fragments within repetitive element families causing the
490 masking of protein coding genes within the ash assembly, the custom repeat libraries were
491 pre-masked using the TAIR10 CDS dataset⁴⁵ (TAIR10_cds_20101214_updated;
492 downloaded from www.arabidopsis.org). First, transposonPSI v2
493 (<http://transposonpsi.sourceforge.net>) was run with the 'nuc' option to identify any TE-related
494 genes within the TAIR10 CDS dataset. Sequences with a significant hit to TE-related
495 sequences (E-value cut-off of $1e-05$) were removed from the TAIR10 CDS file ($n=308$); a
496 further 19 sequences that included the term "transposon" in their annotation, but which did
497 not have a hit using transposonPSI, were also removed. The filtered TAIR10 CDS dataset
498 was used to hard mask the RepeatModeler library, the RepeatExplorer libraries (454 and
499 Illumina) and the library from RepeatMasker using RepeatMasker v. 4.0.5
500 (www.repeatmasker.org) with RMBlast as the search engine and the following parameter
501 settings: -s -no_is -nolow. The four pre-masked libraries were combined into a single
502 custom repeat library; any repeat families annotated as 'rRNA', 'low-complexity' or 'simple'

503 were removed prior to combining the libraries. The combined library was then used to
504 identify repetitive elements in the ash genome assembly with RepeatMasker v. 4.0.5, using
505 the same parameter settings as above. RepeatMasker results were summarised using
506 ProcessRepeats with the species set to 'eudicotyledons' and using the 'nolow' option.
507

508 In addition to the analysis with the combined custom ash repeat library, repeats within the
509 assembly were also annotated by running RepeatMasker separately with each of the four
510 individual repeat libraries with parameter settings as described above. The results were
511 saved in gff format and combined into a single gff file that was then used to inform the
512 process of annotating protein coding genes (see below, "Gene Annotation").
513

514 Although the ash genome assembly covers c. 99% of the expected genome size based on
515 flow cytometry, c. 17% is comprised of Ns. Therefore, the repeat content of the genome
516 assembly may be an underestimate of the actual amount of repetitive DNA within the
517 genome. To test whether the c. 18% of missing sequence includes additional repetitive
518 elements we analysed the repeat content of individual Illumina reads that do not map to the
519 genome assembly. Quality-trimmed and length-filtered reads from the Illumina short insert
520 libraries (Supplementary Table 1) were mapped to the assembly using the 'Map Reads to
521 Reference' tool in the CLC Genomics Workbench, with both similarity match and length
522 match parameters set to 0.90. Unmapped reads from the 200 bp, 300 bp and 500 bp insert
523 libraries (equating to c. 4.8% of all reads from these libraries; see Supplementary Table 1)
524 were searched against the custom library of ash repeats using blastn (see Supplementary
525 Table 5) with an E-value cut-off of 1e-10 and the DUST filter switched off in the BLAST+
526 package (v. 2.2.29+⁴⁴).
527

528 To test for evidence for the expression of TEs, trimmed RNA sequencing reads from five
529 different tissue types (see Supplementary Table 7) were searched against the custom library
530 of ash repeats using blastn as described above for the unmapped DNA sequencing reads.
531

532 Gene Annotation

533
534 Protein coding genes were predicted using an evidence based annotation workflow
535 incorporating protein, cDNA and RNA-Seq alignments. Protein sequences from nine species
536 *Amborella trichopoda*, *Arabidopsis thaliana*, *Fraxinus pennsylvanica*, *Mimulus guttatus*,
537 *Populus trichocarpa*, *Solanum lycopersicum*, *Solanum tuberosum*, *Vitis vinifera* and *Pinus*
538 *taeda* (Supplementary Table 8) were soft masked for low complexity (segmasker-blast-
539 2.2.30) and aligned to the softmasked (for repeats) BATG-0.5 assembly with exonerate⁴⁶
540 protein2genome v-2.2.0 ; alignments were filtered at a minimum 60% identity and 60%
541 coverage, except for *F. pennsylvanica* which were filtered at a minimum of 80% identity and
542 60% coverage. Publically available *F. excelsior* ESTs (12,083 from Genbank) were aligned
543 with GMAP (r20141229)⁴⁷ and filtered at a minimum 95% identity and 80% coverage.
544

545 RNA-Seq reads from the five sequenced RNA samples were filtered for adaptors and quality
546 trimmed, rRNA reads were identified and removed⁴⁸ (trim_galore-0.3.3
547 http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ : -q 20 --stringency 5 --length
548 60; sortmerna-1.9: -r 0.25 --paired-out). RNA-Seq reads were aligned using Tophat
549 (v2.0.13/bowtie-2.2.3)⁴⁹ and transcript assemblies were generated using three alternative
550 methods Cufflinks (v2.2.1)⁵⁰, StringTie (v1.04)⁵¹ and Trinity (genome guided assembly)⁵².
551 Assembled Trinity transcripts were mapped to the *F. excelsior* assembly using GMAP
552 (r20141229) at 80% coverage and 95% identity. A comprehensive transcriptome assembly
553 was created using Mikado (v0.8.5 – in-house tool, manuscript in preparation) based on the
554 GMAP Trinity alignments, Cufflinks and StringTie transcript assemblies. Mikado leverages
555 transcript assemblies generated by multiple methods to improve transcript reconstruction.

556 Loci are first defined across all input assemblies with each assembled transcript scored
557 based on metrics relating to ORF and cDNA size, relative position of the ORF within the
558 transcript, UTR length and presence of multiple ORFs. The best scoring transcript assembly
559 is then returned along with additional transcripts (splice variants) compatible with the
560 representative transcript.

561

562 Protein coding genes were predicted using AUGUSTUS⁵³ by means of a Generalized
563 Hidden Markov Model (GHMM) that takes both intrinsic and extrinsic information into
564 account. An AUGUSTUS *ab initio* model was generated based on a subset of cufflinks
565 assembled transcripts identified by similarity support as containing full-length open reading
566 frames. Gene models were predicted using the trained *ab initio* model with the nine sets of
567 cross species protein alignments, RNA-Seq junctions (defining introns), and Mikado
568 transcripts as evidence hints. RNA-Seq read density was provided as exon hints and repeat
569 information (interspersed repeats) as nonexonpart hints. We generated two alternative
570 AUGUSTUS models by either including or excluding the RNA-Seq read depth information. A
571 set of integrated gene models was derived from the two AUGUSTUS runs along with the
572 transcriptome and protein alignments via EvidenceModeler:r20120625 (EVM)⁵⁴. Weights of
573 evidence were manually set following an initial testing and review process as: AUGUSTUS
574 predictions with RNA-Seq read depth hint, weight 2; AUGUSTUS predictions without RNA-
575 Seq read depth hint, weight 1; protein alignment high confidence (greater than 90%
576 coverage, 60% identity) weight 5; protein alignment low confidence (lower than 90%
577 coverage 60% identity) weight 1; cufflinks transcripts, weight 1; Mikado transcripts, weight
578 10; RNA-seq splice junctions, weight 1. We identified examples of EVM errors resulting from
579 incomplete genes in the AUGUSTUS gene predictions or non-canonical splicing; to rectify
580 these problems we substituted the EVM model for the overlapping AUGUSTUS model (with
581 RNA-Seq read depth hints). To add UTR features and alternative splice variants we ran
582 PASA⁵⁵ with Mikado transcript assemblies and available *F. excelsior* ESTs using the
583 corrected EVM models as the reference annotation.

584

585 The PASA updated EVM models were further refined by removing gene models that showed
586 no expression support (using all available RNA-Seq libraries) or had no support from cross
587 species protein alignments or no blast similarity support with a Viridiplantae (without
588 *Fraxinus excelsior*) protein database (< 50% blast high-scoring segment pair (HSP)
589 coverage) or where the CDS length was less than 100 bp (retaining those transcripts with ≥
590 50% blast HSP coverage). Gene models were also excluded if they aligned with ≥ 30%
591 similarity and 40% coverage to the TransposonPSI (v08222010) library
592 (<http://transposonpsi.sourceforge.net/>) and had ≥ 40% coverage by the
593 RepeatModeler/RepeatMasker derived interspersed repeats. In addition, gene models that
594 had ≥ 30% similarity and 60% coverage to the TransposonPSI library or had ≥ 60%
595 coverage by the RepeatModeler/RepeatMasker derived interspersed repeats were also
596 excluded. The functional annotation of protein coding genes was generated using an in-
597 house pipeline - AnnotF-1.01, that executes and integrates the results from InterProSCAN
598 (version 5) and Blast2GO (version 2.5.0). Completeness of transcript models was classified
599 by Full-lengther Next⁵⁶ and coherence in gene length examined by comparison with single
600 copy gene BLAST hits in monkey flower (Extended Data Fig. 1).

601

602 Transfer RNA (tRNA) genes were predicted by tRNAscanSE-1.3.1 with eukaryote
603 parameters⁵⁷ and rRNAs using rnammer-1.2⁵⁸. miRNA was predicted by BLASTN searches
604 with precursor miRNAs from miRBase⁵⁹ 21.0 against the reference genome sequence
605 (BLAST 2.2.30, E-value 1e-06) and miRCat⁶⁰ using the mature miRNAs from miRBase with
606 default plant parameters, except modifying the flanking window to 200 bp. Putative miRNA
607 precursors from these methods were combined and were folded using RNAfold⁶¹ and mature

608 miRNAs from miRBase were aligned to precursor hairpins using PatMan⁶². These
609 predictions were checked manually for RNA secondary structure.

610

611 Organellar genes were annotated manually using the BLAST tool within the CLC Genomics
612 Workbench v7.5. Mitochondrial genes were identified using CDS from *M. guttatus*, *Nicotiana*
613 *tabacum* and *A. thaliana* (all downloaded from NCBI). Plastid genes were identified using
614 CDS from *Olea europaea* and *N. tabacum* (both downloaded from NCBI). An E-value cut-off
615 of 1E-04 was used. Gene and CDS annotations were added manually to the *F. excelsior*
616 organellar scaffolds using the sequence editing tools available within the CLC Genomics
617 Workbench. In the plastid genome, we annotated 72 protein-coding, 7 putative coding (ycf),
618 rRNA, and tRNA genes. On the mitochondrial scaffolds, we annotated 37 protein-coding,
619 rRNA and tRNA genes.

620

621 **Analysis of whole genome duplications**

622

623 To examine evidence for past whole genome duplication, CDS and protein sequences (one
624 transcript per gene) were taken from our ash genome annotation, and downloaded from
625 Phytozome v10.3 for tomato (*S. lycopersicum*), monkey flower (*M. guttatus*), and grape (*V.*
626 *vinifera*), the CoGe database for bladderwort (*Utricularia gibba*) and coffee-genome.org for
627 coffee (*Coffea canephora*). For olive (*Olea europaea*) we predicted open reading frames
628 from transcriptome data⁶³ using Transdecoder⁵² with all parameters set to defaults (v.2.01
629 <http://transdecoder.github.io>). Olive⁶³ is in the same family as ash (Oleaceae); monkey
630 flower⁸ and bladderwort⁶⁴ in the same order as ash (Lamiales); tomato⁴, and coffee⁶⁵, in
631 different orders (Solanales and Gentianales, respectively), but like ash in the Asterids; and
632 grape⁶⁶, is a Rosid. An all-against-all comparison using protein sequences was carried out
633 on each species separately using BLASTp v2.2.29, with an e-value cutoff of 1E-05. BLAST
634 alignments were further filtered to retain pairs for which the shorter sequence was at least
635 50% of the longer sequence, and the alignment was at least 50% of the shorter sequence. If
636 one sequence had multiple matches meeting the length and e-value thresholds these were
637 grouped into a paralog group, including any other genes that were associated with the
638 matches (e.g. if gene A matches gene B and gene C, and gene C also matches gene D,
639 then one group of A, B, C and D would be formed).

640

641 Next, all possible pairs of protein sequences within each group were aligned using muscle
642 v3.8.31 with default parameters⁶⁷. A nucleotide alignment was generated from the protein
643 alignment using a python script. Synonymous substitutions were estimated using the
644 codeml program from PAML v4.8⁶⁸. The Ks scores within each group were then corrected to
645 remove redundant values; only those representing duplication events within the group were
646 retained (in a group of n genes, there are n-1 possible duplication events) using the method
647 described in Maere et al⁶⁹ and Blanc & Wolfe⁹. These steps are implemented in a python
648 script available online: github.com/EndymionCooper/KSPlotting.

649

650 In order to examine patterns of conserved synteny we constructed syntenic dotplots using
651 the SynMap⁷⁰ with default parameters (Extended Data Fig. 2). The default uses LAST⁷¹ to
652 perform similarity searches, and DAGchainer⁷² to find syntenic regions. By default
653 DAGchainer requires a minimum of five aligned gene pairs with no more than 20 genes
654 between neighbouring pairs.

655

656 Pairs of genes were categorised as 'local' duplications if they were located on the same
657 chromosome or scaffold and resided within 10 genes of each other, and as 'tandem'
658 duplications if they reside directly next to each other. GO term enrichment was performed on
659 ash proteins using the BLAST2GO plugin suite of tools within the CLC Genomics
660 Workbench v8.5. Three separate BLAST searches were run against the RefSeq protein

661 database: firstly using CDS from all genes as queries, secondly using CDS from genes
662 involved in WGD (excluding locally duplicated genes), and thirdly using CDS from locally
663 duplicated genes (genes located within 10 genes of each other). The E-value cut-off for all
664 BLAST runs was 1e-05. BLAST results were annotated with GO terms using the 'Mapping'
665 and 'Annotation' tools within the BLAST2GO plugin, using default parameters except for:
666 Annotation Cutoff = 55 and HSP-Hit Coverage Cutoff = 40. Significantly enriched GO terms
667 were identified using the Fisher Exact Test tool within the plugin, where the reference set
668 was the GO terms for all genes, and an FDR of 0.05 was used.

669

670 **Analysis of gene families**

671

672 The OrthoMCL pipeline (v.2.0.9)⁷³ was used to identify clusters of orthologous and
673 paralogous genes from *F. excelsior* and: *Amborella*⁷⁴, *Arabidopsis*⁷⁵, barrel medic⁷⁶,
674 bladderwort⁶⁴, coffee⁶⁵, grape⁶⁶, loblolly pine⁷⁷, monkey flower⁶, poplar⁷⁸ and tomato⁴
675 (Supplementary Table 10). Input proteomes contained a single transcript per gene and were
676 filtered with orthomclFilterFasta to remove any sequences of < ten amino acids in length
677 and/or > 20% stop codons. Similar sequences were identified via an all versus all BLASTP
678 search for the 362,741 proteins remaining after filtering. The BLAST search was performed
679 in the BLAST+ package⁴⁴ (v.2.2.29+), using an e-value cut-off of 1e-05. BLAST results were
680 filtered with orthomclPairs to retain protein pairs that match across ≥ 50% of the length of the
681 shorter sequence in the pair. Clustering of sequences was carried out with mcl⁷⁹ (v.14.137)
682 using a setting of 1.5 for the inflation parameter. The output from OrthoMCL was
683 summarised using a custom Perl script to obtain counts of the number of sequences from
684 each species belonging to each group. Venn diagrams for selected taxa were generated
685 using InteractiVenn⁸⁰.

686

687 **European Diversity Panel sequencing**

688

689 DNA from the 37 European Diversity Panel trees was sequenced at The Genome Analysis
690 Centre on Illumina HiSeq, using paired-end insert sizes between 100 and 700 bp, and a read
691 length of 150 bp. This generated an average of 63.6 million 150 bp reads (10.9X genome
692 coverage) per tree. Filtering and trimming steps reduced this average to 55.3 million reads.
693 An average of 85.8% of these reads per tree mapped to our reference genome. In addition,
694 DNA reads from Danish Tree35 library '3077' were downloaded from the Open Ash Dieback
695 ftp site (<http://oadb.tsl.ac.uk>); these were 250 bp paired-end reads with an insert size
696 between 200 and 400 bp. Tree35 is given the sample number '38' in all further population
697 analysis.

698

699 **European Diversity Panel genome-wide SNP calling**

700

701 The raw reads from the 37 trees in the European Diversity Panel (Supplementary Table 11)
702 were aligned to the reference genome using bowtie 2.2.5⁸¹. The alignments were converted
703 to the BAM format and the duplicated reads were removed with samtools 1.2⁸². To assign
704 each read to its corresponding tree, the flag 'rg' was added to each BAM file with picard tools
705 1.119 (<http://broadinstitute.github.io/picard/>). SNPs were called with freebayes 1.0.2¹⁰ to
706 produce a VCF file. The SNPs with quality < 300 were filtered with bio-samtools 2.1⁸³.
707 SnpEff 4.1g⁸⁴ was used to predict the effect of the putative SNPs (see Supplementary Table
708 12). Genic regions were within 5kbp from a gene model. Amino acid changes are labelled as
709 missense_variant.

710

711 **SNP calls validation using the KASP platform**

712

713 In order to test the reliability of SNP calls in the genome-wide SNP calling, we designed
714 KASP assays for 53 SNPs, which ranged in their level of confidence (see Supplementary
715 Table 13). None of the SNP calls tested by KASP were present in the reduced SNP set used
716 for population genetic analyses. Primers were designed with a modified version of
717 PolyMarker⁸⁵ including the FAM or HEX tails (FAM tail: 5' GAAGGTGACCAAGTTCATGCT
718 3'; HEX tail: 5' GAAGGTCGGAGTCAACGGATT 3'). The primer mix was prepared as
719 recommended by the manufacturer [46 µL dH₂O, 30 µL common primer (100 µM) and 12 µL
720 of each tailed primer (100 µM)] (<http://www.lgcgroup.com/services/genotyping>). The assays
721 were run on 37 individuals from the European Diversity panel, in 384-well plates as 4µL
722 reactions [2-µL template (10–20 ng of DNA), 1.944 µL of V4 2× Kaspar mix and 0.056 µL
723 primer mix]. PCR was done with the following protocol: hotstart at 95 °C for 15 min, followed
724 by ten touchdown cycles (95 °C for 20 s; touchdown 65 °C, -1 °C per cycle, 25 s) then
725 followed by 30 cycles of amplification (95 °C 10 s; 57 °C 60 s). Fluorescence was detected
726 on a Tecan Safire at ambient temperature. Genotypes were called using Klustercaller
727 software (version 2.22.0.5; LGC Hoddlesdon, UK). Four of the individuals did not amplify and
728 were discarded from the analysis. The result of the calls are in Supplementary Data 7.

729

730 **European Diversity Panel population genetics and history using a reduced set of** 731 **SNPs**

732

733 For population structure analyses and effective population size estimation, variants were
734 only called at SNP sites in the genome where all 38 samples have between 5 and 30x
735 coverage. We refer to this as the 'reduced SNP set'.

736

737 First, all reads were trimmed in the CLC Genomics Workbench to a minimum quality score of
738 0.01 (equivalent to Phred quality score of 20), a minimum length of 50 bp, and were also
739 trimmed of any adaptor and repetitive telomere sequences. Filtered reads were mapped to
740 the reference assembly using the 'Map Reads to Reference' tool in the CLC Genomics
741 Workbench, setting both similarity match and length match parameters to 0.95. Regions with
742 coverage of between 5 and 30 reads in all samples were extracted using the 'Create
743 Mapping Graph', 'Identify Graph Threshold Areas' and 'Calculus Track' tools. These
744 extracted regions totaled 20.6 Mbp (2.3% of the genome)

745

746 Variant calling was performed on a read mapping pooled from all samples, using the 'Low
747 Frequency Variant Caller' tool in the CLC Genomics Workbench, with the coverage-
748 restricted regions from the previous step used as a track of target regions. This prevented
749 variants being called where some samples did not have read coverage, and also in the
750 organellar scaffolds where the read coverage is very high. The following parameters were
751 changed from default: Ignore positions with coverage above = 1000, Ignore broken pairs =
752 no, Ignore non-specific matches = Reads, Minimum Coverage = 190 (38 samples with at
753 least 5 reads each should have a combined total coverage of > 190), Minimum Count = 10,
754 Minimum Frequency = 5%, Base Quality Filter = Yes, Neighbourhood radius = 5, Minimum
755 Central Quality = 20, Minimum neighbourhood quality = 15, Read Direction Filter = yes,
756 Direction Frequency = 5%. As a result 529,812 variants were called, comprising 468,237
757 SNPs, 14,850 equal replacements (where > 1 nucleotides are replaced by an equal number
758 of nucleotides), 26,043 deletions, 19,085 insertions, and 1,597 unequal replacements
759 (where at least one SNP lies directly beside an indel). The average quality of all reads at
760 these variant positions was 36.2.

761

762 To genotype each sample individually at the variant loci called in the previous steps, the
763 'Identify Known Mutations from sample mappings' tool within the CLC Biomedical Genomics
764 workbench was used. The workflow takes a track of known variants as input (such as those
765 called from the pooled read mapping) and reports the presence, absence, coverage, count

766 and other statistics, of each variant locus in the read mapping of another sample (in this
767 case, the read mapping from each of the 38 trees). The 'Identify Candidate Variants' tool
768 was then used to filter variants with a minimum coverage of 5, minimum count of 3 and
769 minimum frequency of 20%. VCF files for each tree were exported from the CLC Workbench
770 and merged into one file using the vcf-merge tool from VCFtools⁸⁶. The merged VCF file was
771 then filtered using vcfutils, to remove indels, multi-allelic loci, and loci with a Minimum Allele
772 Frequency (MAF) < 0.05, with 394,885 SNP loci remaining. This set of high quality SNPs
773 with comprehensive knowledge of the genotype of every sample is referred to as the
774 'reduced SNP set' and is used for further population analyses.

775

776 To visualise similarities and differences among the genomes of the European Diversity
777 Panel, PCA was performed using the SNPRelate v1.4.2⁸⁷ package in R v3.1.2. The filtered
778 VCF file was converted into gds using the snpGDSVCF2GDS command, and was filtered on
779 an LD value of 0.1 using the snpGDSLDPPruning command, leaving 34,607 SNPs. PCA was
780 performed on the pruned set of SNPs using the snpGDSPCA command with default options,
781 and the results of the first three PCs were plotted in R.

782

783 To analyse population structure in the European Diversity Panel, scaffolds were selected
784 that contained 10 or more SNPs in the filtered VCF file (8,955 nuclear scaffolds in total).
785 Three different SNPs were selected at random from each of these scaffolds, and placed into
786 three different files in STRUCTURE input format (26,865 SNPs in total, 8,955 in each set).
787 STRUCTURE v2.3.4⁸⁸ was run with admixture from k=1 to k=20 for each of the three sets of
788 SNPs, with both BURNIN and NUMREPS set to 100,000. All output results were run through
789 Structure Harvester Web v0.6.94⁸⁹, which found k=3 to have the largest delta k value of
790 32.91 (Extended Data Fig. 3). Next, the three runs of k=3 were used as input into CLUMPP
791 v1.1.2⁹⁰ to align the clusters, and samples within each cluster. Aligned results were imported
792 back into STRUCTURE v2.3.4 to generate Q-value bar plots. Average Q-values from the
793 three runs were used to generate a map with pie charts, using Tableau v9.3 (Tableau,
794 Seattle, US) with Tableau base-map country outlines. Each section of the pie represented
795 the average Q-value of the individual belonging to the coloured cluster (Fig. 2b).

796

797 To analyse relationships among plastid sequences in plastid haplotype networks, a
798 consensus sequence of the large single copy plastid region was extracted for each of the 38
799 samples. The sequences were then aligned using the Create Alignment tool in the CLC
800 Genomics Workbench, and the alignment was exported in Phylip format. The alignment was
801 imported into PopArt v1.7 [<http://popart.otago.ac.nz>], where a Median-Joining network was
802 generated. Results were visualised on a map using Tableau v9.3 (Fig. 2a) with Tableau
803 base-map country outlines.

804

805 We estimate the effective population size history of *F. excelsior* using two complementary
806 methods; the PSMC¹⁴ model estimates the history in the non-recent past, whereas by using
807 Linkage Disequilibrium, we can estimate the population size more recently. The Pairwise
808 Sequentially Markovian Coalescent (PSMC) model calculates the effective population size
809 using a Time to Most Recent Common Ancestor (TMRCA) approach. The effective
810 population size history is then estimated from the number of recombination events
811 separating segments of constant TMRCA. The program PSMC 0.6.5¹⁴, takes only a diploid
812 consensus sequence as input. To estimate past effective population size, PSMC analysis
813 was used on the reference tree. DNA reads from the 2451S 200, 300 and 500 bp libraries
814 were mapped to the 2451S reference sequence using CLC Genomics Workbench 'Map
815 Reads to Reference' tool (length fraction = 0.95 and similarity fraction = 0.9). The mapping
816 was exported in bam format, and a consensus sequence was obtained following PSMC
817 recommendations, by using samtools v0.1.18 'mpileup' command with options: -C 50 -A -Q
818 20 -u, bcftools v1.1 to convert the bcf file to vcf format, and finally using vcfutils.pl to convert

819 the vcf file to a consensus sequence where the coverage was between 5 and 200. The
820 PSMC program was then run with default parameters except for: -p "4+25*2+4+6", with one
821 hundred bootstraps. To scale the results, the psmc_plot.pl script was used with default
822 parameters except for the following: -u 7.5e-09 -g 15 -N 0.25 (the mutation rate of *F.*
823 *excelsior* is unknown, so the substitution rate of 7.5e-09 is taken from a study on *Arabidopsis*
824 *thaliana*⁹¹). Effective population size estimates were then plotted in R v3.1.2 (Fig. 2f).

825
826 Effective population size estimation by Linkage Disequilibrium (LD) in the European Diversity
827 Panel was performed using the program SNeP v1.1⁹², which takes genome-wide
828 polymorphism data from several individuals in a population as input. The European Diversity
829 Panel filtered VCF file with the reduced SNP set of 38 trees (same as used in PCA and
830 STRUCTURE analysis) was converted into Map and Ped files. The third column in the Map
831 file (linkage distance in Morgans) was set to zero for all SNPs, as these values were
832 unknown and SNeP calculates this value from each SNP's physical distance. SNeP was
833 then run with a minimum distance between SNPs of 10,000 bp and a maximum of 400,000
834 bp, with Sved's modifier for recombination rate, and with 50 bins. Estimated effective
835 population sizes were plotted in R (Extended Data Fig. 3c), as well as LD decay over
836 distance between 100 and 300,000 bp (Fig. 2e).

837

838 **Simple-sequence repeat analysis**

839

840 To develop accessible population genetic markers, the repeat masked v0.4 2451S genome
841 was mined for simple sequence repeat (SSR) sequences (a repeat motif of 2-5 bp in length
842 repeated a minimum of 5 times) using the QDD-v.3.1 pipeline⁹³. Downstream QDD-v.3.1
843 pipes screened SSR loci (inclusive of the SSR repeat motif and 200 bp forward and reverse
844 flanking regions) for singleton sequences in an all-against-all BLAST (-task blastn -evalue
845 1e-40 -lcase_masking -soft_masking true) and designed primer pairs within 200 bp flanking
846 regions using PRIMER3 software⁹⁴. The c. 31,300 singleton SSR loci identified in the ash
847 genome were screened using RepeatMasker Open-4.0 (<http://www.repeatmasker.org>) in
848 QDD-v.3.1 to eliminate loci which hit known transposable elements in the RepBase
849 Viridiplantae repeat library (<http://www.girinst.org>), leaving c. 28,800 SSR loci. The final
850 primer table output by the QDD-v.3.1 pipeline allows selection of the best primer pair design
851 for each SSR loci. To select candidate markers for further development, these primer pairs
852 were filtered according to parameters provided by QDD-v.3.1. The selected SSR loci had a:
853 maximum primer alignment score of 5; minimum 20 bp forward and reverse flanking region
854 between SSR and primer sequences; high quality primer design (defined by QDD pipeline as
855 an absence of homopolymer, nanosatellite, and microsatellite sequence in primer and
856 flanking sequences), and; minimum number of 7 motif repeats within the SSR sequence.
857 This filtering gave a set of 837 SSR loci, which was screened against the combined custom
858 ash repeat library for v0.5 of the 2451S genome assembly (see above - "Analysis of
859 repetitive DNA") via a blastn search with an E-value of 1e-10 in the BLAST+ package (v.
860 2.2.31+). Elimination of all sequences with a hit to known repetitive elements left 681
861 candidate loci. These were compared to the v0.5 assembly via a blastn search with an E-
862 value cut-off of 1e-10. This returned a set of 664 loci with a unique match to the v0.5
863 assembly for use as population genetic markers (see Supplementary Data 1).

864

865 *In silico* analysis of allelic diversity (i.e. locus polymorphism) of these SSR loci was carried
866 out by screening a subset of loci (366) against a variance table composed of insertions and
867 deletions recorded for the European Diversity Panel. Approximately half (48%) of the loci
868 tested were variable among 37 of the resequenced genomes (sample 38 not included).
869 Twenty candidate SSR loci with the greatest *in silico* allelic diversity were selected for wet
870 lab testing on seven individuals from the European Diversity Panel. Primer pairs with a
871 fluorescent tag on the 5' end of the forward primer (FAM, HEX or TAM) were used. For

872 singleplex PCR, primer aliquots were used at a concentration of 10 pmol/μl . PCR
873 amplification of target regions was carried out in singleplex reactions with a final reaction
874 volume of 10 ul, containing 1 ul genomic DNA, 0.2 ul of each primer (10pmol/ ul), 3 .6 ul of
875 RNase free water, and 5 ul of Qiagen Type-it Multiplex PCR Master Mix, in a G-Storm GS2
876 Multi Block Thermal Cycler. The amplification conditions were as follows: 5 min at 95°C; 18
877 cycles of 30 s at 95 °C, 90 s at 62 °C with a 0.5 °C reduction per cycle, 30 s at 72 °C; 20
878 cycles of 30 s at 95°C, 1 min 30s at 51 °C, 30 s at 72 °C; a final extension step of 30 min at
879 60 °C. PCR samples were diluted to 1:10 with dH₂O and run (on an Applied Biosystems
880 3730xl 96 capillary sequencing instrument with Applied Biosystems GeneScan 400HD Rox
881 dye size standard. Negative control samples were included for each primer pair PCR
882 reaction mix. Allele calling was carried out using GeneMarker v.2.6.4
883 (<http://www.softgenetics.com>).

884
885 Primer pairs which produced interpretable allele peaks from capillary sequencing of
886 singleplex reactions were arranged into four multiplex primer mixes (containing 5 primer
887 pairs each) according to PCR product size and fluorescent tag. Multiplex primer mixes were
888 tested on DNA extractions for a further 14 of the 37 trees from the European Diversity Panel.
889 For each multiplex, primer pair mixes were prepared at a final concentration of 10pmol/ μl
890 and amplified via PCR in 10μl reaction volumes (1 ul genomic DNA, 1 ul primer mix, 3 ul of
891 RNase free water, and 5 ul of Qiagen Type-it Multiplex PCR Master Mix) under the
892 amplification conditions described above. PCR product size range, allele counts, primer
893 design and successful multiplex panels for the 20 wet lab tested candidate SSR markers
894 developed for European ash are described in Supplementary Data 1.

895
896 Further multiplex primer mixes were tested on 7 trees from the European Diversity Panel for
897 amplification of the longest SSR loci (14 or more repeated motifs). Primer pair mixes were
898 prepared at a final concentration of 10pmol/ul and amplified via PCR in 8μl reaction volumes
899 (1 ul genomic DNA from a 1:10 dilution with nuclease free water, 1 ul primer mix, 2 ul of
900 RNase free water, and 4 ul of Qiagen Type-it Multiplex PCR Master Mix.). The amplification
901 conditions were as follows: 5 min at 95°C; 32 cycles of 30 s at 95 °C, 90s at 62 °C with a
902 0.35 °C reduction per cycle, 30 s at 72 °C; a final extension step of 30 min at 60 °C.
903 Amplification was performed in a G-Storm GS2 Multi Block Thermal Cycler. Size fraction
904 analysis of PCR products was carried out for two samples of each tested primer multiplex
905 using a 12 sample DNA1000/7500 chip in an Agilent 2100 Bioanalyzer
906 (<http://www.genomics.agilent.com>). Of the 28 primer pairs tested, 22 successfully amplified
907 across the six primer multiplexes tested (Supplementary Data 1).

908

909 **Association of transcriptomic markers with reduced susceptibility to ash dieback in** 910 **Denmark**

911

912 Sequence reads for the “Danish Scored Panel” of 182 Danish ash accessions (as described
913 in Harper *et al.*, 2016³; sequence reads are available in the European Nucleotide Archive
914 under the study accession number PRJEB10202) were mapped to a reference composed of
915 the complete set of CDS models (including 229 genes identified as possible TEs; see above,
916 Gene Annotation). This provided transcript abundance estimates for 40,133 CDS models
917 (Supplementary Data 2). Transcript abundance was quantified and normalized as reads per
918 kbp per million aligned reads (RPKM). After filtering out models exhibiting negligible
919 expression (mean RPKM value of below 0.4), 33,204 CDS models were analysed as
920 potential gene expression markers (GEMs; Supplementary Data 3). SNPs were called by the
921 meta-analysis of alignments (as described in Bancroft *et al.*⁹⁵) of mRNA-seq reads obtained
922 from each of the 182 accessions. SNP positions were excluded if they did not have a read
923 depth in excess of 20, a base call quality above Q20, missing data below 0.25, and three
924 alleles or fewer. An additional noise threshold was employed to reduce the effect of

925 sequencing errors, whereby ambiguous bases were only allowed to be called if both bases
926 were present at 0.15 or above. This resulted in a final set of 394,006 SNPs (Supplementary
927 Data 4) of which 234,519 had minor allele frequencies in excess of 0.05, and all of which
928 were within the CDS models constituting the GEM panel.

929

930 The SNP dataset for the 182 accessions was entered into the program PSIKO⁹⁶ to produce
931 a Q matrix, which was composed of two population clusters. The SNP genotypes, Q matrix
932 and ash dieback damage scores for these trees³ were incorporated into a compressed
933 mixed linear model⁹⁷ implemented in the GAPIT R package⁹⁸, with missing data imputed to
934 the major allele. The kinship matrix used in this analysis was also generated by GAPIT.

935

936 Gene expression marker (GEM) associations were calculated by a fixed effect linear model
937 in R with RPKM values and the Q matrix inferred by PSIKO as the explanatory variables and
938 damage score the response variable. R^2 , regression coefficients, constants and significance
939 values were outputted for each regression.

940

941 Twenty GEMs were associated with damage scores (Supplementary Data 3). A previous
942 analysis of the gene expression data, based on a simple mRNA transcript reference,
943 identified only 13 GEMs associated with ash dieback damage in ash³, with the strongest
944 associations exhibiting higher P values than the present study (best P values 5.31×10^{-12} and
945 9.83×10^{-13} respectively). The CDS models for the top three GEMs identified in the present
946 study had very high BLAST similarity to the transcripts for two of the GEMs identified in the
947 previous study. FRAEX38873_v2_000173540.4 ($P = 1.95 \times 10^{-10}$) corresponds with
948 Gene_23247_Predicted_mRNA_scaffold3380 from the previous study, but
949 Gene_19216_Predicted_mRNA_scaffold2427 resolved into two distinct CDS models in the
950 present study (FRAEX38873_v2_000261470.1, $P = 9.83 \times 10^{-13}$ and
951 FRAEX38873_v2_000199610.1, $P = 6.01 \times 10^{-12}$). The qRT-PCR primers designed for the
952 previous analysis³ were adequate for assaying FRAEX38873_v2_000173540.4 and
953 FRAEX38873_v2_000261470.1 and new primers were designed for
954 FRAEX38873_v2_000199610.1.

955

956 Two of the 20 significantly associated GEMs in the present study,
957 FRAEX38873_v2_000048360.1 ($P = 1.77 \times 10^{-9}$) and FRAEX38873_v2_000048340.1 ($P =$
958 3.48×10^{-7}), did not have high BLAST similarity to GEMs found in the previous study.
959 However, these GEMs were highly similar to a cDNA transcript containing a predictive A/G
960 SNP (termed a cSNP) identified previously, where presence of a G allele was associated
961 with low damage scores. Both of these GEMs contained the “less susceptible” G variant. A
962 third paralogous gene in this family with the A variant was also found
963 (FRAEX38873_v2_000184430.1), and was not identified as a GEM associated with damage
964 score ($P = 0.02$). The present study therefore resolves this cSNP marker into three
965 paralogous genes, two fixed for a “less susceptible” G nucleotide, and one a “susceptible” A
966 nucleotide.

967

968 These five GEMs were applied using qRT-PCR, and in the case of
969 FRAEX38873_v2_000048360.1 and FRAEX38873_v2_000048340.1 RT-PCR, to a small
970 test panel of 58 Danish accessions (henceforth “Danish Test Panel”) to assess their
971 predictive capabilities in a similar way as in Harper et al.³ Unlike this previous study
972 however, ratios between the bases of the FRAEX38873_v2_000048360.1 and
973 FRAEX38873_v2_000048340.1 were scored by eye (instead of simply scoring the presence
974 or absence of the “less susceptible” nucleotide), in order to estimate levels of gene
975 expression for the “less susceptible” paralog, whilst maintaining the simplicity of the assay.
976 These ratios and the qRT-PCR assays for the other three GEMs were combined into a
977 single predicted damage score for each of the Danish Test Panel, which could then be

978 compared with the observed damage scores for these trees. The combined prediction was
979 correlated with the log mean damage scores for 2013-14 ($R^2=0.25$, $P=6.9 \times 10^{-5}$) which gave
980 a small improvement in predictive power from the previous analysis ($R^2=0.24$, $p<8.4 \times 10^{-5}$).

981

982 **Screening of UK *F. excelsior* accessions for markers of reduced susceptibility to ash** 983 **dieback**

984

985 Four markers were selected for predictive marker assays based on this analysis and
986 previous work on the Danish Test Panel of 58 trees³. The three GEM markers most highly
987 associated with disease damage were assayed by qRT-PCR using the following primer
988 combinations: FRAEX38873_v2_000261470.1 (GTCGAGGAGGATGGTCAGTCAT,
989 AATCTTGCGGAGGACCTATCG), FRAEX38873_v2_000199610.1
990 (GGTGAGAGGAAAGGTTCAAATGA, TGC GTTTT GAGAAGGAAACCA),
991 FRAEX38873_v2_000173540.4 (AGGGCAAGGCTTGGAAACAT,
992 TAGGCTTTTTCTAGCTGCTTGCTCA) and GAPDH reference
993 (CTGGGATCGCTCTTAGCAAGA, CGATCAAATCAATCACACGAGAA).

994

995 Using RNA extracted from the British Screening Panel, qRT-PCR reactions were performed
996 with SYBR Green fluorescence detection in a qPCR thermal cycler (ViiA™ 7, Applied
997 Biosystems, San Francisco, CA) using optical grade 384-well plates, allowing all reactions to
998 be performed simultaneously for each target gene. Each reaction was prepared using 3 µl
999 from a 2 ng/µl dilution of cDNA derived from the RT reaction, 5 µl of SYBR® Green PCR
1000 Master Mix (Applied Biosystems®), 200 nM forward and reverse primers, in a total volume of
1001 10 µl. The cycling conditions were: 2 min at 50°C, 10 min at 95°C, followed by 40 cycles of
1002 95°C for 15 sec and 60°C for 1 min with the final dissociation at 95°C for 15 sec, 60°C for 1
1003 min and 95°C for 15 sec. Three technical replicates were used for quantification analysis.
1004 Melting curve analysis was performed to evaluate the presence of non-specific PCR
1005 products and primer dimers. The specificity and uniqueness of the primers and the
1006 amplicons were verified by amplicon sequencing (GATC Biotech LIGHTrun). The results
1007 were exported as raw data, and the LinRegPCR⁹⁹ software was used for baseline correction.
1008 The resulting means of triplicate N_0 -values, representing initial concentrations of a target
1009 and reference genes were used to analyse gene expression. For each marker, the set of
1010 qRT-PCR quantifications were standardized and rescaled to better emulate the range of
1011 RPKM values observed in the original association panel, and then predicted damage scores
1012 generated using the regression coefficient and constant from the GEM associations.

1013

1014 An additional GEM marker was assayed as a cSNP by PCR using 1ul undiluted cDNA, 11.5
1015 ul Thermo Scientific Fermentas PCR Master Mix (2X), 200 nM forward
1016 (GGTTTCTCTTCTGCAGCGAG) and reverse (TCCATGATCATCTTGCTGAG) primers in a
1017 total volume of 25 µl. The touchdown PCR was performed in using a BIORAD Tetrad PCR
1018 machine with the following cycling conditions: 5 min at 94°C, followed by 15 cycles of 94°C
1019 for 30 sec, 63°C for 30 sec -1°C/cycle, 72°C for 1 min, and 30 cycles of 94°C for 30 sec,
1020 53°C for 30 sec, 72°C for 1 min and a final elongation step at 72°C for 7 mins.

1021

1022 Sanger sequences obtained using the forward primer co-amplify GEM
1023 FRAEX38873_v2_000048360.1, which is highly associated with ash dieback disease
1024 damage, and another member of the gene family that is not. Due to a polymorphism
1025 between the two (at position 203 of the CDS model mentioned above), the relative
1026 abundance of the G nucleotide found in the highly associated GEM can be scored by eye
1027 relative to the A nucleotide found in the other paralog as a cSNP. Previously (Harper *et al.*,
1028 2016), this marker was scored in the Danish Test Panel as the presence or absence of a G
1029 nucleotide at this position, but predictions using this method did not incorporate the dynamic
1030 range of the gene expression observed, so for this analysis G:A peak height ratios were

1031 approximated directly from the sequence chromatograms using Softgenetics Mutation
1032 Surveyor® software for the British Screening Panel and the Danish Test Panel. These ratios
1033 were then standardized and rescaled to the RPKM values for
1034 FRAEX38873_v2_000048360.1 in order to predict damage scores as before.

1035

1036 Combined predictions were made by ranking and standardizing the individual predictions for
1037 all four markers, and then calculating the mean rank score for each individual tree
1038 (Supplementary Data 6). Combined predictions were calculated for the Danish Test Panel
1039 and compared to the observed ash dieback damage scores to ensure that the assay was
1040 predictive (Fig. 3).

1041

1042 The four assays were applied in the same way to analyse a panel of 130 accessions
1043 originating from across the UK range of *F. excelsior* ("British Screening Panel"). Strikingly,
1044 when assayed by RT-PCR, expression of the "G" variant paralogs was seen at much higher
1045 frequency in the British Screening Panel than in the Danish panels and the mean G:A ratio
1046 across the British Screening Panel was 0.67 compared to a mean of 0.03 observed in the
1047 Danish Test Panel. Likewise, the gene expression estimates for the British Screening Panel
1048 exhibited wider ranges and were more favourable in terms of their expected effect on
1049 damage scores. The qRT-PCR results for the GEMs negatively correlated with disease
1050 damage (FRAEX38873_v2_000261470.1 and FRAEX38873_v2_000199610.1) exhibited
1051 higher mean expression in the UK (0.1 ± 0.11 and 0.12 ± 0.14) versus the Danish Test Panel
1052 (0.09 ± 0.08 , 0.12 ± 0.11), and the positively correlated FRAEX38873_v2_000173540.4 was on
1053 average expressed at a lower level in the British Screening Panel (0.48 ± 0.26) than the
1054 Danish Test Panel (0.59 ± 0.17). As expected, this translated to lower combined predictions
1055 for ash dieback damage in the British Screening Panel. Only 9% of the Danish Test Panel
1056 accessions were predicted to have a low damage score (defined as 25% canopy damage or
1057 less) compared with 25% of the British Screening Panel (Fig. 3).

1058

1059 **Analysis of predictive genes**

1060

1061 In order to predict the susceptibility of the reference tree 2451S to ADB we calculated RPKM
1062 values for the five GEM marker CDS models (FRAEX38873_v2_000173540.4,
1063 FRAEX38873_v2_000048340.1, FRAEX38873_v2_000048360.1,
1064 FRAEX38873_v2_000261470.1 and FRAEX38873_v2_000199610.1) from leaf
1065 transcriptome read data. We also did this for each of the trees in the Danish Scoring Panel,
1066 and the average of these predictions taken to provide combined predictions. The top and
1067 bottom quartiles from the distribution of predicted scores, which represent the trees with the
1068 most susceptible and least susceptible gene expression patterns at these five loci, were then
1069 correlated with the RPKM values for the genome sequenced tree 2451S (Extended Data Fig.
1070 4).

1071

1072 RPKM data were also generated for four tissue types: leaf, flower, cambium and root, of the
1073 parent of sequenced tree 2451S by mapping raw reads to the CDS reference as before.
1074 RPKM data for the 20 CDS models found to be significantly associated with susceptibility to
1075 ADB in the GEM analysis were selected and compared for the four tissue types.

1076

1077 The five CDS models represented in the ADB susceptibility predictions were translated using
1078 the standard codon usage table and were searched against the nr database in GenBank
1079 using BLASTP with default settings to identify top hits to protein sequences in *A. thaliana*:
1080 FRAEX38873_v2_000199610.1 and FRAEX38873_v2_000261470.1 show high similarity to
1081 AGAMOUS-LIKE 42/FOREVER YOUNG FLOWER (AGL42/FYF; AT5G62165);
1082 FRAEX38873_v2_000173540.4, FRAEX38873_v2_000048340.1 and
1083 FRAEX38873_v2_000048360.1 have top hits to SHORT VEGETATIVE PHASE/AGAMOUS-

1084 LIKE 22 (SVP/AGL22; AT2G22540). Both AGL42/FYF and SVP/AGL22 are encoded by
1085 type II MADS-box genes¹⁶. To find potential orthologues from other species, we examined
1086 the results of the OrthoMCL analysis for clusters containing AGL42/FYF and SVP/AGL22; all
1087 sequences from these clusters were extracted and added to the appropriate *F. excelsior*
1088 sequences to create two datasets, one of AGL42/FYF-like sequences and one of
1089 SVP/AGL22-like sequences. To ensure adequate representation of putative orthologues, we
1090 further expanded these datasets to include sequences from the OrthoMCL clusters
1091 containing *A. thaliana* proteins from closely related MADS lineages, as identified by previous
1092 phylogenetic analyses of type II MADS-box sequences^{16,17}.

1093
1094 Preliminary phylogenetic analysis of these datasets revealed that, despite showing high
1095 sequence similarity in BLAST searches, FRAEX38873_v2_000048340.1 and
1096 FRAEX38873_v2_000048360.1 do not fall within the clade containing SVP/AGL22 and
1097 related *A. thaliana* sequences. Therefore, to identify potentially more closely related
1098 sequences we performed a BLASTP search of FRAEX38873_v2_000048340.1 and
1099 FRAEX38873_v2_000048360.1 against the complete set of 362,741 protein sequences
1100 used for the OrthoMCL analysis (see Supplementary Table 10), using the BLAST+
1101 package⁴⁴ (v.2.2.31+) with an e-value cut-off of 1e-05 (FRAEX38873_v2_000048340.1 and
1102 FRAEX38873_v2_000048360.1 were not included in the OrthoMCL analysis because they
1103 were flagged as putative TE-related genes during annotation). This identified several highly
1104 similar sequences from other species with better ranking BLAST hits than those to the *A.*
1105 *thaliana* proteins. These sequences belong to a single OrthoMCL cluster, and include a
1106 tomato (*S. lycopersicum*) sequence from the apparent orthologue of the potato (*S.*
1107 *tuberosum*) *StMADS11* gene; all sequences from this cluster were added to the
1108 SVP/AGL22-like dataset, along with the potato *StMADS11* protein (GenBank accession
1109 ACH53556.1).

1110
1111 Sequences for both datasets were aligned using M-Coffee¹⁰⁰, via the T-Coffee web server
1112 (www.tcoffee.org; last accessed 01.06.16) with the following parameter settings:
1113 Mpcma_msa Mmafft_msa Mclustalw_msa Maligntx_msa Mpoa_msa Mmuscle_msa
1114 Mprobcons_msa Mt_coffee_msa -output=score_html clustalw_aln fasta_aln score_ascii
1115 phylip -tree -maxnseq=150 -maxlen=2500 -case=upper -seqnos=on -outorder=input -
1116 run_name=result -multi_core=4 -quiet=stdout. Positions in the alignments with consensus
1117 scores of <6 from M-Coffee were removed; filtered alignments were then run through the
1118 TCS tool¹⁰¹ via the T-Coffee web server and any positions with a reliability score of <6 were
1119 removed. Recombination was tested for in the filtered alignments using GARD¹⁰². Analyses
1120 were run via the Datamonkey server (www.datamonkey.org; last accessed 01.06.16) under
1121 the best-fit model of evolution (selected with the corrected Akaike's Information Criterion,
1122 AICc¹⁰³) with β - Γ rate variation and three rate classes. No breakpoints with significant
1123 topological incongruence at $p \leq 0.05$ were detected for either dataset. Phylogenetic analysis
1124 of each dataset was conducted using Bayesian inference in MrBayes and maximum
1125 likelihood in RAxML; input alignments are provided in Supplementary Data 8. MrBayes
1126 (v.3.2.5¹⁰⁴) was run using the mixed amino acid model, to allow models of protein sequence
1127 evolution to be fit automatically across the alignments; the following parameter settings were
1128 used for each dataset: prset aamodelpr = mixed, mcmc nruns = 2, nchains = 4, ngen =
1129 1000000, samplefreq = 1000. Parameter values from both runs for each dataset were
1130 viewed in TRACER v1.6 (<http://beast.bio.ed.ac.uk/Tracer>) to confirm that effective sample
1131 sizes of >200 had been obtained for each parameter and stationarity reached. Trees
1132 sampled during the first 100000 generations of each run were discarded as the burn-in; trees
1133 and parameter values were summarised in MrBayes using the sumt and sump commands.
1134 RAxML (v.8.2.8¹⁰⁵) was run using the option to automatically determine the best protein
1135 substitution model, with 1000 replicates of the rapid bootstrap algorithm; parameter settings
1136 were as follows: raxmlHPC -f a -x 13102 -p 29503 -# 1000 -m PROTGAMMAAUTO.

1137

1138 The phylogenetic analysis suggested that FRAEX38873_v2_000173540.4 is a likely
1139 orthologue of the *A. thaliana* *SVP/AGL22* gene, or possibly *AGL24*, whereas
1140 FRAEX38873_v2_000048340.1 and FRAEX38873_v2_000048360.1 appear orthologous to
1141 the potato *StMADS11* gene (Extended Data Fig. 5). These all belong to the SVP/StMADS11
1142 group¹⁶ of type II MADS-box genes. FRAEX38873_v2_000261470.1 and
1143 FRAEX38873_v2_000199610.1 cluster with the *A. thaliana* SUPPRESSOR of
1144 OVEREXPRESSION of CONSTANS 1(SOC1)-like proteins AGL42, AGL71 and AGL72
1145 (Extended Data Fig. 5). The two other major clades within the phylogenetic tree include the
1146 AGL20/SOC1 protein and the AG14 and AGL19 proteins (Extended Data Fig. 5); together,
1147 the AGL42/AGL71/AGL72, AL20 and AGL14/AGL19 containing clades are known as the
1148 SOC1/TM3 group of type II MADS-box proteins^{16,17}.

1149

1150 In *A. thaliana*, *AGL42*, *AGL71* and *AGL72* have redundant functions in controlling flowering
1151 time and appear to be regulated by *AGL20/SOC1*²⁰. In turn, *AGL20/SOC1* is regulated by
1152 both *AGL22/SVP* and *AGL24*^{18,19}, which are floral meristem identity genes with redundant
1153 functions during early stages of flower development²¹. The *StMADS11* gene does not appear
1154 to have a direct orthologue in *A. thaliana*, but in potato (*S. tuberosum*) *StMADS11* is
1155 expressed in vegetative tissues¹⁰⁶. Despite their well-known roles in floral regulation,
1156 *SVP/StMADS11* and *SOC1/TM3* proteins are likely to have wider functions. In *A. thaliana*, it
1157 is suggested that *AGL22/SVP* is also required for age-related resistance (ARR), which gives
1158 older tissues of plants enhanced pathogen tolerance or resistance²⁴. The *Brassica rapa*
1159 *BrMADS44* gene, which appears orthologous to *AGL42*, shows differential expression in
1160 response to cold and drought stress; some *B. rapa* genes belonging to the SVP/StMADS11
1161 clade are also differentially expressed in response to these stresses, indicating a potential
1162 role in stress resistance²². Furthermore, many genes involved in regulation of flowering time
1163 in *A. thaliana* are involved in controlling phenology in perennial trees species and genes
1164 belonging to the SVP/StMADS11 clade have potential roles in growth cessation, bud set and
1165 dormancy²³.

1166

1167 **Metabolomic profiling**

1168

1169 In order to understand if trees with low and high susceptibility vary in their metabolite profiles
1170 as well as their transcriptomes, we undertook untargeted metabolite profiling on a subset of
1171 the Danish Test Panel. Untargeted metabolomics has not previously been applied to natural
1172 populations but has the potential to identify small molecules (or small molecule associations)
1173 that directly contribute to tolerance or resistance. We compared triplicate samples from five
1174 low-susceptibility Danish trees (R-14164C, R-14184A, R-14193A, R-14198B, R-14181) and
1175 five high-susceptibility trees (R-14169, R-14127, R-14156 R-14120, 25UTaps).

1176

1177 Three leaflets from each triplicate sample were freeze dried and gently crushed to mix
1178 tissue. Approximately 100-150mg was ground to a fine powder using a TissueLyser
1179 (Qiagen), and 10mg was extracted in 400µl 80% MeOH containing d5-IAA internal standard
1180 at 2.5ng/ml ([²H₅] indole-3-acetic acid; OIChemIm Ltd, Czech Republic), centrifuged
1181 (10,000g, 4°C, 10 min) and the pellet re-extracted in 80% MeOH. The pooled supernatants
1182 were filtered through a 0.2µm syringe filter (Phenomenex, UK).

1183

1184 These leaf extracts (5 µl) were analysed using a Polaris C18 1.8 µm, 2.1 x 250 mm reverse
1185 phase analytical column (Agilent Technologies, Palo Alto, USA) and samples resolved on an
1186 Agilent 1200 series Rapid Resolution HPLC system coupled to a quadrupole time-of-flight
1187 QToF 6520 mass spectrometer (Agilent Technologies, Palo Alto, USA). Buffers were as
1188 follows: positive ion mode; mobile phase A (5% acetonitrile, 0.1% formic acid), mobile phase
1189 B (95% acetonitrile with 0.1% formic acid). Negative ion mode; mobile phase A (5%

1190 acetonitrile with 1mM ammonium fluoride), mobile phase B (95% acetonitrile). The following
1191 gradient was used: 0 - 10 min – 0% B; 10-30 min – 0 - 100% B; 30 - 40 min – 100% B. The
1192 flow rate was 0.25 ml min⁻¹ and the column temperature was held at 35 °C throughout. The
1193 source conditions for electrospray ionisation were as follows: gas temperature was 325 °C
1194 with a drying gas flow rate of 9 l min⁻¹ and a nebuliser pressure of 35 psig. The capillary
1195 voltage was 3.5 kV in both positive and negative ion mode. The fragmentor voltage was 115
1196 V and skimmer 70 V. Scanning was performed using the autoMS/MS function at 4 scans
1197 sec⁻¹ for precursor ion surveying and 3 scans sec⁻¹ for MS/MS with a sloped collision energy
1198 of 3.5 V/100 Da with an offset of 5 V.

1199
1200 Positive and negative ion data was converted into mzData using the export option in Agilent
1201 MassHunter. Peak identification and alignment was performed using the Bioconductor R
1202 package xcms¹⁰⁷ and features were detected using the centWave method¹⁰⁸ for high
1203 resolution LC/MS data in centroid mode at 30 ppm. Changes to the default parameters were:
1204 mzdifff=0.01, peakwidth=10-80, noise=1000, prefilter=3,500. Peaks were matched across
1205 samples using the density method with a bw=5 and mzwid=0.025 and retention time
1206 correlated using the obiwrap algorithm with profStep=0.5. Missing peak data was filled in the
1207 peaklists generated from the ADB low susceptibility ash leaf samples compared to the
1208 peaklists generated from the ADB susceptible leaves. The resulting peaklists were
1209 annotated using the Bioconductor R package, CAMERA¹⁰⁹. The peaks were grouped using
1210 0.05 % of the width of the full width at half maximum (FWHM) and groups correlated using a
1211 p-value of 0.05 and calculating correlation inside and across samples. Isotopes and adducts
1212 were annotated using a 10 ppm error.

1213
1214 Statistical analysis and modelling was performed using MetaboAnalyst v3.0 with the
1215 following parameters. Missing values were replaced using a KNN missing value estimation.
1216 Data was filtered (40%) to remove non-informative variables using the interquartile range
1217 (IQR). Samples were normalised using the internal standard d5-IAA (POS: M181T1448;
1218 NEG: M179T1382). Data was auto-scaled.

1219
1220 Peaks from the three replicates were aligned with xcms for both positive and negative mode
1221 and features tested for practical significance to determine the differences between the
1222 tolerant and susceptible genotypes. In addition, PLS-DA was performed using
1223 MetaboAnalyst allowing the discrimination of tolerant and susceptible genotypes based on
1224 their metabolic profiles (Fig. 4a).

1225
1226 The individual features (putative metabolites) that contribute to the separation between the
1227 different classes were further characterised. We first applied a range of univariate and
1228 multivariate statistical tests to determine the importance of these features. This included
1229 variable influence on the projection (VIP) values derived from PLS-DA scores, practical
1230 significance, t-test, p-value, Benjamini and Hochberg FDR (False Discovery Rate) p-value,
1231 effect size and Random Forest analysis, and MS/MS fragmentation network analysis. For
1232 example, using Random Forest, significant features were ranked by mean decrease in
1233 classification accuracy with 14/15 susceptible samples (OOB error: 0.033; class error 0.07)
1234 and 15/15 tolerant samples correctly classified.

1235
1236 For all further analyses we chose to use statistical and practical significance (Response
1237 screening, JMP version 12) to identify features with a practical significance for identification.
1238 A combination of k-means clustering was used to group features by patterns of abundance
1239 and also by retention time. This enabled the clustering of base peaks with their associated
1240 isotopes and adducts. Product ions were identified using MS/MS data in Agilent MassHunter
1241 Qualitative Analysis version 4.

1242

1243 Identification was not possible for those features with no fragmentation, or lacking significant
1244 supporting adducts. Many features of interest were identified but require further work to
1245 provide confident attributions, while some features did not provide fragmentation patterns.
1246 We thus restricted further identification and characterisation to a highly discriminatory class
1247 of compounds of the iridoid glycoside class and predominantly compounds previously
1248 recorded in Oleaceae, summarised in Extended Data Figs 6-9 and Supplementary Data 9.
1249 We validated these identifications using three methods: MS/MS fragmentation networking
1250 (Fig 4c), MS/MS mirror plot (Extended Data Figure 6) and accurate mass MS/MS product ion
1251 structure correlation (Extended Data Figure 7). The MS/MS fragmentation network was
1252 generated after extracting the m/z of the MS/MS product ions from the discriminatory
1253 features using MassHunter Qualitative Analysis Version 4 and visualized using Cytoscape
1254 indicating product ion masses which have been previously reported from fragmentation of
1255 iridoid glycosides¹¹⁰. Further validation was performed through a mirror plot comparing the
1256 MS/MS spectra of four features (N2-5) detected in negative mode with an ESI-TOF/IT-MS
1257 spectra of elenolic acid glucoside taken from the literature¹¹². Finally, the accurate mass of
1258 MS/MS product ions from four discriminatory features identified in negative mode (N1-N4)
1259 were correlated with the structure of the putatively identified compound using MassHunter
1260 Molecular Structure Correlator (Agilent).

1261
1262 A timeline for the project may be found in Supplementary Table 14.

1263

1264

1265 **URLS**

1266

1267 Genome website: www.ashgenome.org

1268

1269 **Data availability**

1270 The reference tree is growing at Earth Trust with accession number: 2451S. Trimmed DNA
1271 and RNA reads and the final assembly for the 2451S genome sequence, as well as RNA reads
1272 for parent tree and raw reads and consensus read mappings of the European diversity panel
1273 trees have been deposited in European Nucleotide Archive (EMBL-EBI) with the project
1274 accession code "PRJEB4958" (<http://www.ebi.ac.uk/ena/data/view/PRJEB4958>).

1275 Metabolomic data that support the findings of this study have been deposited in

1276 MetaboLights with the accession code "MTBLS372"

1277 (www.ebi.ac.uk/metabolights/MTBLS372).

1278

1279

1280

1281

1282

1283

1284

Literature Cited in Methods

- 1285
1286
1287 31. Boshier, D. *et al.* *Ash species in Europe: biological characteristics and practical*
1288 *guidelines for sustainable use.* (Oxford Forestry Institute, University of Oxford, 2005).
1289 32. Doyle, J. J. A rapid DNA isolation procedure for small quantities of fresh leaf tissue.
1290 *Phytochem bull* **19**, 11–15 (1987).
1291 33. Obermayer, R., Leitch, I. J., Hanson, L. & Bennett, M. D. Nuclear DNA C₀ values in 30
1292 Species Double the Familial Representation in Pteridophytes. *Ann. Bot.* **90**, 209–217
1293 (2002).
1294 34. Dolezel, J., Greilhuber, J. & Suda, J. Estimation of nuclear DNA content in plants using
1295 flow cytometry. *Nat. Protoc.* **2**, 2233–2244 (2007).
1296 35. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve
1297 genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
1298 36. Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D. & Caccamo, M. NextClip: an
1299 analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics*
1300 **30**, 566–568 (2014).
1301 37. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-
1302 assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
1303 38. English, A. C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-
1304 read sequencing technology. *PLoS One* **7**, e47768 (2012).
1305 39. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of
1306 occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
1307 40. Gomez-Alvarez, V., Teal, T. K. & Schmidt, T. M. Systematic artifacts in metagenomes
1308 from complex microbial communities. *ISME J.* **3**, 1314–1317 (2009).
1309 41. Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-
1310 based web server for genome-wide characterization of eukaryotic repetitive elements
1311 from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).
1312 42. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
27

- 1313 search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- 1314 43. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic*
1315 *Acids Res.* **27**, 573–580 (1999).
- 1316 44. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
1317 (2009).
- 1318 45. Lamesch, P. *et al.* The *Arabidopsis* Information Resource (TAIR): improved gene
1319 annotation and new tools. *Nucleic Acids Res.* **40**, D1202–10 (2012).
- 1320 46. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence
1321 comparison. *BMC Bioinformatics* **6**, 31 (2005).
- 1322 47. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for
1323 mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
- 1324 48. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal
1325 RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
- 1326 49. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with
1327 RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
- 1328 50. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals
1329 unannotated transcripts and isoform switching during cell differentiation. *Nat.*
1330 *Biotechnol.* **28**, 511–515 (2010).
- 1331 51. Perteira, M. *et al.* StringTie enables improved reconstruction of a transcriptome from
1332 RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
- 1333 52. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the
1334 Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512
1335 (2013).
- 1336 53. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for
1337 gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–12 (2004).
- 1338 54. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using
1339 EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**,

- 1340 R7 (2008).
- 1341 55. Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal
1342 transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
- 1343 56. Lara, A. J. *et al.* in *Innovations in Hybrid Intelligent Systems* 361–368 (Springer Berlin
1344 Heidelberg, 2007).
- 1345 57. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer
1346 RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
- 1347 58. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes.
1348 *Nucleic Acids Res.* **35**, 3100–3108 (2007).
- 1349 59. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs
1350 using deep sequencing data. *Nucleic Acids Res.* **42**, D68–73 (2014).
- 1351 60. Stocks, M. B. *et al.* The UEA sRNA workbench: a suite of tools for analysing and
1352 visualizing next generation sequencing microRNA and small RNA datasets.
1353 *Bioinformatics* **28**, 2059–2061 (2012).
- 1354 61. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
- 1355 62. Prüfer, K. *et al.* PatMaN: rapid alignment of short sequences to large databases.
1356 *Bioinformatics* **24**, 1530–1531 (2008).
- 1357 63. Muñoz-Mérida, A. *et al.* *De novo* assembly and functional annotation of the olive (*Olea*
1358 *europaea*) transcriptome. *DNA Res.* **20**, 93–108 (2013).
- 1359 64. Ibarra-Laclette, E. *et al.* Architecture and evolution of a minute plant genome. *Nature*
1360 **498**, 94–98 (2013).
- 1361 65. Denoeud, F. *et al.* The coffee genome provides insight into the convergent evolution of
1362 caffeine biosynthesis. *Science* **345**, 1181–1184 (2014).
- 1363 66. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization
1364 in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
- 1365 67. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high
1366 throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

- 1367 68. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**,
1368 1586–1591 (2007).
- 1369 69. Maere, S. *et al.* Modeling gene and genome duplications in eukaryotes. *Proc. Natl.*
1370 *Acad. Sci. U. S. A.* **102**, 5454–5459 (2005).
- 1371 70. Lyons, E., Pedersen, B., Kane, J. & Freeling, M. The Value of Nonmodel Genomes and
1372 an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the
1373 Rosids. *Trop. Plant Biol.* **1**, 181–190 (2008).
- 1374 71. Kielbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame
1375 genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
- 1376 72. Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: a tool for
1377 mining segmental genome duplications and synteny. *Bioinformatics* **20**, 3643–3646
1378 (2004).
- 1379 73. Li, L., Stoeckert, C. J., Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for
1380 eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
- 1381 74. Amborella Genome Project. The *Amborella* genome and the evolution of flowering
1382 plants. *Science* **342**, 1241089 (2013).
- 1383 75. Kaul, S. *et al.* Analysis of the genome sequence of the flowering plant *Arabidopsis*
1384 *thaliana*. *Nature* **408**, 796–815 (2000).
- 1385 76. Young, N. D. *et al.* The *Medicago* genome provides insight into the evolution of rhizobial
1386 symbioses. *Nature* **480**, 520–524 (2011).
- 1387 77. Zimin, A. *et al.* Sequencing and assembly of the 22-gb loblolly pine genome. *Genetics*
1388 **196**, 875–890 (2014).
- 1389 78. Tuskan, G. A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. &
1390 Gray). *Science* **313**, 1596–1604 (2006).
- 1391 79. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale
1392 detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
- 1393 80. Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P. & Minghim, R. InteractiVenn:

- 1394 a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics*
1395 **16**, 169 (2015).
- 1396 81. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat.*
1397 *Methods* **9**, 357–359 (2012).
- 1398 82. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
1399 2078–2079 (2009).
- 1400 83. Etherington, G. J., Ramirez-Gonzalez, R. H. & MacLean, D. bio-samtools 2: a package
1401 for analysis and visualization of sequence and alignment data with SAMtools in Ruby.
1402 *Bioinformatics* **31**, 2565–2567 (2015).
- 1403 84. Cingolani, P. *et al.* A program for annotating and predicting the effects of single
1404 nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster*
1405 strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
- 1406 85. Ramirez-Gonzalez, R. H., Uauy, C. & Caccamo, M. PolyMarker: A fast polyploid primer
1407 design pipeline. *Bioinformatics* **31**, 2038–2039 (2015).
- 1408 86. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158
1409 (2011).
- 1410 87. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal
1411 component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
- 1412 88. Hubisz, M. J., Falush, D., Stephens, M. & Pritchard, J. K. Inferring weak population
1413 structure with the assistance of sample group information. *Mol. Ecol. Resour.* **9**, 1322–
1414 1332 (2009).
- 1415 89. Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: a website and program for
1416 visualizing STRUCTURE output and implementing the Evanno method. *Conserv.*
1417 *Genet. Resour.* **4**, 359–361 (2011).
- 1418 90. Jakobsson, M. & Rosenberg, N. A. CLUMPP: a cluster matching and permutation
1419 program for dealing with label switching and multimodality in analysis of population
1420 structure. *Bioinformatics* **23**, 1801–1806 (2007).

- 1421 91. Buschiazzo, E., Ritland, C., Bohlmann, J. & Ritland, K. Slow but not low: genomic
1422 comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to
1423 angiosperms. *BMC Evol. Biol.* **12**, 8 (2012).
- 1424 92. Barbato, M., Orozco-terWengel, P., Tapio, M. & Bruford, M. W. SNeP: a tool to estimate
1425 trends in recent effective population size trajectories using genome-wide SNP data.
1426 *Front. Genet.* **6**, 109 (2015).
- 1427 93. Megléc, E. *et al.* QDD version 3.1: a user-friendly computer program for microsatellite
1428 selection and primer design revisited: experimental validation of variables determining
1429 genotyping success rate. *Mol. Ecol. Resour.* **14**, 1302–1313 (2014).
- 1430 94. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist
1431 programmers. *Methods Mol. Biol.* **132**, 365–386 (2000).
- 1432 95. Bancroft, I. *et al.* Dissecting the genome of the polyploid crop oilseed rape by
1433 transcriptome sequencing. *Nat. Biotechnol.* **29**, 762–766 (2011).
- 1434 96. Popescu, A.-A., Harper, A. L., Trick, M., Bancroft, I. & Huber, K. T. A novel and fast
1435 approach for population structure inference using kernel-PCA and optimization.
1436 *Genetics* **198**, 1421–1431 (2014).
- 1437 97. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association
1438 studies. *Nat. Genet.* **42**, 355–360 (2010).
- 1439 98. Lipka, A. E. *et al.* GAPIT: genome association and prediction integrated tool.
1440 *Bioinformatics* **28**, 2397–2399 (2012).
- 1441 99. Ruijter, J. M. *et al.* Amplification efficiency: linking baseline and bias in the analysis of
1442 quantitative PCR data. *Nucleic Acids Res.* **37**, e45 (2009).
- 1443 100. Di Tommaso, P. *et al.* T-Coffee: a web server for the multiple sequence alignment of
1444 protein and RNA sequences using structural information and homology extension.
1445 *Nucleic Acids Res.* **39**, W13–7 (2011).
- 1446 101. Chang, J.-M., Di Tommaso, P. & Notredame, C. TCS: a new multiple sequence
1447 alignment reliability measure to estimate alignment accuracy and improve phylogenetic

- 1448 tree reconstruction. *Mol. Biol. Evol.* **31**, 1625–1637 (2014).
- 1449 102. Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. W.
1450 GARD: a genetic algorithm for recombination detection. *Bioinformatics* **22**, 3096–3098
1451 (2006).
- 1452 103. Sugiura, N. Further analysts of the data by akaike's information criterion and the finite
1453 corrections: Further analysts of the data by akaike's. *Communications in Statistics-*
1454 *Theory and Methods* **7**, 13–26 (1978).
- 1455 104. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model
1456 choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
- 1457 105. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
1458 large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
- 1459 106. Carmona, M. J., Ortega, N. & Garcia-Maroto, F. Isolation and molecular characterization
1460 of a new vegetative MADS-box gene from *Solanum tuberosum* L. *Planta* **207**, 181–188
1461 (1998).
- 1462 107. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing
1463 mass spectrometry data for metabolite profiling using nonlinear peak alignment,
1464 matching, and identification. *Anal. Chem.* **78**, 779–787 (2006).
- 1465 108. Tautenhahn, R., Böttcher, C. & Neumann, S. Highly sensitive feature detection for high
1466 resolution LC/MS. *BMC Bioinformatics* **9**, 504 (2008).
- 1467 109. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: an
1468 integrated strategy for compound spectra extraction and annotation of liquid
1469 chromatography/mass spectrometry data sets. *Anal. Chem.* **84**, 283–289 (2012).
- 1470 110. Li, C.-M. *et al.* Structural characterization of iridoid glucosides by ultra-performance
1471 liquid chromatography/electrospray ionization quadrupole time-of-flight tandem mass
1472 spectrometry. *Rapid Commun. Mass Spectrom.* **22**, 1941–1954 (2008).
- 1473 111. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements.
1474 *Nat. Rev. Genet.* **8**, 973–982 (2007).

1475 112. Gupta, S. D. *Reactive oxygen species and antioxidants in higher plants*. (CRC Press,
1476 2010).

1477

1478 **Acknowledgments**

1479 Eurofins MWG GmbH provided a discounted service for Illumina and 454 sequencing of the
1480 reference genome, funded by NERC Urgency Grant, NE/K01112X/1 to R.J.A.B.. The
1481 associative transcriptomic and metabolomic work was part of the the “Normex” project led by
1482 J.A.D. funded jointly by UK BBSRC (BBS/E/J/000CA5323) and DEFRA. The Genome
1483 Analysis Centre, Norwich, UK sequenced the European diversity panel, funded by the
1484 “Normex” project and a National Capability in Genomics (BB/J010375/1) grant. William
1485 Crowther (Queen Mary University of London) assisted with DNA extractions for the KASP
1486 assay; The John Innes Centre contributed KASP analyses. Juliana Fabiana Miranda
1487 assisted with RNA extractions and qRT-PCR at University of York. Dr Hannah Florance,
1488 Prof. Nicholas Smirnoff and the Exeter Metabolomics Facility developed metabolomic
1489 methods and ran samples, and Dr Tom Howard for helped with statistics. L.J.K and R.J.A.B.
1490 were partly funded by LWEC Tree Health and Plant Biosecurity Initiative - Phase 2 Grant,
1491 BB/L012162/1 to R.J.A.B., S.L. and Paul Jepson (University of Oxford) funded jointly by a
1492 grant from BBSRC, Defra, ESRC, the Forestry Commission, NERC and the Scottish
1493 Government, under the Tree Health and Plant Biosecurity Initiative. G.W. was funded by
1494 Teagasc Walsh Fellowship 2014001 to R.J.A.B. and G. C. D.. E.D.C was funded by Marie
1495 Skłodowska-Curie Individual Fellowship “FraxiFam” (grant agreement 660003) to E.D.C. and
1496 R.J.A.B. E.S.A.S. and J.Z. are funded by the Marie-Curie Initial Training Network
1497 INTERCROSSING. R.H.R.G. is supported by a Norwich Research Park PhD Studentship
1498 and The Genome Analysis Centre Funding and Maintenance Grant. This research utilised
1499 Queen Mary's MidPlus computational facilities, supported by QMUL Research-IT and funded
1500 by EPSRC grant EP/K000128/1 and NERC EOS Cloud.

1501

1502

1503

1504

1505

1506
1507

1508 **Author information**

1509

1510 **These authors contributed equally to this work**

1511 Elizabeth SA Sollars, Andrea L Harper, Laura J Kelly, Christine Sambles

1512

1513 **Affiliations**

1514

1515 **School of Biological and Chemical Sciences, Queen Mary University of London, Mile**
1516 **End Road, London E1 4NS, UK**

1517 Elizabeth SA Sollars, Laura J Kelly, Gemma Worwick, Jasmin Zohren, Endymion D
1518 Cooper, Richard JA Buggs

1519

1520 **QIAGEN Aarhus A/S, Silkeborgvej 2, Prismet, 8000 Aarhus C., Denmark**

1521 Elizabeth SA Sollars

1522

1523 **Royal Botanic Gardens Kew, Richmond, Surrey, TW9 3AB, UK**

1524 Richard JA Buggs

1525

1526 **Centre for Novel Agricultural Products, University of York, Heslington, York, YO10**
1527 **5DD, UK**

1528 Andrea L Harper, Lenka Havlickova, Yi Li, Zhesi He, Alison Fellgett, Ian Bancroft

1529

1530 **The Genome Analysis Centre, Norwich Research Park, Norwich, NR4 7UH, UK**

1531 Ricardo Ramirez-Gonzalez, David Swarbreck, Gemy Kaithakottil, Mario Caccamo

1532

1533 **Biosciences, College of Life and Environmental Sciences, University of Exeter, Exeter,**
1534 **UK**

1535 Christine Sambles, David J Studholme, Deborah L. Salmon, Murray Grant

1536

1537 **School of life Sciences, Gibbet Hill Campus, University of Warwick, Coventry, UK**

1538 Murray Grant (Present address)

1539

1540 **Earth Trust, Little Wittenham, Abingdon, Oxfordshire, OX14 4QZ, UK**

1541 Jo Clark

1542

1543 **Department of Plant Sciences, University of Oxford, OX1 3RB, UK**

1544 David Boshier

1545

1546 **Forest Research, Northern Research Station, Roslin, Midlothian, EH25 9SY, UK**

1547 Steve Lee

1548

1549 **Department of Geosciences and Natural Resource Management, University of**
1550 **Copenhagen, Denmark**

1551 Lea Vig McKinney, Lene Rostgaard Nielsen, Erik Dahl Kjær

1552

1553 **Teagasc, Agriculture and Food Development Authority, Ashtown, Dublin, D15 KN3K,**
1554 **Ireland**

1555 Gerry C Douglas

1556

1557 **John Innes Centre, Norwich Research Park, NR4 7UH, UK**

1558 Cristobal Uauy, J. Allan Downie

1559

1560 **National Institute of Agricultural Botany, Cambridge, CB3 0LE, UK**

1561 Mario Caccamo

1562

1563 The *F. excelsior* 2451S genome sequence and final assembly, as well as RNA reads for
1564 parent tree and raw reads and consensus read mappings of the European diversity panel
1565 trees were submitted to European Nucleotide Archive (EMBL-EBI) under project accession
1566 PRJEB4958.

1567

1568 **Author Contributions**

1569 R.J.A.B, M.C., D.S., M.G., J.A.D. and I.B. are the lead investigators. R.J.A.B. coordinated
1570 the project and directed work on the reference genome. E.S.A.S. assembled the reference
1571 genome and organellar genomes, and analysed: gene and genome duplications, European
1572 population structure, past effective population sizes. L.J.K. extracted HMW DNA for the
1573 European diversity panel and conducted repetitive element, OrthoMCL and phylogenetic
1574 analyses. G.W. conducted SSR analyses. J.Z. extracted HMW DNA and RNA for the
1575 reference genome, E.D.C. analysed genome duplication in the reference genome. D.S. and
1576 G.K. carried out bioinformatic analyses to annotate the reference genome. M.C. conceived
1577 of and, with R.J.A.B., oversaw the European-wide diversity panel sequencing. R.R.-G.,
1578 E.S.A.S. and M.C. carried out SNP calling on the European-wide diversity panel, and KASP
1579 genotyping. C.U. conducted KASP genotyping. B.J.C. conceived of and oversaw the
1580 NEXTERA sequencing on the reference tree genome. M.C., J.A.D. and B.J.C. generated the
1581 first-pass "Tree 35" Illumina reads included in the European-wide SNP analysis. E.D.K.,
1582 L.R.N. and L.V.M., generated, selected and collected Danish samples. D.B. generated and
1583 J.C. maintained and sampled the reference tree. J.C., D.B, G. C. D. and S.L. generated,
1584 selected and collected U.K. and European-wide diversity panel samples.
1585 For the associative transcriptomics: I.B and A.L.H. conceived and planned the study; A.L.H.,
1586 L.H., and A.F. performed experiments; bioinformatics was executed by Y.L. and Z.H and
1587 A.L.H. completed the data analysis. For the metabolomics: C.S., D.J.S., and M.G. conceived
1588 and conducted the analyses; C.S. developed methodology, and D.L.S. processed and
1589 extracted samples and ran the mass-spectrometer.

1590

1591 **Competing financial interests**

1592 The authors declare no competing financial interests.

1593

1594 **Corresponding author**

1595 Correspondence to Richard Buggs <r.buggs@qmul.ac.uk>

1596

1597

1598

1599

1600 Extended Data Figure Legends

1601

1602 **Extended Data Figure 1 | Completeness and coherence of annotation models. a,**
1603 Assessment of transcript completeness for the *F. excelsior* gene set. Transcripts were
1604 classified as full-length, 5'-end, 3'-end, internal, coding (ORF predicted but no blast
1605 support), unknown (no blast support), mis-assembled and putative ncRNA using Full-
1606 lengtherNEXT (v0.0.8), 76.43% of transcript models were identified as complete. **b,**
1607 Coherence in gene length between *F. excelsior* and *M. guttatus* proteins. Blast analysis (1e-
1608 5) identified 2,576 proteins that had reciprocal best hits to 2,605 *Mimulus guttatus* proteins
1609 identified as single copy in *Mimulus guttatus*, *Solanum lycopersicum*, *Solanum tuberosum*
1610 and *Vitis vinifera* (Phytosome). A high coherence in gene length was found between
1611 *Fraxinus excelsior* and *Mimulus guttatus* $r > 0.917$.

1612

1613 **Extended Data Figure 2 | Synteny between ash and monkey flower.** Syntenic dotplot
1614 between ash (vertical axis) and monkey flower (horizontal axis) showing regions of multiple
1615 synteny. Scaffolds equal to approximately 75% of the ash genome assembly for which
1616 syntenic blocks were not detected are not shown. For clarity small scaffold names are
1617 omitted.

1618

1619 **Extended Data Figure 3 | Population structure of *F. excelsior* in Europe. a,** Results
1620 from STRUCTURE; three replicates were run for $k=3$, with each replicate using a different
1621 set of 8,955 SNPs as input. Numbers refer to samples, whose locations are given in
1622 Supplementary Table 11. **b,** Delta K values for three runs of STRUCTURE of each value of k
1623 between $k=2$ and $k=19$. $k=3$ has the highest Delta K value of 32.91. **c,** Effective population
1624 size history estimated using the SNeP program, with genotype information from all 38
1625 diversity panel samples at 394,885 SNP loci.

1626

1627 **Extended Data Figure 4 | Prediction of susceptibility of reference tree.** RPKM values for
1628 leaf material from the low heterozygosity reference tree 2451S for 5 CDS models predictive
1629 for ADB. These are shown next to expression profiles for the Danish Scoring Panel with the
1630 least susceptible and most susceptible expression patterns according to the GEM analysis.

1631

1632 **Extended Data Figure 5 | Investigation of the function of GEM markers for low**
1633 **susceptibility to ash dieback.** Unrooted maximum likelihood (ML) trees from the RAxML
1634 analyses. **a,** Best scoring ML tree from the phylogenetic analysis of SVP/AGL22 and
1635 StMADS11-like sequences. **b,** Best scoring ML tree for the SOC1-like sequences. Nodes with
1636 bootstrap support values of ≥ 70 from the ML analysis and posterior probabilities of ≥ 0.95
1637 from the Bayesian analysis are indicated with asterisks. *Fraxinus excelsior* sequences are
1638 shown in blue; *A. thaliana* sequences in red. Four-letter taxon codes at the start of sequence
1639 names, where present, follow those in Extended Data Table 1. Sequence names are those
1640 from the original data files used for the orthoMCL analysis (see Supplementary Table 10),
1641 with the exception of the StMADS11 protein from potato, where the GenBank accession
1642 number is given. Common names for selected genes/proteins are annotated on the trees.
1643 Scale bars indicate the mean number of substitutions per site.

1644

1645 **Extended Data Figure 6 | MS-MS Mirror plot of elenolic acid glucoside (ESI-TOF/IT-MS)**
1646 **compared to four negative mode features (N2, N3, N4 and N5).** The spectra share four
1647 product ions in common, m/z 179, 223, 371 and 403 (elenolic acid glucoside molecular ion).
1648 These product ions correspond to a loss of a methyl and hydroxyl group (403-371), loss of
1649 hexose (403-223) which is followed by a loss of CO₂ (223-179). Elenolic acid corresponds to
1650 the secoiridoid part of oleuropein-related compounds suggesting that these four
1651 compounds are secoiridoids¹¹².

1652
1653 **Extended Data Figure 7 | Identification of MS-MS product ions for four iridoid glycoside**
1654 **related features observed in negative mode.** Predicted structure for key m/z peaks using
1655 Molecular Structure Correlator (Agilent) and the structure of putative IDs. Bonds and atoms
1656 in black are present in that product ion, whereas gray indicates loss.

1657
1658 **Extended Data Figure 8 | Identification of iridoid glycoside related metabolites in positive**
1659 **mode.** Box plots showing abundance (log₂ transformed) of features in positive mode
1660 discriminating between 5 different genotypes of high (TOL) and low (SUS) susceptibility ash
1661 trees.

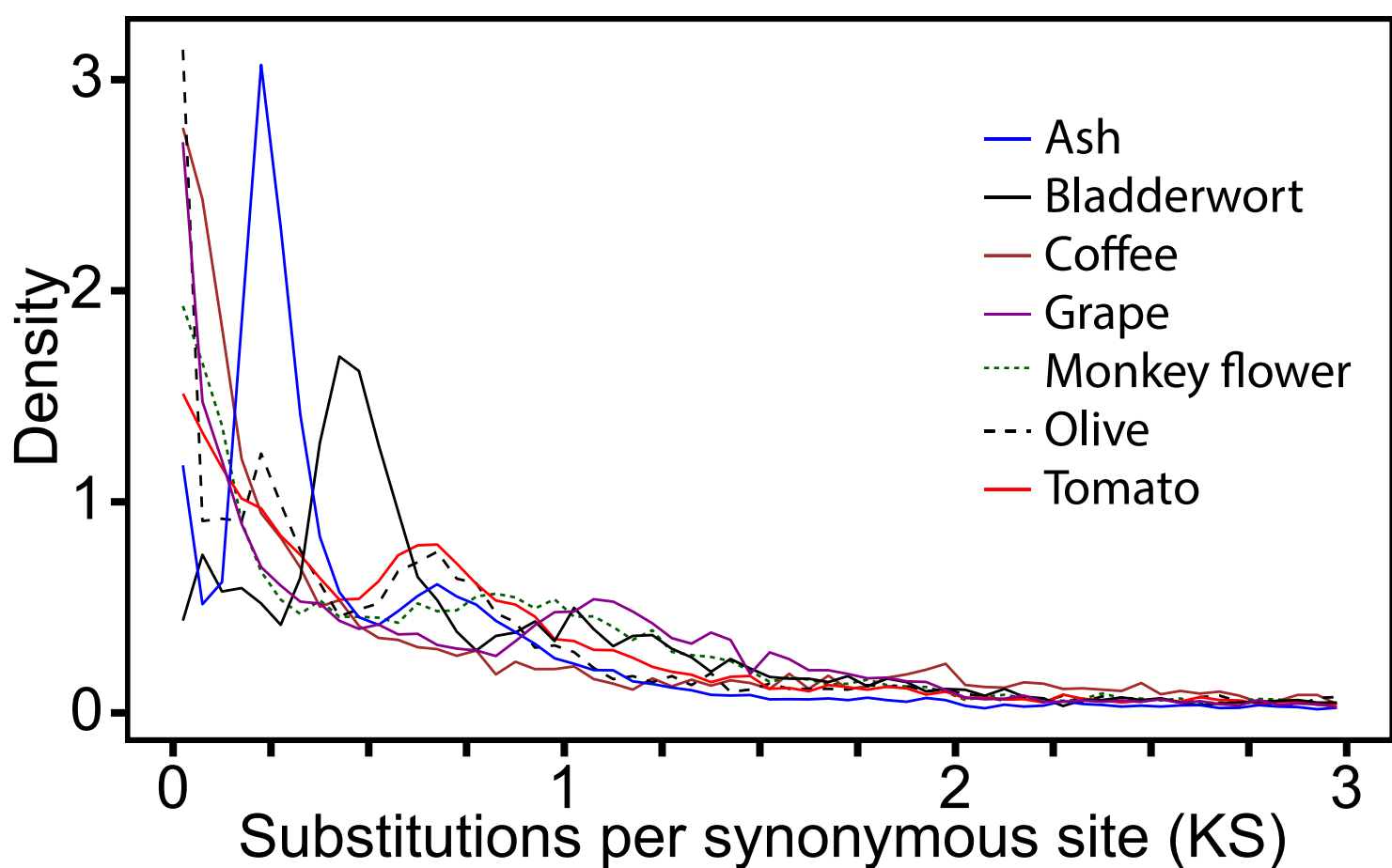
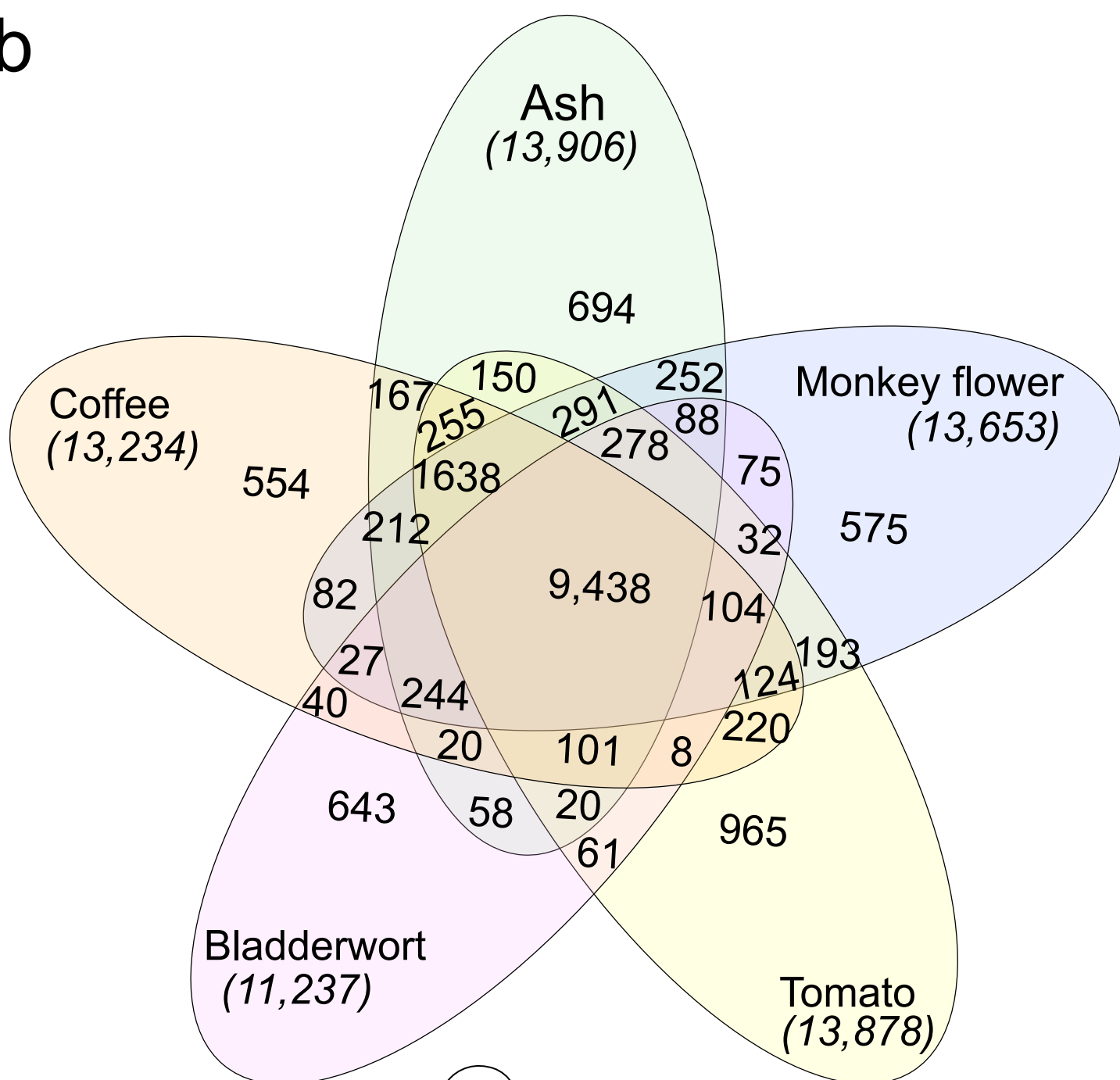
1662
1663 **Extended Data Figure 9 | Identification of metabolites.** MS/MS fragmentation product ion
1664 data of features discriminating between five different genotypes of high (TOL) and low (SUS)
1665 susceptibility ash trees in positive mode. Corresponding box-plots are presented in
1666 Extended Data Fig. 8.

1667
1668

1669 **Extended Data Table Legend**

1670
1671 **Extended Data Table 1 | The 20 largest clusters in *Fraxinus excelsior* from the**
1672 **OrthoMCL analysis of 11 species**, showing the number of sequences from each species
1673 belonging to the clusters. Clusters containing at least five more sequences from *F. excelsior*
1674 than for the other Asterid species (underlined) are shown in bold. FEXC = *Fraxinus*
1675 *excelsior*; ATHA = *Arabidopsis thaliana*; ATRI = *Amborella trichopoda*; CCAN = *Coffea*
1676 *canephora*; MGUT = *Mimulus guttatus*; MTRU = *Medicago truncatula*; PITA = *Pinus taeda*;
1677 PTRI = *Populus trichocarpa*; SLYC = *Solanum lycopersicum*; UGIB = *Utricularia gibba*; VVIN
1678 = *Vitis vinifera*. Details of gene families in column two are inferred from the gene family
1679 membership/function of *A. thaliana* genes (according to The Arabidopsis Information
1680 Resource; www.arabidopsis.org) belonging to these clusters. It should be noted that
1681 OrthoMCL clusters are not necessarily equivalent to gene families as a single gene family
1682 may be split over multiple clusters and multiple gene families may be grouped into a single
1683 cluster.

1684

a**b****c**